PatchScene: Patch-based Voxel Diffusion for Large-Scale Scene Completion

Qingdong Xu[1,*] Jiajun Zhu[1,2,*] Shilin Zhu[4,*]
Xinjing He[5] Chao Lu[2] Huanran Wang[2] Jiyao Zhang[3,†]

[1]MEGVII Technology [2]Qianli Technology [3]Peking University
[4]Northeastern University, China [5]Northwest Polytechnical University, Xi'an
{2301010,2301062}@stu.neu.edu.cn mr.zhujiajun@gmail.com jiyaozhang@stu.pku.edu.cn
[*]Equal contribution [†]Corresponding author

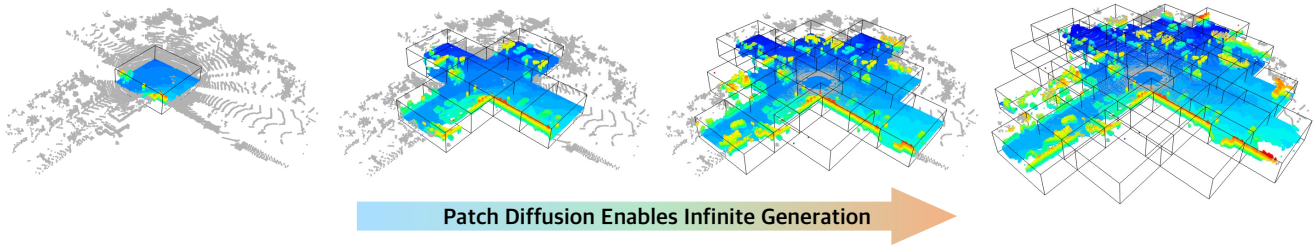

Figure 1. We introduce PatchScene, a novel diffusion framework based on the divide-and-conquer paradigm, which directly generates point clouds in explicit voxel space. It achieves temporally consistent, high-fidelity, and spatially infinite scene completion through a cycle of patch-based completion, fusion, and diffusion.

Abstract

We propose PatchScene, a novel diffusion-based framework for large-scale LiDAR scene completion. Unlike existing methods that rely on global latent representations or dense voxel grids, PatchScene adopts a patch-based voxel diffusion paradigm that explicitly generates fine-grained geometry within localized 3D regions. To ensure coherent reconstruction at both spatial and temporal scales, we introduce a confidence-guided spatio-temporal fusion mechanism that integrates overlapping patches and adjacent frames in a unified generative process. Furthermore, we design an Annular-Flow diffusion strategy that leverages the radial density pattern of LiDAR scans to progressively propagate high-fidelity information from near-range to far-range regions, enabling spatially unbounded scene completion. Extensive experiments on the SemanticKITTI benchmark demonstrate that PatchScene achieves state-of-the-art performance across all standard metrics, surpassing previous approaches in both geometric accuracy and temporal consistency. Remarkably, the model trained on 20 m LiDAR ranges generalizes effectively to 50 m scenes without retraining, highlighting its strong scalability and generalization capability for real-world autonomous driving applications.

1. Introduction

LiDAR is a pivotal sensor in autonomous driving and robotics, providing high-fidelity 3D geometric structures and absolute scale information, which are crucial for robust scene understanding [7, 16]. However, the inherent scanning patterns of LiDAR result in spatially non-uniform point distributions, characterized by density decreasing rapidly with distance [31, 34]. Furthermore, pervasive occlusions inherent to line-of-sight physical measurement prevent the capture of surfaces in occluded regions. These limitations significantly hinder the application of LiDAR point clouds in crucial tasks such as environmental perception, navigation planning, and scene reconstruction for autonomous vehicles [15, 42]. In contrast, dense point clouds provide richer geometric details and sharper object boundaries, thereby markedly improving the accuracy of downstream algorithms like object detection, tracking, and semantic segmentation [27, 28, 36]. They serve as a fundamental prerequisite for safe navigation and high-precision localization. Nevertheless, sparse-to-dense scene completion remains a formidable challenge in the 3D vision domain, primarily due to the vast scale variations and the irregular, unordered nature of point cloud data in outdoor environments [17, 18, 35].

A primary challenge in point cloud completion for au-

tonomous driving stems from the large-scale nature of the scenes [22, 33]. This makes achieving efficient generation while maintaining high geometric fidelity exceptionally difficult. To address these challenges, methods such as [2, 20, 38] employ techniques like sparse convolution to mitigate the substantial computational overhead. However, the irregular nature of point cloud representations, combined with the difficulty of predicting precise coordinate offsets in continuous 3D space, often results in completed surfaces plagued by artifacts or excessive smoothing, failing to restore sharp geometric structures. XCube [23] attempts to solve the large-scale prediction problem by compressing voxels into a low-dimensional latent space and performing scene completion via latent diffusion, followed by multi-stage refinement. Nonetheless, this multi-stage encoding-decoding process inevitably incurs information loss and error accumulation, leading to degraded geometric detail in the reconstruction [11]. A second significant challenge lies in guaranteeing temporal consistency, which is crucial for maintaining continuous target trajectories and smooth geometric surfaces in dynamic environments. Yet, prevailing methods are confined to independent, single-frame completion, thereby disregarding the inherent temporal correlations within the sensor data sequence [37, 41].

To overcome these limitations, we propose PatchScene, a novel diffusion framework built upon a divide-and-conquer paradigm. It achieves temporally consistent, high-fidelity, and spatially infinite scene completion through a cycle of patch-based completion, fusion, and diffusion, as illustrated in Fig. 1. Specifically, we first discretize the full scene and perform diffusion directly within the explicit space of local voxel patches. We introduce an Annular-Flow Patch Diffusion strategy, which guides generation by progressively diffusing outwards in an annular pattern. This effectively propagates dense information from inner, well-observed regions to sparse, distant locations, enabling high-precision completion and extrapolation to infinite spatial domains. Furthermore, we propose a confidence-guided spatio-temporal patch fusion mechanism to enhance inter-frame consistency, realizing coherent and seamless scene completion across temporal sequences. Experimental results demonstrate that our method surpasses previous approaches on the SemanticKITTI dataset [1], achieving state-of-the-art (SOTA) performance. The model's ability to train on 20m data and generalize effectively to a 50m range further validates its flexible spatial scalability. Finally, the integration of temporal information allows our method to significantly suppress the inter-frame flickering artifacts common in single-frame approaches, resulting in a substantial quality improvement.

In summary, our contributions are as follows:

- We propose PatchScene, a novel framework that employs diffusion model to explicitly generate fine-grained point clouds in voxel space, flexibly extending to infinite spatial ranges.
- We introduce a novel Patch Spatio-temporal Fusion method that ensures holistic continuity and geometric consistency across both spatial and temporal dimensions of the completed point clouds.
- PatchScene achieves state-of-the-art results on the SemanticKITTI benchmark, setting a new standard for large-scale scene level point cloud completion.

2. Related Works

2.1. Discriminative Models for Scene Completion

Early scene completion methods were mostly based on discriminative models, which learn a mapping from partial observations to complete 3D voxel occupancy or Signed Distance Fields (SDFs). These approaches typically represent scenes using regular voxel grids, converting the completion process into a dense 3D prediction problem. SDF-based methods model the scene as a signed distance field, where each voxel encodes the shortest distance to the nearest surface along with its sign. By regressing continuous distance values, such methods can produce smooth and structured geometric completions even under sparse input conditions [6, 21]. Another mainstream line of work directly predicts the occupancy state (occupied / free) and semantic labels of each voxel in the 3D space, leveraging 3D convolutional networks to extract contextual features and perform end-to-end voxel occupancy prediction, thereby achieving scene-level semantic completion [24–26, 29]. Subsequent research introduced Transformer architectures to enhance feature extraction and fusion capabilities, improving global consistency and representation quality [8, 30]. However, all of these methods essentially follow a one-shot prediction mechanism, where optimization relies solely on regression losses. As a result, they struggle to capture uncertainty and fail to model the inherent geometric diversity of complex scenes in a generative manner.

2.2. Diffusion Models for 3D Generation

Given the remarkable success of diffusion models in image generation, several recent studies have explored their extension to 3D point cloud generation. Some works represent LiDAR point clouds as range images, adopting image-generation paradigms to produce diverse and upsampled point clouds [19, 22, 41]. However, due to inherent occlusion in range representations, such approaches fail to reconstruct complete 3D scenes. To enable scene completion, subsequent works have performed diffusion directly in 3D space, which can be broadly categorized into three types: point-based, voxel/SDF-based, and latent-based methods.

Point-based methods directly apply diffusion models in the raw point cloud space by predicting offset vectors for

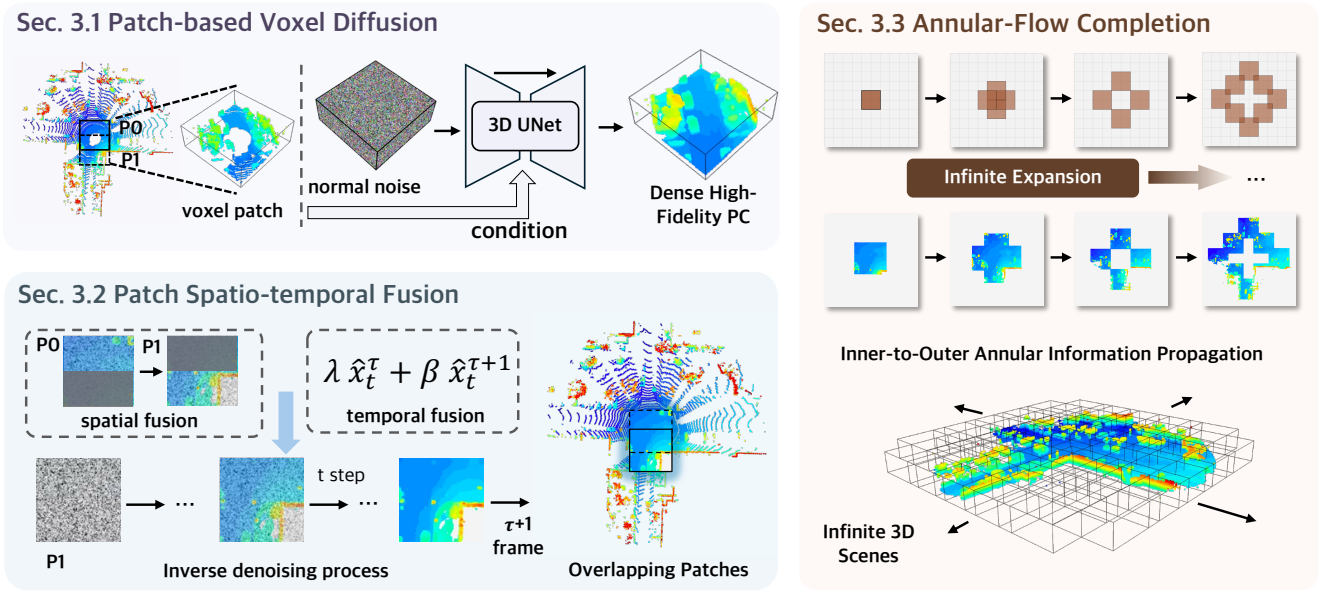

Figure 2. Overview of PatchScene. The voxel space is first divided into overlapping local patches, each processed independently through diffusion-based denoising to generate local point clouds. Spatial and temporal fusion then merges these patches into a coherent global point cloud. Finally, an annular outward diffusion strategy extends completion across the entire scene, handling near-dense and far-sparse LiDAR distributions for large-scale, temporally consistent reconstruction.

noisy points, enabling randomly sampled noise points to move toward missing regions of the scene and thus achieve completion within the point set domain [2, 14, 20]. Building upon this paradigm, model distillation is introduced in ScoreLiDAR to accelerate and optimize the generative process [38]. Although operating directly in point space preserves the original geometric structure without voxelization artifacts, the irregular nature of point clouds often leads to unstructured or uneven generation results. Consequently, these methods tend to exhibit structural inconsistency, with holes and oversmoothed fine details in the completed scenes. In voxel/SDF-based approaches, some works primarily applied to single objects or indoor scenes, employing cascaded diffusion models to achieve high-resolution point cloud generation [9, 10, 39]. For outdoor scenes, self-supervised methods like MID recover geometric consistency from sparse LiDAR scans without semantic labels [31]. However, the memory and computational footprint of explicit 3D grid representations grows cubically, severely limiting the scalability of these methods when applied to large-scale outdoor environments.

For large-scale and high-resolution data generation, latent-based approaches provide a common and effective solution. These methods typically employ a multi-stage VAE framework to encode raw point clouds into a lower-dimensional latent space, where diffusion-based generation is then performed. This strategy enables the synthesis of large-scale and highly dense point clouds [4, 23]. However, such systems are often complex, requiring cascaded VAE training, which is computationally expensive, difficult to optimize, and inherently lossy due to compression.

3. Method

To address the challenges of large-scale scenes and temporal multi-view continuous consistency in autonomous driving point cloud completion, we propose PatchScene, which divides the voxel space for local patch generation, followed by spatio-temporal global fusion, and an Annular-Flow Diffusion Completion strategy. The overall architecture is illustrated in the Fig. 2. In Sec. 3.1, we introduce the partitioning of the complete voxel space into mutually overlapping regular patches and independently executing a diffusion denoising process to generate local point clouds. Subsequently, in Sec. 3.2, we detail our spatial overlap fusion and temporal consistency fusion. These mechanisms simultaneously merge the disparate patches across both temporal and spatial dimensions into a detail-rich, complete, and unified global dense point cloud. Finally, in Sec. 3.3, leveraging the physical characteristic of LiDAR scan points of being dense at near ranges and sparse afar, we use an annular, outward progressive diffusion generation method to extend the point cloud completion to infinite spatial domains.

3.1. Patch-based Voxel Diffusion

Mapping point cloud space directly into a discrete, regular voxel space significantly simplifies the training process

and yields more uniform and accurate point cloud completions. However, high-resolution voxelization leads to a cubic growth in computational cost, rendering traditional voxel-based methods infeasible for large-scale scene-level completion. To address this, we partition the entire voxel space using a set of overlapping patches with predefined dimensions (height, width, and depth). We then apply a diffusion model independently to each patch for denoising and completion [5, 13]. This strategy effectively reduces the task to object-level completion per patch, significantly mitigating the computational and memory burdens. Consequently, we can utilize a high voxel resolution for individual patches, which are subsequently fused spatio-temporally to reconstruct a complete, unified, and high-precision large-scale scene-level point cloud. The diffusion denoising process for our patch-based voxel completion is detailed as follows:

Voxel Patching. We represent a complete, dense occupancy frame as $\mathbf{X}_0 \in \{0,1\}^{H\times W\times D}$, and its corresponding sparse observation as $\widetilde{\mathbf{X}}$. For the k-th patch, we extract a local patch from the global voxel space using predefined dimensions (h, w, d) and a starting index (i_k, j_k, l_k). This patching operation is denoted as:

$$\mathbf{x}_0^{(k)} = \text{Patch}(\mathbf{X}_0, k), \quad \tilde{\mathbf{x}}^{(k)} = \text{Patch}(\tilde{\mathbf{X}}, k) \tag{1}$$

where $\mathbf{x}_0^{(k)}, \tilde{\mathbf{x}}^{(k)} \in \{0,1\}^{h\times w\times d}$. Here, $\text{Patch}(\mathbf{X}, k)$ signifies the k-th local patch extracted from the global voxel $\mathbf{X}$. The patches are extracted with a specific stride to ensure sufficient spatial overlap, resulting in a set of N patches that collectively cover the entire scene.

Forward Diffusion. During the forward process in the training phase, noise is independently added to each local patch $\mathbf{x}_0^k$ [3]:

$$\mathbf{x}_t^k = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^k + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad t = 1, \ldots, T \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ are predefined noise schedule coefficients.

Inverse Denoising. In the inverse denoising process, we train a deep neural network $f_\theta(\mathbf{x}_t^k, t, \mathbf{p}_k)$ to predict the original data $\mathbf{x}_0$ directly, which enhances the stability and quality of the generated samples. The network accurately recovers the estimated noise-free data $\hat{\mathbf{x}}_0^k$ from the current noisy patch $\mathbf{x}_t^k$ and is conditioned on a learnable position encoding $\mathbf{p}_k$ to ensure the process is sensitive to spatial context.

$$\hat{\mathbf{x}}_0^k = f_\theta(\mathbf{x}_t^k, t, \mathbf{p}_k) \tag{3}$$

Based on the closed-form solution of the forward diffusion process Eq. 2, we substitute the network's prediction $\hat{\mathbf{x}}_0^k$ to derive the corresponding predicted noise $\hat{\epsilon}$:

$$\hat{\epsilon} = \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(\mathbf{x}_t^k - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0^k\right) \tag{4}$$

Subsequently, using this predicted noise $\hat{\epsilon}$, we can compute the sample for the previous timestep:

$$\mathbf{x}_{t-1}^k = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t^k - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}\right) + \sigma_t\mathbf{z} \tag{5}$$

where $\alpha_t = 1 - \beta_t$.

Training Objective. During the training process, for each ground-truth sample, we randomly select a patch k and a timestep t. The model f_θ is then used to predict the corresponding scene patch $\hat{\mathbf{x}}_0^k$ from the noisy input $\mathbf{x}_t^k$. We optimize the model by minimizing the Mean Squared Error (MSE) between the prediction $\hat{\mathbf{x}}_0^k$ and the ground-truth occupancy voxel $\mathbf{x}_0^k$:

$$\mathcal{L}_{\text{patch}}(\theta) = \mathbb{E}_{k,t}\left[||\mathbf{x}_0^k - \hat{\mathbf{x}}_0^k||_2^2\right] \tag{6}$$

3.2. Patch Spatio-temporal Fusion

Spatial Fusion. Performing the inverse denoising process independently on all local patches can lead to boundary discontinuities and noticeable artifacts in the overlapping regions, thereby compromising the global consistency of the scene. To resolve this issue, we introduce a stochastic interference spatial fusion mechanism during the inverse diffusion process. This ensures that the denoising outcomes in the overlapping areas of adjacent patches remain smooth and semantically coherent. After obtaining the predicted noise $\hat{\boldsymbol{\epsilon}}^k$ for the current patch at step t according to Eq. 4, we project and aggregate the noise predictions from all adjacent patches into a global noise field $\hat{\boldsymbol{\epsilon}}^{\text{global}}$. The fusion operation is exclusively performed within the overlapping regions $\mathcal{O}_k$ between patches. For any point $p \in \mathcal{O}_k$ within an overlap zone, the fused noise for the next sampling step, $\hat{\boldsymbol{\epsilon}}_{\text{fused}}^k$, is generated through the following probabilistic weighting scheme:

$$\hat{\boldsymbol{\epsilon}}_{\text{fused}}^k(p) = B(p)\cdot\hat{\boldsymbol{\epsilon}}^{\text{global}}(p) + (1-B(p))\cdot\hat{\boldsymbol{\epsilon}}^k(p) \tag{7}$$

where $B \in \{0,1\}$ is a binary random mask, with each element $B(p)$ drawn from an independent Bernoulli distribution, $B(p) \sim \mathcal{B}(0.5)$. This mechanism mandates that, within the overlapping regions, the noise of the current patch has a 50% probability of adopting the estimate from the global context. Such random coupling effectively integrates local denoising information with the global semantic context, preventing the blurring effects often associated with deterministic fusion methods. This ensures the generation of denoised results with strong spatial coherence in the subsequent inverse step.

Temporal Fusion. While spatial fusion ensures intra-frame coherence, we also aim to fuse information across frames to achieve more natural and geometry-preserving point cloud

completions. Consecutive frames are temporally continuous, meaning their observed point clouds are visually coherent and maintain semantic consistency, albeit with slight viewpoint variations. Leveraging the completed dense point cloud from the previous frame can enhance the current frame's understanding of scene semantics and improve the model's implicit grasp of viewpoint information, thereby refining the completion to be detail-rich. During the denoising generation process for frame τ, we cache its prediction at each denoising step. For the subsequent frame $\tau+1$, we first perform ICP [32] registration between the sparse observations $\widetilde{\mathbf{X}}^{\tau}$ and $\widetilde{\mathbf{X}}^{\tau+1}$ to obtain the rigid transformation matrix $\mathbf{T}_{\tau\to\tau+1}$, which is used to transform the cached noise into the new coordinate system. When generating the denoised result for frame $\tau+1$ at step t, the model relies not only on the observation $\widetilde{\mathbf{X}}^{\tau+1}$ but also incorporates guidance from the preceding frame. This temporal fusion is formulated as:

$$\hat{\mathbf{x}}_t^{\tau+1} = \lambda \cdot \hat{\mathbf{x}}_t^{\tau} + (1-\lambda)\cdot \hat{\mathbf{x}}_t^{\tau+1} \tag{8}$$

Here, $\lambda \in [0,1]$ is the cross-frame guidance scale. The value of λ is adaptively determined on the Bird's-Eye View (BEV) voxel grid based on local density consistency, dynamically adjusting the weighting between the two consecutive frames.

$$\lambda(p) = \min\left(\frac{\rho^{\tau+1}(p)}{\rho^{\tau}(p)+\epsilon}, 1.0\right), \quad p \in \mathcal{V} \tag{9}$$

where $\rho(p)$ denotes the local point density at voxel position p, and $\mathcal{V}$ represents the set of all voxels in the BEV grid. This design allows the model to inherit more information from the previous frame in geometrically consistent regions while relying more on the current frame's prediction in areas with structural changes or sparsity. This achieves effective temporal fusion across frames in the BEV space.

3.3. Annular-Flow Diffusion Completion

Given the overlapping nature of our patches, the generation order becomes a critical factor, especially when leveraging completed patches to guide the generation of their neighbors. We observe that 3D LiDAR scenes exhibit a significant physical characteristic: point cloud density is high in proximity to the sensor and becomes progressively sparser with increasing radial distance. Consequently, voxel regions near the sensor are densely populated and relatively complete, whereas distant regions are sparse and often contain large areas of missing data. Since our completion method is conditioned on these observations, the quality of the generated output is inherently higher in the central, denser regions. Therefore, leveraging this physical property, we devise a completion strategy where patches with high-quality completions in the inner regions are used to progressively guide the generation process outwards via spatial fusion, continuing until all patches cover the entire scene. This approach ensures a uniformly dense and globally consistent completion.

Within the scene's voxel space $\mathbf{X}$, we group the patches into a series of concentric annular regions $\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_L\}$ based on their distance from the sensor's center. Here, $\mathcal{R}_1$ corresponds to the high-density region near the LiDAR, while $\mathcal{R}_L$ represents the distant, sparse region. Our proposed Annular-Flow process begins with the innermost ring $\mathcal{R}_1$ and proceeds outwards. For each region $\mathcal{R}_\ell$, we perform patch-based conditional diffusion sampling to obtain local completions $\hat{\mathbf{x}}_0^k$. Critically, the denoising process for patches in $\mathcal{R}_\ell$ is guided by the already completed results from the adjacent inner ring $\mathcal{R}_{\ell-1}$ at every timestep t. This center-outward guided generation allows high-fidelity information to continuously flow from the core to the periphery, enabling the sparse outer regions to leverage the rich semantic context of the inner scene. This center-outward flow endows our method with the inherent capability to extend to unbounded scenes, in principle achieving completion for infinite-scale environments.

4. Experiment

4.1. Setup

We conduct training and evaluation on the SemanticKITTI dataset, where the LiDAR scanning range is set to 50 m, consistent with previous scene-level methods. The voxel resolution is 0.15625 m. During training, we initialize the learning rate at 4×10^{-4} and employ the AdamW optimizer [12] for 100 epochs. For the diffusion process, a cosine noise schedule is applied with $\beta_{10} = 0.0001$ and $\beta_T = 0.02$, using $T = 1000$ time steps. When dividing the scene into patches, each patch covers a 20 m $\times$ 20 m area. After the denoising generation process, we upsample the completed results by a factor of two, yielding approximately 900,000 points per frame in the final point cloud.

4.2. Results

Table 1 provides a comprehensive quantitative comparison between our approach and previous state-of-the-art point cloud completion methods. For fair comparison, all methods are evaluated using the same metric protocols with the original point clouds as ground truth (GT). Our method also does not use any additional temporal information and relies solely on single-frame completion. The evaluation includes Chamfer Distance (CD), Jensen–Shannon Divergence measured in both BEV and full 3D space (JSD-BEV and JSD-3D), and voxel IoU. CD captures the average nearest-neighbor discrepancy between two point sets and reflects local geometric accuracy. JSD-BEV and JSD-3D measure the similarity between global spatial distributions in 2D and 3D, indicating structural plausibility. Voxel IoU

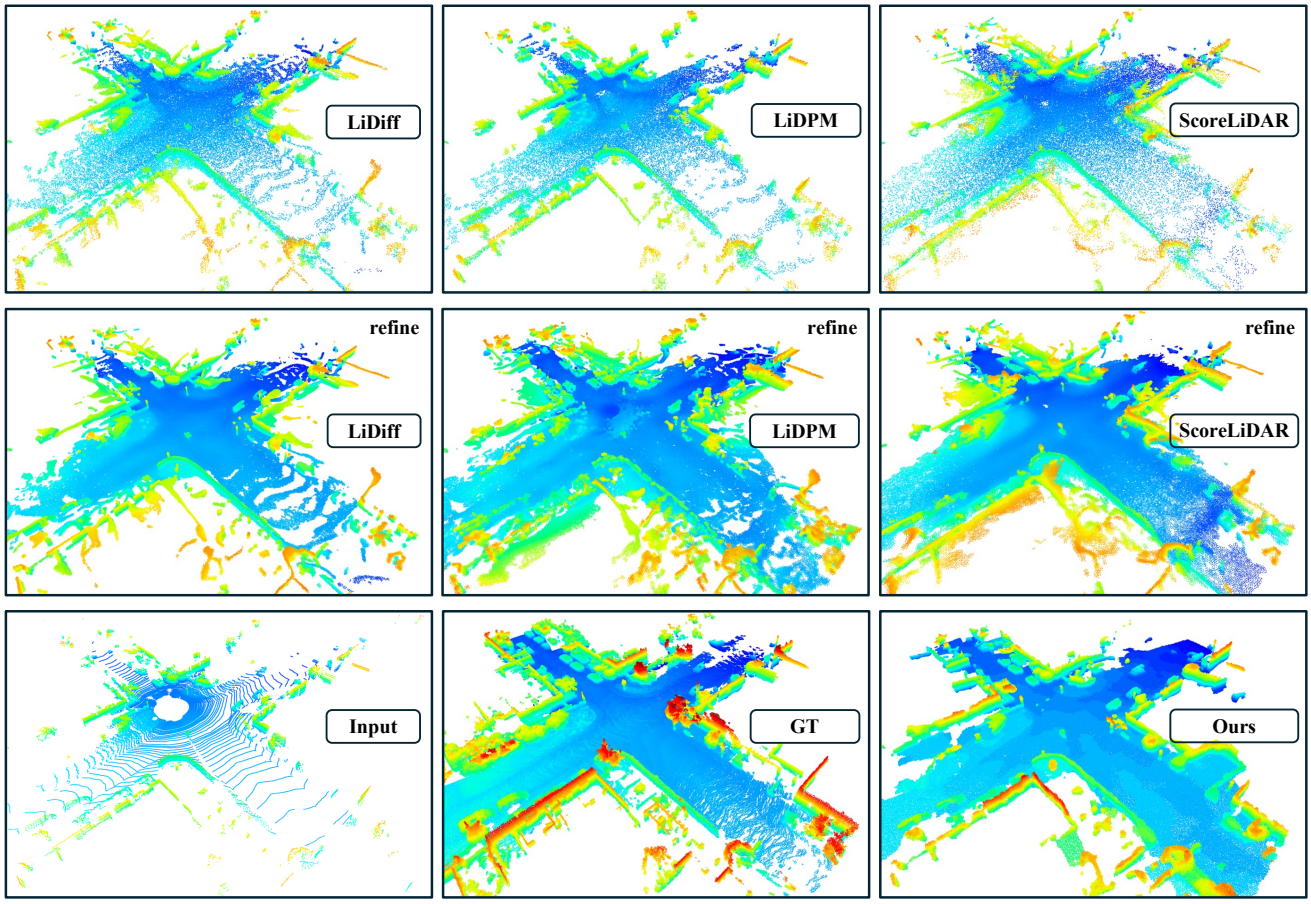

Figure 3. Comparison of our method with LiDiff, LiDPM, and ScoreLiDAR on the same scene in SemanticKITTI. All point clouds are height-normalized and color-mapped within the same range. The completion results produced by LiDiff, LiDPM, and ScoreLiDAR remain relatively sparse, and although refinement yields denser predictions, it introduces holes and inconsistent hallucinated points. In contrast, our PatchScene achieves completion results that are closest to the ground truth.

evaluates volumetric consistency and large-scale completeness. Based on these complementary metrics, our method demonstrates clear and consistent improvements over all previous approaches, showing its ability to recover fine-grained local geometry while preserving coherent global structure and geometric integrity.

To further examine the qualitative behavior of different methods, Fig. 3 illustrates representative visual comparisons on the SemanticKITTI dataset. LiDiff directly predicts pointwise 3D offsets on raw inputs, which often results in noticeable holes in regions with extremely sparse measurements and produces ambiguous object shapes (e.g., vehicles) due to limited semantic awareness. LiDPM benefits from its global diffusion design, which helps preserve high-level semantic completeness. However, the resulting completions are still coarse near sparse boundaries. ScoreLiDAR generally produces more uniformly distributed completions but frequently introduces hallucinated structures, particularly around object edges, leading to deviations from physically plausible geometry. In contrast, our approach integrates a patch-based voxel diffusion framework with Spatio-Temporal Fusion, enabling more reliable scene-level semantic inference and improved multi-view consistency across temporal sequences. As a result, the generated completions contain substantially fewer scattered noisy points, exhibit stronger structural coherence across both object-level and scene-level regions, and avoid the boundary distortions observed in prior methods. Collectively, these improvements allow our method to achieve clearly superior completion quality in terms of both global scene geometry and local geometric detail.

4.3. Analysis of 3D Scene Generation

Temporal Consistency. To assess temporal continuity, we compute the bidirectional RMSE between adjacent frames. This evaluation captures temporal asymmetries from viewpoint changes or dynamic objects, offering a comprehensive understanding of temporal stability. Table. 2 presents Patch-

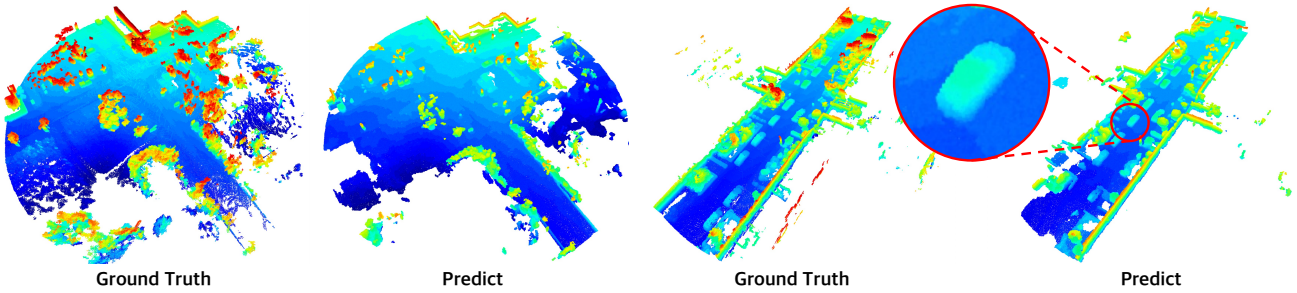

Figure 4. We train PatchScene on scenes with a LiDAR sensing range of 20 meters and directly apply it to point cloud completion with an extended range of 50 meters. Whether in open environments or narrow spaces, our completed point clouds consistently maintain high geometric fidelity, accurately preserve object boundaries, and simultaneously ensure strong global scene coherence.

Method	CD↓	JSD 3D↓	JSD BEV↓	Voxel IoU↑		
				0.5	0.2	0.1
LMSCNet [25]	0.641	-	0.431	30.8	12.1	3.7
LODE [6]	1.029	-	0.451	33.8	16.4	5.0
MID [31]	0.503	-	0.470	31.6	22.7	13.1
PVD [40]	1.256	-	0.498	15.9	4.0	0.6
LiDiff [20]	0.434	0.564	0.444	31.5	16.8	4.7
LiDiff(refine)	0.376	0.573	0.416	32.4	23.0	13.4
LiDPM [14]	0.446	0.532	0.440	34.1	19.5	6.3
LiDPM(refine)	0.377	0.542	0.403	36.6	25.8	14.9
ScoreLidar† [38]	0.412	0.589	0.425	29.7	13.0	3.6
ScoreLidar(refine)†	0.342	0.590	0.399	32.0	19.9	9.4
Ours	**0.319**	**0.444**	**0.371**	**45.3**	**38.2**	**19.7**

Table 1. Comparison of our method with existing approaches on SemanticKITTI. Baselines, metrics, and ground truth are from LiDPM, with results marked † independently reproduced and evaluated.

Scene's performance with and without temporal fusion. The results show that temporal fusion significantly reduces both forward and backward RMSE, indicating markedly improved inter-frame coherence and smoother transitions. Crucially, this enhancement in temporal consistency is achieved while largely preserving geometric accuracy. Although there is a marginal increase in JSD BEV, the structural degradation is negligible, with CD and JSD 3D maintain competitive performance. These findings collectively demonstrate that propagating multi-view and temporal information across frames successfully enhances global temporal stability without compromising the overall fidelity of the completed point clouds.

Method	CD↓	JSD 3D↓	JSD BEV↓	RMSE↓	
				t_0 to t_1	t_1 to t_0
w/o temporal	0.319	0.444	0.371	0.155	0.159
temporal fusion	0.309	0.432	0.372	0.086	0.081

Table 2. Analyzing Temporal Consistency with the RMSE Metric

Infinite-Space Generation. To verify the scalability of our method in infinite spatial extension, we train PatchScene on scenes with a LiDAR range of 20 m and directly apply it to point cloud completion with a 50 m range. As illustrated in Fig. 4, we showcase both open and narrow scenes to demonstrate the completion performance of PatchScene under different real-world conditions. In open scenes, the global structure of the completion results produced by PatchScene remains highly consistent with the ground truth. Remarkably, in sparse regions at the right boundary, our method even generates more continuous and complete structures than the ground truth, revealing its strong potential for maintaining global consistency. In narrow scenes, the method effectively completes vehicle point clouds with clear and well-defined shape boundaries, demonstrating that the local diffusion processes, combined with spatial–temporal fusion, preserve both fine-grained details and overall structural integrity. These results indicate that PatchScene can generalize learned scene priors to larger-scale, unseen environments while maintaining reasonable structural accuracy.

Denoising Timestep. We observe that the number of denoising timesteps has a noticeable impact on point cloud completion performance. As shown in Table 3, we report the completion metrics under different timestep settings. Interestingly, increasing the number of diffusion steps beyond a moderate value does not necessarily improve performance, since overly large timesteps can degrade quantitative metrics by introducing excessive stochasticity that pushes generated points away from the conditioning inputs. Conversely, using fewer steps can yield slightly higher scores on some metrics, but it limits the effectiveness of patch-wise fusion, leaving visible boundaries between adjacent patches, as illustrated in Fig. 5. To balance these effects, we adopt 10 timesteps as it provides a favorable trade-off between completion accuracy and inter-patch consistency, achieving the best overall performance across metrics while producing visually coherent reconstructions.

Method	CD↓	JSD 3D↓	JSD BEV↓	Voxel IoU↑		
				0.5	0.2	0.1
timestep = 5	0.310	0.437	0.371	45.7	38.9	20.3
timestep = 10	0.319	0.444	0.371	45.4	38.2	19.7
timestep = 20	0.334	0.450	0.376	44.6	37.6	19.4
timestep = 30	0.349	0.453	0.378	44.3	37.3	19.2
timestep = 50	0.357	0.456	0.380	44.1	37.1	19.1

Table 3. Analyzing the Impact of Denoising Timesteps on Point Cloud Completion Accuracy

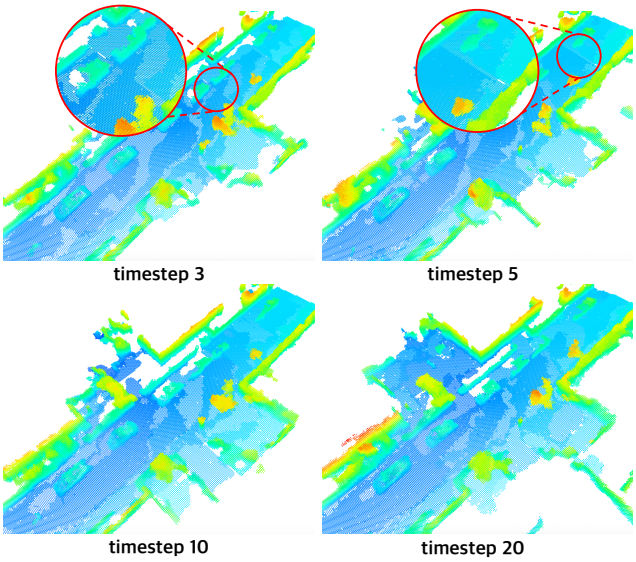

Figure 5. The effect of the denoising timestep on completion performance. Less timesteps lead to visible boundaries in the completed scenes, while larger timesteps enable more fusion iterations and yield stronger inter-patch consistency.

4.4. Ablation Study

Spatial Fusion. To validate the necessity and effectiveness of the spatial fusion design in PatchScene, we conduct a comprehensive ablation study in Table 4. Specifically, we evaluate four distinct variants for handling overlapping regions: (i) a naive baseline without spatial fusion; (ii) direct averaging; (iii) weighted averaging prioritizing inner-layer voxels; and (iv) our proposed random coupling strategy. Empirical results reveal that the absence of fusion or the use of deterministic averaging often leads to suboptimal transitions or over-smoothed representations. In contrast, our random sampling strategy, which randomly retains 50% of overlapping voxels, seamlessly aligns with the underlying statistical properties of the diffusion model. Consequently, it yields the best performance on CD and JSD-BEV metrics while maintaining highly competitive results on JSD-3D, demonstrating its robustness in synthesizing high-fidelity and globally consistent scene completions.

Generation Direction. We further investigate the impact of different diffusion directions for patch fusion, as shown in Table 5. We compare the completion performance under three diffusion strategies: Annular inward, Annular outward, and Linear diffusion, evaluated using CD, JSD-BEV, and JSD-3D metrics. The Linear Diffusion strategy follows a left-to-right and top-to-bottom propagation manner, which may be intuitive for 2D images but does not align with the radial symmetry characteristic of LiDAR scans, resulting in the poorest performance. Both Annular inward and Annular outward strategies consider the ring-shaped scanning pattern of LiDAR; however, the Annular outward diffusion yields better results, demonstrating the effectiveness of propagating high-fidelity information from the dense central region toward the sparse peripheral areas.

Method	CD↓	JSD 3D↓	JSD BEV↓	Voxel IoU↑		
				0.5	0.2	0.1
w/o fusion	0.348	0.451	0.383	43.9	37.4	19.7
average addition	0.351	0.439	0.381	44.6	38.5	20.3
weight addition	0.345	0.438	0.379	44.9	38.6	20.3
random coupling	0.319	0.444	0.371	45.3	38.2	19.7

Table 4. Ablation of Spatial Fusion

Method	CD↓	JSD 3D↓	JSD BEV↓	Voxel IoU↑		
				0.5	0.2	0.1
linear diffusion	0.451	0.528	0.399	38.2	28.7	13.0
Annular inward	0.391	0.461	0.389	43.3	36.7	19.0
Annular outward	0.319	0.444	0.371	45.3	38.2	19.7

Table 5. Ablation of Generation Direction

5. Conclusion

In this work, we introduced PatchScene, a diffusion-based framework that addresses the long-standing challenges of large-scale LiDAR scene completion—namely, the trade-off between geometric fidelity, temporal coherence, and computational efficiency. By decomposing the global voxel space into overlapping local patches, PatchScene enables high-resolution completion within a tractable computational budget. The proposed spatio-temporal fusion mechanism ensures globally smooth and temporally consistent reconstructions, while the Annular-Flow diffusion process naturally aligns with the physical scanning characteristics of LiDAR, allowing information to flow outward and support infinite-range scene synthesis.

Through extensive evaluations and ablation studies, PatchScene demonstrates substantial improvements over state-of-the-art methods on the SemanticKITTI dataset, achieving superior Chamfer Distance, JSD, and voxel IoU scores. The framework's ability to generalize across different spatial scales further validates its robustness and scalability.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2

[2] Helin Cao and Sven Behnke. Diffssc: Semantic lidar scan completion using denoising diffusion probabilistic models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2185–2192. IEEE, 2025. 2, 3

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[4] Xiaoliang Ju, Zhaoyang Huang, Yijin Li, Guofeng Zhang, Yu Qiao, and Hongsheng Li. Diffindscene: Diffusion-based high-quality 3d indoor scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4526–4535, 2024. 3

[5] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 4

[6] Pengfei Li, Ruowen Zhao, Yongliang Shi, Hao Zhao, Jirui Yuan, Guyue Zhou, and Ya-Qin Zhang. Lode: Locally conditioned eikonal implicit scene completion from sparse lidar. *arXiv preprint arXiv:2302.14052*, 2023. 2, 7

[7] You Li, Julien Moreau, and Javier Ibanez-Guzman. Emergent visual sensors for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):4716–4737, 2023. 1

[8] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 2

[9] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023. 3

[10] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10083, 2024. 3

[11] Shang Liu, Chenjie Cao, Chaohui Yu, Wen Qian, Jing Wang, and Fan Wang. Earthcrafter: Scalable 3d earth generation via dual-sparse latent diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7260–7268, 2026. 2

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[13] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2837–2845, 2021. 4

[14] Tetiana Martyniuk, Gilles Puy, Alexandre Boulch, Renaud Marlet, and Raoul de Charette. Lidpm: Rethinking point diffusion for lidar scene completion. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 555–560. IEEE, 2025. 3, 7

[15] Benedikt Mersch, Tiziano Guadagnino, Xieyuanli Chen, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Building volumetric beliefs for dynamic environments exploiting map-based moving object segmentation. *IEEE Robotics and Automation Letters*, 8(8):5180–5187, 2023. 1

[16] Amir Meydani. State-of-the-art analysis of the performance of the sensors utilized in autonomous vehicles in extreme conditions. In *International Conference on Artificial Intelligence and Smart Vehicles*, pages 137–166. Springer, 2023. 1

[17] Björn Michele, Alexandre Boulch, Gilles Puy, Tuan-Hung Vu, Renaud Marlet, and Nicolas Courty. Saluda: Surface-based automotive lidar unsupervised domain adaptation. In *2024 International Conference on 3D Vision (3DV)*, pages 421–431. IEEE, 2024. 1

[18] Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in neural information processing systems*, 36:67960–67971, 2023. 1

[19] Kazuto Nakashima and Ryo Kurazume. Lidar data synthesis with denoising diffusion probabilistic models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14724–14731. IEEE, 2024. 2

[20] Lucas Nunes, Rodrigo Marcuzzi, Benedikt Mersch, Jens Behley, and Cyrill Stachniss. Scaling diffusion models to real-world 3d lidar scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14770–14780, 2024. 2, 3, 7

[21] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[22] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14738–14748, 2024. 2

[23] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4209–4219, 2024. 2, 3

[24] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021. 2

[25] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 7

[26] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8):1978–2005, 2022. 2

[27] Jules Sanchez, Jean-Emmanuel Deschaud, and François Goulette. Domain generalization of 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18077–18087, 2023. 1

[28] Yiming Shan, Yan Xia, Yuhong Chen, and Daniel Cremers. Scp: Scene completion pre-training for 3d object detection. *arXiv preprint arXiv:2309.06199*, 2023. 1

[29] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2

[30] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023. 2

[31] Ignacio Vizzo, Benedikt Mersch, Rodrigo Marcuzzi, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Make it dense: Self-supervised geometric scan completion of sparse 3d lidar scans in large outdoor environments. *IEEE Robotics and Automation Letters*, 7(3):8534–8541, 2022. 1, 3, 7

[32] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters*, 8(2):1029–1036, 2023. 5

[33] Cho-Ying Wu and Ulrich Neumann. Scene completeness-aware lidar depth completion for driving scenario. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2490–2494. IEEE, 2021. 2

[34] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5418–5427, 2022. 1

[35] Yuwen Xiong, Wei-Chiu Ma, Jingkang Wang, and Raquel Urtasun. Learning compact representations for lidar completion and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083, 2023. 1

[36] Li Yi, Boqing Gong, and Thomas Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15363–15373, 2021. 1

[37] Qingyang Yu, Lei Chu, Qi Wu, and Ling Pei. Grayscale and normal guided depth completion with a low-cost lidar. In *2021 ieee international conference on image processing (icip)*, pages 979–983. IEEE, 2021. 2

[38] Shengyuan Zhang, An Zhao, Ling Yang, Zejian Li, Chenye Meng, Haoran Xu, Tianrun Chen, AnYang Wei, Perry Pengyun Gu, and Lingyun Sun. Distilling diffusion models to efficient 3d lidar scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2025. 2, 3, 7

[39] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023. 3

[40] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5826–5835, 2021. 7

[41] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, pages 17–35. Springer, 2022. 2

[42] Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6055–6062. IEEE, 2025. 1