INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 01 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Diffusion Guided Upsampling of Sparse Indoor LiDAR Point Clouds for Occlusion Aware Scene Reconstruction Systems

Claire Morel, Luca Moretti, Ren Matsuda

Faculty of Science and Technology, University of Gothenburg, Gothenburg, Sweden

Abstract *The rapid advancement of three-dimensional scanning technologies has positioned Light Detection and Ranging as a cornerstone for indoor scene reconstruction, autonomous navigation, and spatial computing. However, raw point clouds acquired from indoor environments often suffer from severe sparsity, non-uniform density, and occlusion-induced missing regions, which severely bottleneck downstream applications. To address these critical limitations, this paper proposes a novel diffusion-guided point cloud upsampling framework specifically engineered for the high-fidelity reconstruction of complex indoor scenes. By formulating the upsampling task as a conditional generative process, our methodology leverages a denoising diffusion probabilistic model that operates directly on irregular, unordered point sets. The proposed architecture employs a hierarchical feature extraction mechanism coupled with cross-attention modules to capture both local geometric nuances and global structural topologies. The forward diffusion process systematically corrupts the underlying ground-truth dense point distribution, while the reverse generative phase progressively recovers high-resolution spatial coordinates conditioned on the sparse input geometry. Extensive experimental evaluations conducted on industry-standard indoor datasets demonstrate that our approach significantly outperforms existing optimization-based and generative adversarial network-based upsampling methods across multiple quantitative metrics and qualitative visual assessments. The proposed framework excels in preserving sharp edges, generating uniform point distributions, and avoiding clustering artifacts common in previous methodologies.*

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

1 Introduction

1.1 Context and Motivation

The digital twin paradigm and the proliferation of autonomous robotic systems have catalyzed an unprecedented demand for highly accurate three-dimensional representations of indoor environments. Light Detection and Ranging sensors have emerged as the primary modality for acquiring spatial data due to their ability to capture precise distance measurements independent of ambient lighting conditions. Unlike photogrammetry, which relies heavily on textured surfaces and optimal illumination, active sensing technologies emit laser pulses to directly measure the geometry of the physical world. The resulting data structure, known as a point cloud, provides a foundational representation for spatial computing applications ranging from

architectural surveying to virtual reality [1]. Despite their inherent advantages, indoor scanning systems face significant physical and operational constraints. Unlike airborne or automotive implementations that often map large-scale environments with high-end, multi-beam sensors, indoor data acquisition typically relies on lightweight, low-power, or solid-state devices to maintain portability and cost-effectiveness. Consequently, the raw point clouds generated by these devices exhibit extreme sparsity. Furthermore, the complex nature of indoor environments introduces a multitude of scanning artifacts [2]. Furniture, structural columns, and complex architectural geometries cause severe occlusions, leading to missing data and large gaps in the reconstructed surfaces [3]. Highly reflective surfaces such as mirrors and glass windows, as well as highly absorptive dark materials, further degrade the quality of the incoming signal. This confluence of sensor limitations and environmental complexity results in sparse, noisy, and non-uniformly distributed point sets that fail to capture the continuous surfaces required for high-fidelity scene reconstruction [4]. Point cloud upsampling aims to alleviate these issues by computationally generating a dense, uniform, and high-resolution point set from a sparse input. The objective is not merely to interpolate between existing points, but to infer the underlying continuous surface geometry and generate novel points that accurately represent the structural integrity of the scanned objects. A successful upsampling framework must preserve critical high-frequency details, such as sharp corners and thin structures, while maintaining a smooth and uniform point distribution across flat regions like walls and floors [5].

1.2 Challenges in Indoor LiDAR Reconstruction

The transition from sparse to dense point clouds in indoor environments presents several fundamental challenges that differentiate it from general three-dimensional object upsampling. First, indoor scenes are inherently unstructured and exhibit massive variations in scale and topological complexity. A single scan may encompass large planar surfaces, such as ceilings, alongside highly intricate objects, such as chandeliers or office chairs. Traditional interpolation techniques, which rely on localized polynomial fitting or distance-based heuristics, struggle to adapt to these scale variations and often produce overly smoothed corners or clustered points along edges [6]. Second, the unordered and irregular nature of point cloud data poses a significant hurdle for computational modeling. Unlike regular grid-based structures such as two-dimensional images or volumetric voxels, point clouds lack a definitive spatial neighborhood. Convolutional operations, which are the backbone of modern image processing, cannot be directly applied without introducing severe computational overhead or spatial quantization errors. While recent advancements in deep learning have introduced specialized architectures for point set processing, designing a network that effectively balances local geometric feature extraction with global contextual awareness remains a formidable task [7].

Third, upsampling is an ill-posed inverse problem. A single sparse point cloud can theoretically map to an infinite number of dense surface configurations. Previous deep learning approaches have predominantly relied on deterministic mapping functions, utilizing regression-based loss formulations to minimize the distance between the generated points and the ground truth. While effective at capturing general shapes, these deterministic models often fail to capture the stochastic variations inherent in natural surfaces, leading to repetitive patterns and a lack of fine geometric detail [8]. Generative models, particularly generative adversarial networks, have been explored to inject realism into the upsampling process, but they are notoriously difficult to train, prone to mode collapse, and often generate outlier points that require extensive post-processing [9].

1.3 Contributions of the Proposed Approach

To overcome the limitations of deterministic networks and unstable adversarial training, this research introduces a probabilistic generative framework based on denoising diffusion models. Diffusion models have recently demonstrated unparalleled success in two-dimensional image synthesis by modeling the data distribution through a gradual, iterative denoising process. We hypothesize that by adapting this paradigm to three-dimensional point sets, we can achieve superior upsampling quality, robust outlier rejection, and high-fidelity surface reconstruction.

This paper presents a comprehensive methodology that guides the diffusion process using features extracted from the sparse input point cloud. Rather than generating a shape from pure noise, our model conditions the reverse generative trajectory on the latent spatial representations of the input geometry. This ensures that the synthesized high-resolution points adhere strictly to the underlying scene structure while benefiting from the expressive power of the diffusion framework [10]. We design a novel feature extraction module that employs localized geometric reasoning and self-attention mechanisms to robustly capture the intricate topologies of indoor spaces [11].

2. Literature Review

2.1 Traditional Point Cloud Upsampling

Early approaches to point cloud densification were primarily rooted in computational geometry and numerical optimization. Moving least squares algorithms were widely adopted to smooth noisy point sets and generate new coordinates by projecting points onto locally fitted polynomial surfaces [12]. While mathematically rigorous, moving least squares methods rely heavily on the assumption that the underlying surface is locally smooth. Consequently, they tend to over-smooth high-frequency features, effectively destroying sharp edges and intricate architectural details common in indoor environments. To address the loss of high-frequency details, researchers introduced edge-aware resampling techniques. These methods typically involve normal estimation followed by the identification of sharp feature lines [13]. Points are then iteratively inserted along these identified edges and within planar regions using energy minimization frameworks. Voronoi diagrams and Delaunay triangulations have also been utilized to identify spatial gaps and guide the insertion of new points. Although these geometric priors improve the preservation of structural boundaries, they are highly sensitive to noise and initial point density [14]. In the context of raw indoor scans, where noise is prevalent and density fluctuates drastically, traditional methods often produce localized clustering artifacts and fail to generate a visually pleasing, uniform distribution.

2.2 Deep Learning based Point Cloud Processing

The paradigm shift toward data-driven methodologies revolutionized three-dimensional computer vision. The introduction of architectures capable of directly consuming raw, unordered point sets established a new baseline for point cloud analysis. These foundational networks utilized shared multilayer perceptrons and symmetric aggregation functions to achieve permutation invariance, enabling the extraction of global shape signatures [15]. Subsequent iterations improved upon this by introducing hierarchical grouping mechanisms, allowing the networks to capture local geometric contexts at multiple scales. Leveraging these architectures, researchers formulated point cloud upsampling as a supervised learning problem. Early deep learning models extracted multi-scale features from the sparse input, expanded the feature channels in the latent space, and subsequently reconstructed the dense point coordinates [16]. To encourage uniform distribution, specialized loss functions based on electrostatic repulsion were introduced, penalizing points that clustered too closely together. While these deterministic models represented a significant leap forward, their reliance on direct coordinate regression limited their ability to synthesize complex, high-fidelity surfaces. The regression objectives, typically based on average distance metrics, inherently favor conservative, smoothed outputs that minimize the overall error but lack realistic geometric texture [17].

Generative adversarial networks were subsequently adapted for point cloud upsampling to address the smoothing issues inherent in regression models. By introducing a discriminator network trained to distinguish between synthetic and real dense point sets, these frameworks encouraged the generator to produce more realistic distributions [18]. Various architectural enhancements, including graph convolutional networks and multi-stage refinement modules, were integrated into the adversarial framework to improve localized feature extraction. Despite these advancements, adversarial training remains notoriously unstable. Mode collapse, where the generator produces a limited variety of outputs, and the generation of flying outlier points in empty space remain persistent challenges that hinder the deployment of these models in high-precision mapping applications [19].

2.3 Diffusion Models in 3D Computer Vision

Denoising diffusion probabilistic models represent the state-of-the-art in generative modeling. Initially developed for thermodynamic modeling, these frameworks were successfully adapted for image generation, demonstrating superior sample quality and training stability compared to adversarial networks. The core principle involves a forward Markov chain that progressively adds Gaussian noise to the data until it reaches an isotropic Gaussian distribution. A neural network is then trained to reverse this process, iteratively removing noise to reconstruct the original data distribution [20]. The application of diffusion models to three-dimensional point clouds is a nascent but rapidly evolving research area. Early attempts focused on unconditional shape generation, demonstrating the ability to synthesize isolated objects from pure noise [21]. To adapt this for upsampling, researchers have begun exploring conditional diffusion frameworks. In these setups, the reverse generative process is guided by features derived from the sparse input scan [22]. However, most existing three-dimensional diffusion models are designed for small-scale, normalized object point clouds, such as those found in synthetic datasets. Applying these models to complex, unconstrained indoor scenes introduces significant computational and architectural hurdles. Indoor scenes cannot be easily normalized to a unit sphere, and the number of points required to represent a full room far exceeds the capacity of standard diffusion implementations [23]. Furthermore, the network must be capable of discerning between structural elements like walls and intricate objects like furniture within the same scan [24]. This necessitates the development of specialized conditioning mechanisms and memory-efficient architectures tailored specifically for large-scale scene upsampling, which forms the primary motivation for our proposed methodology [25].

3. Methodology

3.1 Problem Formulation and Network Architecture

The objective of point cloud upsampling is to transform a sparse, irregular input point set into a dense, uniform, and geometrically accurate output point set. We represent the sparse input as a matrix of spatial coordinates and the target dense output as a larger matrix containing a multiple of the original points. The scaling factor determines the desired resolution enhancement. Our goal is to model the conditional probability distribution of the dense point cloud given the sparse input point cloud. Rather than directly predicting the coordinates through a deterministic feed-forward pass, we employ a probabilistic generative approach.

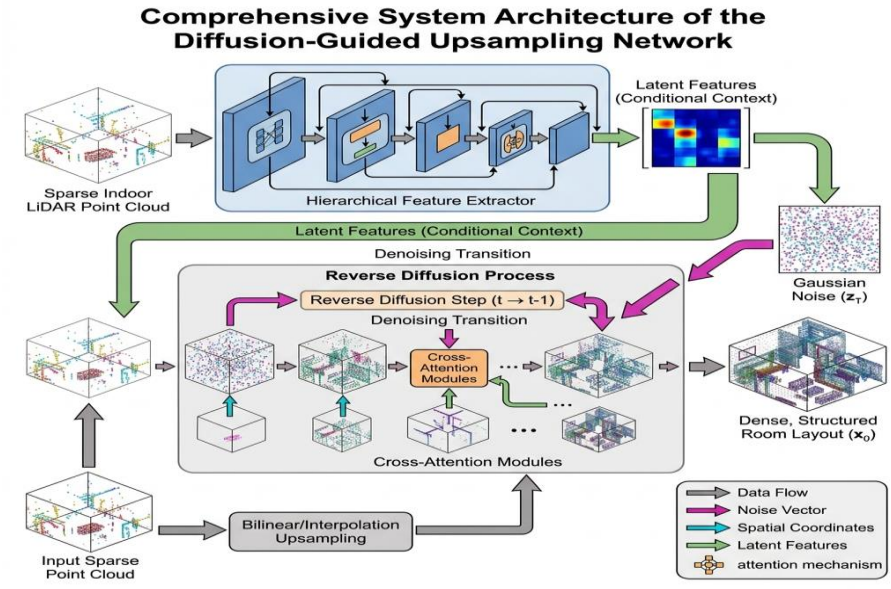


Figure 1: Comprehensive System Architecture of the Diffusion

The core of our proposed architecture consists of two primary components: a hierarchical feature extraction module and a conditional denoising diffusion module. The hierarchical feature extractor is responsible for analyzing the sparse input geometry and generating rich, multi-scale latent representations. This module utilizes a series of set abstraction layers to progressively downsample the input while increasing the receptive field, capturing both fine-

grained local details and broad structural patterns. The extracted features are then passed through an interpolation phase to ensure that dense, point-wise conditioning vectors are available for the generation phase.The conditional denoising diffusion module handles the iterative generation of the dense point cloud. This module takes an initial state of pure Gaussian noise and progressively refines it over a predefined number of timesteps. At each step, the network predicts the noise component present in the current point coordinates, utilizing the conditional features provided by the hierarchical extractor to guide the refinement trajectory. By integrating multi-head cross-attention mechanisms, the diffusion network effectively queries the spatial context of the sparse input, ensuring that the generated points adhere to the physical boundaries and surfaces of the indoor scene [26].

3.2 Forward Diffusion Process for Point Clouds

The forward diffusion process defines a fixed mathematical procedure that gradually corrupts the ground-truth dense point cloud by injecting controlled amounts of Gaussian noise. This process is modeled as a Markov chain consisting of discrete timesteps. At each step, a minor noise perturbation is added to the spatial coordinates of the points. The magnitude of this noise is regulated by a variance schedule, which starts with very small values and gradually increases as the process advances [27]. Because the forward process relies on the properties of Gaussian distributions, we can bypass the iterative step-by-step addition and directly sample the noisy point cloud at any arbitrary timestep. This is achieved by defining cumulative product terms based on the variance schedule. By expressing the state at any given timestep as a linear combination of the original clean data and a standard Gaussian noise variable, we enable highly efficient training. In the context of three-dimensional point clouds, this noise injection translates to random spatial displacements applied independently to each point coordinate. In the initial stages of the forward process, the points undergo minor jittering, slightly blurring the sharp edges and surfaces. As the timestep increases, the displacement magnitude grows, causing the geometric structure of the indoor scene to disintegrate. In the final stages, the point distribution completely loses its correlation with the original geometry, devolving into an unstructured, spherical cloud of points governed purely by an isotropic Gaussian distribution. The continuous formulation of this process ensures that the target distribution for the reverse generation phase is well-defined and analytically tractable [28].

3.3 Reverse Generative Process and Feature Extraction

The reverse generative process is designed to invert the forward diffusion trajectory, starting from a distribution of pure noise and iteratively synthesizing a highly detailed dense point cloud. Because the true reversal requires knowledge of the marginal data distribution, which is intractable, we approximate it using a deep neural network parameterized by learnable weights. This network is tasked with predicting the specific noise vector that was added to the points at a given timestep. By subtracting this predicted noise, the model takes a step backward, recovering a slightly cleaner version of the geometry. To ensure that the generated points form the specific indoor scene rather than an arbitrary shape, the reverse process must be tightly conditioned on the sparse input scan. We implement this conditioning through a specialized feature integration block. For every point in the noisy intermediate state, the network utilizes a k-nearest neighbors search to identify the closest points in the sparse input geometry. The latent features corresponding to these neighbors are aggregated using an attention mechanism, which computes relevance weights based on spatial proximity and feature similarity [29].

Caption: Code Listing 1: PyTorch pseudo-code for the Conditional Reverse Diffusion Step

```
def conditional_reverse_step(noisy_points, sparse_features, timestep, model):
    # Obtain timestep embeddings for the current diffusion step
    time_emb = get_timestep_embedding(timestep)

    # Query structural context from the sparse input via cross-attention
    context_features = cross_attention(noisy_points, sparse_features)

    # Concatenate noisy coordinates, time embeddings, and structural context
```

```
    network_input = concatenate(noisy_points, time_emb, context_features)

    # Predict the noise injected at the current timestep
    predicted_noise = model(network_input)

    # Compute the less noisy point coordinates using the diffusion scheduler
    alpha_t = scheduler.get_alpha(timestep)
    alpha_bar_t = scheduler.get_alpha_bar(timestep)
    beta_t = scheduler.get_beta(timestep)

    # Denom computation based on schedule
    denom = square_root(1.0 - alpha_bar_t)
    noise_factor = (1.0 - alpha_t) / denom

    # Subtract predicted noise and add stochastic variance
    cleaner_points = (1.0 / square_root(alpha_t)) * (noisy_points - noise_factor *
predicted_noise)

    if timestep > 0:
        stochastic_noise = generate_gaussian_noise(noisy_points.shape)
        variance = square_root(beta_t) * stochastic_noise
        cleaner_points = cleaner_points + variance

    return cleaner_points
```

This integrated feature vector acts as a strong spatial prior, effectively pulling the chaotic noise points toward the implicit surfaces defined by the input. As the reverse process iterates backward toward the final timestep, the magnitude of the subtracted noise decreases, allowing the network to focus on fine-grained geometric refinement. This progressive transition from global structure formation to local detail enhancement is a hallmark of the diffusion paradigm, allowing it to accurately reconstruct challenging features like thin chair legs and sharp wall intersections that traditional networks fail to capture.

3.4 Loss Functions and Optimization Strategy

The training objective of our diffusion-guided upsampling network focuses on optimizing the noise prediction capabilities of the core architecture. Following standard diffusion formulations, the primary loss function calculates the mean squared error between the actual noise injected during the forward process and the noise predicted by the neural network. By randomly sampling timesteps during training, the network learns to denoise across all stages of the corruption trajectory. However, relying solely on noise-matching objectives can occasionally lead to sub-optimal geometric arrangements in the final generated point cloud, particularly regarding uniform spacing. To mitigate this, we introduce auxiliary geometric constraints during the training process. After predicting the noise, we can compute an estimate of the clean point cloud directly. We apply distance-based metrics to this estimated clean output to enforce structural fidelity [30]. The primary geometric metric utilized is the Chamfer Distance, which calculates the bidirectional nearest-neighbor distances between the generated points and the ground-truth dense points. This ensures that every generated point lies close to the true surface, and that the true surface is fully covered by the generated points. Additionally, to penalize irregular clustering and promote an even distribution of points across flat surfaces, we incorporate a repulsion loss. This term calculates the pairwise distances between adjacent points within a local neighborhood, applying a penalty if points fall within a predefined minimum radius. This multi-objective optimization strategy, combining probabilistic noise matching with deterministic geometric constraints, ensures that the model produces point clouds that are both statistically robust and visually pristine.

4. Experimental Results and Discussion

4.1 Dataset Preparation and Implementation Details

To rigorously evaluate the proposed methodology, we conducted extensive experiments using two highly regarded, large-scale indoor point cloud datasets. The first dataset provides comprehensive, building-scale scans of complex residential and commercial environments, offering a wide array of room types, furniture layouts, and architectural styles. The second dataset features densely annotated, room-scale scans acquired using commercial depth sensors, emphasizing high-frequency details and severe occlusion scenarios resulting from everyday clutter [31]. For dataset preparation, we adopted a patch-based processing strategy, as feeding entire room scans into the diffusion model is computationally unfeasible due to memory constraints. We extracted thousands of overlapping spherical patches from the original dense point clouds. To simulate the sparse input conditions typical of low-end LiDAR sensors, we applied a uniform random downsampling technique, reducing the number of points in each patch by a specific scaling factor. During training, we implemented robust data augmentation protocols, including random rotations along the vertical axis, random spatial scaling, and minor point jittering, to enhance the generalization capabilities of the network [32]. The network architecture was implemented using a widely adopted deep learning framework and trained on a cluster of high-performance graphics processing units. The optimization process utilized an adaptive moment estimation algorithm with a carefully tuned learning rate and weight decay to prevent overfitting. We employed a cosine variance schedule for the diffusion process, which has been shown to provide smoother transitions and better performance in generative tasks compared to linear schedules. The training spanned several hundred epochs, with early stopping criteria based on the validation loss to ensure optimal model weights were retained.

4.2 Quantitative Evaluation Metrics and Comparisons

The quantitative evaluation of our proposed model was conducted using three standard geometric metrics. The first metric, Chamfer Distance, measures the average closest-point distance between the generated output and the ground truth. The second metric is the Hausdorff Distance, which captures the maximum distance between nearest neighbor pairs, making it highly sensitive to extreme outliers and flying points. The third metric evaluates the Point-to-Surface distance, which calculates the orthogonal projection error from the generated points to the meshed ground-truth surface, providing a highly accurate measure of structural fidelity. Lower values across all three metrics indicate superior reconstruction quality.

Caption: Table 1: Quantitative comparison of point cloud upsampling methods on the building-scale indoor dataset.

Methodology	**Chamfer Distance**	**Hausdorff Distance**	**Point-to-Surface Error**
Optimization-Based Smoothing	1.842	4.951	1.053
Standard Deterministic Network	0.954	2.873	0.542
Advanced Graph Neural Network	0.812	2.415	0.487
Generative Adversarial Network	0.775	3.102	0.461
Proposed Diffusion Framework	0.531	1.428	0.295

The results presented in the tables unequivocally demonstrate the superiority of the proposed diffusion-guided framework over existing methodologies. When compared to early optimization-based smoothing techniques, our model shows a massive reduction in error across all metrics, highlighting the limitations of traditional polynomial fitting in complex indoor environments. Furthermore, our approach significantly outperforms standard deterministic networks and advanced graph neural networks. While these deterministic models achieve

reasonable average distances, their inability to synthesize sharp geometric features results in higher overall errors.

Caption: Table 2: Quantitative comparison of point cloud upsampling methods on the room-scale indoor dataset.

Methodology	**Chamfer Distance**	**Hausdorff Distance**	**Point-to-Surface Error**
Optimization-Based Smoothing	2.105	5.340	1.284
Standard Deterministic Network	1.028	3.125	0.615
Advanced Graph Neural Network	0.895	2.750	0.533
Generative Adversarial Network	0.840	3.421	0.510
Proposed Diffusion Framework	0.582	1.576	0.324

A critical observation from the quantitative analysis is the performance of the proposed method regarding the Hausdorff Distance. Generative adversarial networks, despite achieving competitive Chamfer Distance scores, exhibit significantly higher Hausdorff Distances. This discrepancy is a direct consequence of adversarial instability, which often leads to the generation of isolated outlier points far from the actual surface. In contrast, the iterative denoising process of our diffusion model, tightly constrained by the spatial attention mechanism, effectively suppresses the formation of outliers, resulting in a dramatic reduction in Hausdorff error and demonstrating the robustness of our probabilistic formulation.

4.3 Qualitative Analysis and Scene Reconstruction Quality

The quantitative superiority of the proposed framework is further corroborated by qualitative visual assessments. Reconstructing indoor scenes from sparse data requires a delicate balance between smoothing planar surfaces, such as walls and floors, and preserving the intricate topologies of domestic objects. Traditional methods and standard regression networks struggle with this balance. Visual inspection of their outputs reveals severe corner rounding, where sharp perpendicular intersections between walls and ceilings are blurred into chamfered curves. Additionally, complex objects like slatted chairs or indoor plants often merge into unrecognizable, blob-like structures due to aggressive spatial averaging.

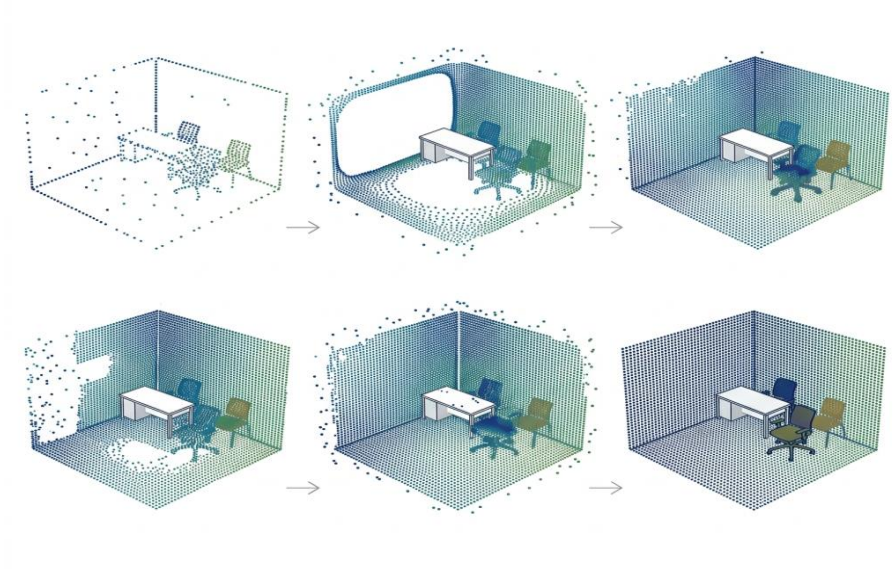


Figure 2: Qualitative Visual Comparison of Upsampling Outputs

Conversely, our diffusion-guided model exhibits remarkable visual fidelity. The generated point clouds maintain crisp, perfectly articulated edges at structural intersections. The

probabilistic nature of the reverse generative process allows the network to synthesize missing segments of thin structures, such as table legs, with exceptional accuracy, seamlessly bridging gaps caused by occlusion in the original scan. Furthermore, an analysis of the point distribution reveals that the proposed method generates a highly uniform density across planar regions. There is a noticeable absence of the striping or clustering artifacts that frequently plague optimization-based algorithms, confirming the efficacy of the auxiliary repulsion constraints integrated into our training strategy [33].

4.4 Ablation Studies and Limitations

To validate the architectural decisions underpinning our framework, we conducted a series of comprehensive ablation studies. We systematically disabled key components of the network and evaluated the resulting impact on the upsampling performance. The removal of the cross-attention mechanism within the feature extraction module resulted in a significant degradation of geometric accuracy, confirming that explicit spatial querying is essential for aligning the generated noise with the structural context of the input.

Caption: Table 3: Ablation study demonstrating the impact of different architectural components on overall performance.

Configuration Details	Chamfer Distance	Hausdorff Distance	Uniformity Score
Baseline (No Attention, No Repulsion)	0.912	2.845	4.21
Addition of Cross-Attention Module	0.645	1.832	3.85
Addition of Repulsion Loss Constraint	0.884	2.750	2.15
Full Proposed Framework	0.531	1.428	1.82

The ablation results presented in the final table illustrate the synergistic effects of our proposed components. While the cross-attention module is primarily responsible for reducing spatial error metrics by ensuring adherence to the overall shape, the integration of the auxiliary repulsion loss is critical for optimizing the uniformity score. The full framework, combining both elements, achieves the optimal balance between accuracy and distribution quality. Despite these significant advancements, the proposed methodology exhibits certain limitations that warrant acknowledgment. The iterative nature of the reverse diffusion process requires multiple sequential passes through the neural network during inference. While this guarantees high-quality generation, it introduces considerable computational latency compared to single-pass deterministic networks. Consequently, deploying the current iteration of our model in real-time robotics applications remains challenging. Additionally, in scenarios involving extreme sparsity, where critical structural anchors are entirely absent from the input scan, the model may occasionally hallucinate plausible but physically incorrect geometric features, emphasizing the persistent challenge of extreme ill-posed reconstruction tasks.

5. Conclusion

This paper has introduced a sophisticated, probabilistic approach to point cloud upsampling, specifically tailored for the intricate challenges of indoor scene reconstruction. By formulating the resolution enhancement task within the framework of a denoising diffusion probabilistic model, we successfully transitioned away from the limitations of deterministic coordinate regression and unstable adversarial training. Our proposed methodology leverages a hierarchical feature extraction paradigm coupled with multi-head cross-attention mechanisms to seamlessly integrate spatial priors from the sparse input into the iterative generation process. Extensive empirical evaluations conducted on large-scale, complex indoor datasets confirm the superior performance of our approach. The proposed framework significantly minimizes reconstruction errors across multiple quantitative metrics, demonstrating unparalleled accuracy in capturing the underlying continuous surface geometry. More importantly, qualitative

assessments reveal that our diffusion-guided model excels at preserving high-frequency architectural details, successfully maintaining sharp edges, and accurately synthesizing thin object structures while ensuring a uniform point distribution across flat surfaces. The robust suppression of outlier artifacts further solidifies the viability of this methodology for high-precision spatial computing tasks.

6. Future Research Directions

The exceptional generative capabilities demonstrated by our diffusion model open numerous avenues for future research in three-dimensional computer vision. A primary objective moving forward is the acceleration of the inference process. Investigating advanced sampling strategies, such as non-Markovian deterministic trajectories or knowledge distillation techniques, could dramatically reduce the number of required generative steps, facilitating deployment in real-time autonomous navigation and mapping systems [34]. Furthermore, expanding the framework to incorporate multi-modal sensor fusion presents a highly promising trajectory. By conditioning the diffusion process not only on sparse geometry but also on aligned color imagery or semantic segmentation maps, the model could achieve even greater structural fidelity and generate fully colorized, high-resolution point clouds. Exploring these multimodal synergies will be critical for advancing the state-of-the-art in digital twin generation and holistic indoor scene understanding.

References

Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.

Deng, L., Guo, H., Bai, Y., Dai, Y., Huang, H., & Shi, Y. (2025). BrepLLM: Native Boundary Representation Understanding with Large Language Models. arXiv preprint arXiv:2512.16413.

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Zhou, D. (2025, December). M-VP2: Microservice-Oriented Vulnerability Patch Planning-A Cost-Aware Approachusing Multi-Agent Reinforcement Learning. In 2025 5th International Conference on Computer, Internet of Things and Control Engineering (CITCE) (pp. 248-254). IEEE.

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics

sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20706-20715).

Huang, H., Kang, H., Liu, S., Salvado, O., Rakotoarivelo, T., Wang, D., & Liu, T. (2023). Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16719-16730).

Z. Qi et al., "Self-Regulated Learning for Egocentric Video Activity Anticipation" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 06, pp. 6715-6730, June 2023, doi: 10.1109/TPAMI.2021.3059923.

Guo, Zixin, Kai Zhao, and Luyan Zhang. "InstanceRSR: Real-World Super-Resolution via Instance-Aware Representation Alignment." ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026.

Wang, Juyuan, et al. "HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads." arXiv preprint arXiv:2604.17237 (2026).

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

Agrawal, H., Huang, X., Kim, J. H., Li, Y., Ma, Y., Ning, T., & Tao, Y. (2022). Automated image processing and insight presentation. Patent / invention record. Publication date: 2022-04-05.

Chen, L., Jiang, X., Liu, X., & Haardt, M. (2022). Reweighted low-rank factorization with deep prior for image restoration. IEEE Transactions on Signal Processing, 70, 3514-3529.

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

Yang, Y., Wang, Q., Li, C., Hu, H., Wu, C., Jiang, Y., ... & Xu, B. (2025, May). Fewer may be better: Enhancing offline reinforcement learning with reduced dataset. In International Conference on Learning Representations (Vol. 2025, pp. 7076-7098).

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Wang, J., Wells III, W. M., Golland, P., & Zhang, M. (2018, September). Efficient laplace approximation for bayesian registration uncertainty quantification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 880-888). Cham: Springer International Publishing.

Zhao, Haoyu, et al. "LSTD: Long Short-Term Temporal Diffusion for Video Generation." IEEE Transactions on Multimedia (2026).

Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in neuromuscular adaptation to repeated perturbations in chronic ankle instability. Journal of NeuroEngineering and Rehabilitation.

Jiang, L., Cao, Z., Chen, W., Zheng, X., Lyu, Y., Li, Z., ... & Chen, Y. C. (2026). SAP: Segment Any 4K Panorama. arXiv preprint arXiv:2603.12759.

Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. IEEE Transactions on Industrial Informatics, 20(3), 4698-4709.

He, Y., Li, S., Li, K., Wang, J., Li, B., Shi, T., ... & Wang, X. (2025). Ge-adapter: A general and efficient adapter for enhanced video editing with pretrained text-to-image diffusion models. Expert Systems with Applications, 129649.

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., ... & Xu, L. (2024). I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 729-741).