This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SG-LDM: Semantic-Guided LiDAR Generation via Latent-Aligned Diffusion

Zhengkang Xiang Zizhao Li Amir Khodabandeh Kourosh Khoshelham
The University of Melbourne

https://github.com/YiGuYT/SG-LDM

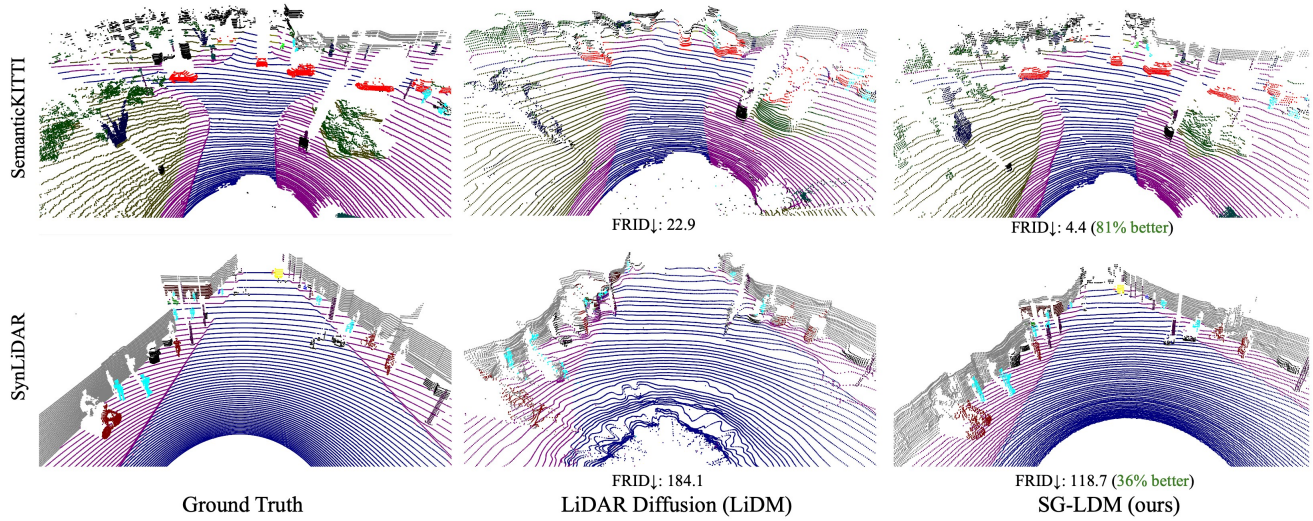

Figure 1. Quantitative comparison of our model and LiDM [45] on semantic-to-lidar conditional generation. Both models are trained on SemanticKITTI. Our approach achieves an improvement of 81% in FRID score on the SemanticKITTI validation set and demonstrates robust generalization, yielding a 36% FRID improvement on the SynLiDAR dataset.

Abstract

Lidar point cloud synthesis based on generative models offers a promising solution to augment deep learning pipelines, particularly when real-world data is scarce or lacks diversity. By enabling flexible object manipulation, this synthesis approach can significantly enrich training datasets and enhance discriminative models. However, existing methods focus on unconditional lidar point cloud generation, overlooking their potential for real-world applications. In this paper, we propose SG-LDM, a Semantic-Guided Lidar Diffusion Model that employs latent alignment to enable robust semantic-to-lidar synthesis. By directly operating in the native lidar space and leveraging explicit semantic conditioning, SG-LDM achieves state-of-the-art performance in generating high-fidelity lidar point clouds guided by semantic labels. Moreover, we propose the first diffusion-based lidar translation framework based on SG-LDM, which enables cross-domain translation as a domain adaptation strategy to enhance downstream perception performance. Systematic experiments demonstrate that SG-LDM significantly outperforms existing lidar diffusion models and the proposed lidar translation framework further improves data augmentation performance in the downstream lidar segmentation task.

1. Introduction

Lidar has clear advantages over RGB cameras in driving scene perception, including geometric accuracy and robustness to poor visibility and weather conditions. However, learning-based lidar perception systems require large-scale annotated datasets [23, 24, 62–64, 74], and manually assigning semantic labels or bounding boxes to each point (or cluster of points) is time-consuming and far more onerous than annotating 2D images. Additionally, existing real-world lidar datasets are naturally imbalanced, i.e., common classes like roads, buildings, and vehicles dominate the point clouds, while important but rarer classes (e.g. pedestrians, cyclists, traffic signs) are underrepresented [3, 6]. This imbalance can bias learning algorithms, which tend to be overly tuned to frequent classes and struggle with minority classes. Due to these, models trained exclusively on real-world datasets suffer from biases and generalization gaps.

To mitigate data scarcity and imbalance, leveraging syn-

thetic data as a complementary resource has received considerable attention in recent years [66, 67]. Advances in simulation platforms and generative modeling now enable the automated generation of large volumes of synthetic data, eliminating the need for labor-intensive data collection. Two primary approaches have emerged:

- **Physics-based simulation in virtual environments**: Leveraging high-fidelity simulators (e.g. CARLA [10], AirSim [51]) built on game engines, this approach utilizes a virtual lidar sensor in a digital world to generate point clouds via ray-casting. The virtual environment can be populated with diverse 3D assets (roads, buildings, vehicles, pedestrians, vegetation, etc.), and since all objects are known, every point is automatically labeled. This approach eliminates the manual labeling effort while allowing complete control over the scene content.
- **Data-driven generative models**: Beyond scripted simulation, deep generative models offer a novel way to synthesize lidar point clouds by learning the underlying data distribution. Recent works have explored using generative adversarial networks (GANs) [5, 37] and diffusion models [20, 36, 45, 65, 77] to produce realistic 3D point patterns that mimic real lidar scans. These models can effectively generate synthetic lidar point clouds without requiring any pre-built 3D assets.

Despite their potential, both approaches face notable limitations. Virtual environments often exhibit a substantial domain gap compared to real-world data, necessitating additional training pipelines such as adversarial training [29, 31, 56, 58, 59, 70] or self-training [49, 71] for domain adaptation. Existing lidar generative models [5, 20, 77] focus on either unconditional data generation, which generates point clouds without incorporating semantic labels, or conditional point cloud upsampling and completion, where the process is guided by partial input data. However, both approaches lack the explicit semantic information necessary for effective data augmentation. Recently, LiDM [45] explored semantic-to-lidar synthesis, however, its performance remains underdeveloped. Specifically, the synthesized point clouds often display coarse geometric details, structural inconsistencies, and significant noise, which undermine their fidelity compared to real-world data. These limitations are clearly illustrated in Figure 1.

We consider semantic-to-lidar a critical task, which has the potential to revolutionize 3D scene perception by enabling controllable, on-demand annotation of diverse scenarios. By conditioning lidar generative models on semantic segmentation maps, we directly specify the spatial arrangement of vehicles, pedestrians, and other scene elements, allowing for the synthesis of rich and complex lidar point clouds that align with predefined scene structures. Such controllable synthesis is particularly advantageous for addressing scenarios that are either rare or impractical for data collection in real-world driving, such as accidents, collisions, and ephemeral road incidents [26]. This level of customizability and diversity would significantly enhance training datasets, improving model robustness and safety in autonomous systems. However, current methods for synthetic lidar data generation lack the ability to effectively utilize the rich semantic information available in both real and virtual environments.

To bridge this gap, we propose SG-LDM, a **S**emantic-**G**uided **L**idar **D**iffusion **M**odel. To effectively integrate classifier-free guidance, we propose a novel semantic alignment technique in the latent space that facilitates diffusion training in both unconditional and conditional modes. Prior diffusion-based approaches [20, 45] rely on latent diffusion architectures with a carefully crafted variational autoencoder (VAE) and the diffusion model performs in the latent space. This incurs substantial compression loss and limited transferability, as these VAEs are typically trained with lidar data from a single source. In contrast, our approach discards the latent diffusion architecture, leading to significant performance improvements and better generalization. Figure 1 compares the lidar point clouds generated by a latent diffusion architecture (LiDM [45]) and our SG-LDM. LiDM exhibits noisy points around the ego vehicle due to suboptimal compression and fails to generalize across domains (trained on SemanticKITTI [3] and tested on SynLiDAR [67]).

Moreover, we propose the first diffusion-based lidar translation framework built upon SG-LDM that leverages the inherent properties of the diffusion process to bridge the domain gap between real and synthetic lidar data. Unlike GAN-based translation frameworks [67], our approach provides a more stable solution by effectively aligning both semantic and geometric features across domains. In summary, the contributions of this paper are as follows:

- We present SG-LDM, a novel semantic-guided lidar diffusion model that establishes a new state-of-the-art in semantic-to-lidar generation.
- We propose a semantic alignment module in the latent space which improves performance of the diffusion model and enables effective classifier-free guidance.
- We introduce the first diffusion-based lidar translation framework built upon SG-LDM to bridge the domain gap between real and synthetic lidar data, offering a more stable alternative to GAN-based approaches.
- Through systematic evaluation of data generation and augmentation performance on the SemanticKITTI and SynLiDAR datasets, we demonstrate that SG-LDM significantly outperforms existing lidar generative models.

2. Related work

Generative Modeling of 3D Point Clouds Generative modeling of 3D point clouds has been an active research

area for several years [2, 30, 57, 61, 69, 72, 75]. More recent studies have proposed methods to condition point cloud generation on auxiliary modalities such as text [38] or images [32]. Moreover, synthetic data produced by these generative models have been demonstrated to effectively augment downstream object recognition tasks [66].

For the generation of outdoor lidar point clouds, most methods require an initial transformation of the point clouds into a range map [5, 20, 36, 45, 65, 77] or a bird's eye view representation [68]. Although these representations typically contain geometry without semantics, LiDM [45] is the only method that addresses the semantic-to-lidar task using a latent diffusion framework.

Diffusion Model Diffusion models have become the dominant paradigm in computer vision generative tasks for both 2D and 3D datasets, underpinning many successful applications in image [9, 16, 18, 39, 40, 44, 46, 48], video [4, 12, 17, 19, 52, 60], and 3D content generation [21, 28, 30, 34, 42, 57]. Since the introduction of the original denoising diffusion probabilistic model (DDPM) [16], many methods have been proposed to enhance the diffusion process. Among the most influential improvements are latent diffusion, denoised diffusion implicit models (DDIM), and classifier-free guidance (CFG). Latent diffusion [46] leverages a pre-trained variational autoencoder (VAE) to perform the diffusion process in a lower-dimensional latent space, thereby reducing computational complexity. DDIM [53] introduces a deterministic sampling procedure as opposed to the stochastic sampling in standard DDPMs and typically allows using significantly fewer inference steps (e.g., going from 1000 down to 50 or fewer) without completely sacrificing image quality. CFG [15] strengthens the conditioning signal, enabling a controllable trade-off between fidelity and diversity in the generated outputs.

Lidar Translation Despite the extensive study of generative modeling-based image-to-image translation [1, 7, 11, 13, 22, 41, 73, 76], analogous techniques for lidar data remain largely underexplored. One strategy in lidar translation involves reconstructing a mesh from sequences of raw point clouds and subsequently employing ray casting to generate data in the target distribution [25]. However, mesh reconstruction introduces further domain discrepancies, primarily due to differences between the mesh representation and the intrinsic properties of raw lidar data. Xiao et al. [67] proposed a GAN-based data translation method that uses two conditional GANs to translate the appearance and sparsity of lidar point clouds, respectively. Yuan et al. [71] proposed a domain adaptive segmentation method by statistically transferring the density of the source point clouds to mimic the density distribution of the target point cloud. Several approaches explicitly model the lidar drop effect in real-world settings to enable synthetic-to-real domain adaptation [35, 37]. However, since these methods focus on a specific task, they cannot effectively address the general domain gap between the synthetic and real point clouds.

3. Method

In this section, we begin by outlining the problem formulation in Section 3.1. Next, we introduce the core components of our diffusion model in Section 3.2. We then introduce our semantic alignment module and detail the refined training process in Section 3.3. Finally, we present a simple lidar translation framework based on our SG-LDM in Section 3.4. Figure 2 presents the overview of the training and inference process of SG-LDM.

3.1. Problem Formulation

Semantic-to-Lidar Generation: Given a labeled point cloud dataset $\mathcal{D} = \{X, Y\}$, where $X = \{\mathbf{x}_i \in \mathbb{R}^3\}_{i=1}^N$ is the set of points and $Y = \{\mathbf{y}_i \in \{1, \ldots, K\}\}_{i=1}^N$ is the set of associated semantic labels, with each $\mathbf{y}_i$ taking one of K possible class values, the task is to learn a generative model parameterized by θ such that the conditional density $p_\theta(X \mid Y)$ accurately captures the distribution of the points given the semantic labels.

Lidar Data Representation: Following existing work on lidar scene generative modeling [5, 20, 36, 45, 77], we adopt the projected range image [33] as our lidar representation. The detailed range image and point cloud conversion is formulated in Section 9. This approach relaxes the problem from 3D conditional generation to 2D conditional generation, enabling us to build our method on a mature 2D diffusion model. Rather than directly learning the conditional density of lidar point clouds $q(X \mid Y)$, we learn the conditional density of their 2D projections $q(\tilde{X} \mid \tilde{Y})$. For clarity, the random variables X and Y in the remainder of this paper refer to the projected range images and labels.

Lidar Translation: Given labeled lidar point cloud datasets from synthetic and real environments, denoted as $\mathcal{D}_s = \{X_s, Y_s\}$ and $\mathcal{D}_r = \{X_r, Y_r\}$, respectively, our goal is to translate synthetic point clouds X_s into $\hat{X}_r$ so that they more closely resemble the real data X_r. Consequently, a model trained with the translated data $\{\hat{X}_r, Y_s\}$ for a downstream task like lidar segmentation is expected to achieve better performance in predicting Y_r compared to a model trained solely on the raw synthetic data $\{X_s, Y_s\}$.

3.2. Revisiting Diffusion Model

We employ denoised diffusion probalistic model (DDPM) [16] as the main training target and classifier-free guidance (CFG) [15] during inference. In DDPM, a model is trained to reverse the noising steps of Markov chain, which is defined as the diffusion process by adding noise to clean data, also called forward diffusion process:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t \mathbf{I}\big) \tag{1}$$

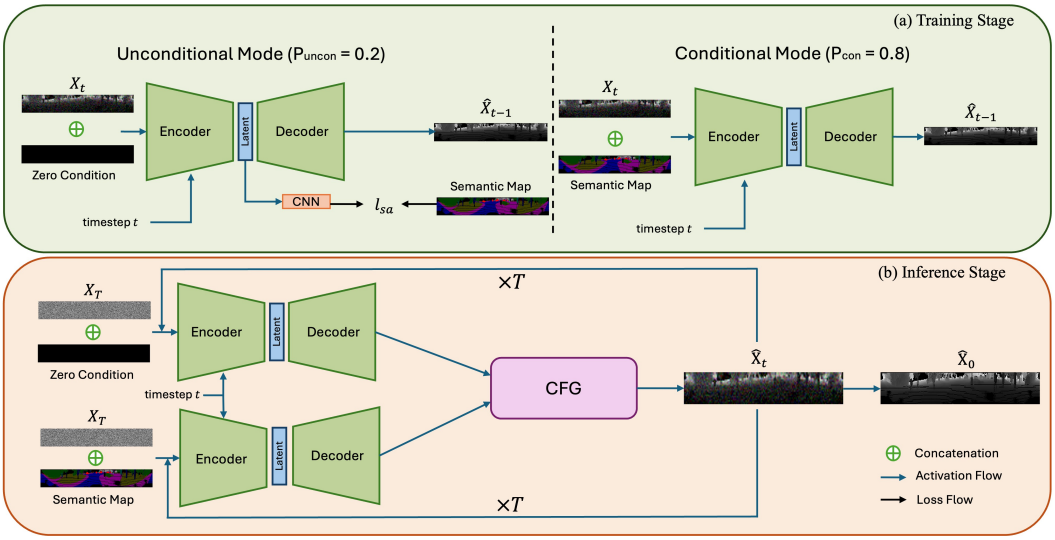

Figure 2. Overview of SG-LDM. The model is trained in both conditional and unconditional modes, enabling classifier-free guidance during inference. In the unconditional mode, we incorporate a semantic alignment strategy in the latent space during training, which enhances the performance of unconditional generation and improves overall results.

Here, $t \in \{1, \ldots, T\}$ denotes the noising steps in the Markov chain, and the noise level at each step is governed by the parameters $\beta_1, \ldots, \beta_T$ generated according to the variance schedule. Given $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, the original clean image $\mathbf{x}_0$ can be used to generate the noisy image at any time step t via the following relation:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big) \tag{2}$$

A diffusion model parameterized by θ is trained to learn the reversed diffusion process:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\big) \tag{3}$$

Given $\epsilon \sim \mathcal{N}(0, I)$, the training objective of the diffusion model is derived by optimizing the variational lower bound, which leads to the following simplified loss function:

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t + \boldsymbol{\epsilon}, t) \right\|^2 \tag{4}$$

Classifier-free guidance (CFG) [15] is an essential technique in diffusion models that enhances conditional generation by amplifying the influence of conditioning information during the reverse diffusion process. Instead of relying solely on a conditional model $p_\theta(\mathbf{X} \mid \mathbf{Y})$, CFG also trains an unconditional model $p_\theta(\mathbf{X})$. During sampling, the noise predictions from both models are linearly combined:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{y}) = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) \tag{5}$$

where we have guidance scale $= w + 1$. By increasing w, the reverse diffusion process is more strongly tied to condition $\mathbf{y}$, resulting in outputs with higher fidelity at the cost of reduced diversity.

To train a diffusion model with both conditional and unconditional modes, a common practice is to set a probability for the unconditional mode, e.g. $\mathbf{P}_{uncon} = 0.2$, which means that the training process will have a 20% chance of having the semantic condition $\mathbf{y}$ replaced by a null condition $\varnothing$.

3.3. Semantic Alignment

During experiments with CFG on a standard diffusion model, we observed a marked decline in performance after incorporating CFG. A closer analysis revealed that the model's ability to generate content unconditionally was severely impaired when trained alongside its conditional counterpart. We hypothesize that the model was inadvertently optimized to leverage conditioning cues, so that when these cues are absent, it lacks sufficient information to produce coherent outputs. To address this, we introduce a semantic alignment module for the unconditional mode, ensuring that it can extract and utilize semantic information from raw inputs even without the explicit condition.

We incorporate a three-layer convolutional neural network h_ϕ that operates on the latent features produced by the diffusion encoder f_θ. This network projects these features into a space that matches the dimensions of a downscaled semantic map. We then enforce semantic consistency by applying a cosine similarity loss between the projected features and the semantic map. This alignment loss guides the model to preserve robust semantic representations, even when explicit conditioning information is absent:

$$\mathcal{L}_{\text{SA}}(\theta, \phi) = -\mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t} \left[\cos(\mathbf{y}_i, h_\phi(f_\theta(\mathbf{x}_t)) \right] \tag{6}$$

Additionally, to ensure that the semantic alignment is only

applied when it is meaningful, we modulate its influence dynamically. Since it is uninformative for the encoder to extract features from predominantly noisy inputs at large timesteps, we introduce a dynamic weight $\lambda = 1 - \frac{\mathbf{t}}{\mathbf{T}}$. This schedule gradually decreases the alignment strength as the noise level increases. The final training objective for our SG-LDM in the unconditional mode becomes:

$$\mathcal{L} = \mathcal{L}_{\text{DDPM}} + \lambda \mathcal{L}_{\text{SA}} \tag{7}$$

Combining DDPM and semantic alignment, the training algorithm of our SG-LDM becomes:

Algorithm 1 DDPM Training with Semantic Alignment

repeat
Sample $(x_0, y) \sim q(x_0, y)$
Sample $t \sim \text{Uniform}(\{1, \ldots, T\})$
Sample $\epsilon \sim \mathcal{N}(0, I)$
Compute noisy sample:
$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$
Sample $u \sim \text{Uniform}(0, 1)$
if $u < 0.8$ **then**
$\mathcal{L} = \left\| \epsilon - \epsilon_\theta(x_t, t, y) \right\|^2$
Take gradient step on $\nabla_\theta \mathcal{L}$
else
$\mathcal{L}_{\text{DDPM}} = \left\| \epsilon - \epsilon_\theta(x_t, t, \varnothing) \right\|^2$
$\mathcal{L}_{\text{SA}} = -\cos(\mathbf{y}_i, h_\phi(f_\theta(\mathbf{x}_t)))$
$\lambda = 1 - \frac{\mathbf{t}}{\mathbf{T}}$
$\mathcal{L} = \mathcal{L}_{\text{DDPM}} + \lambda \mathcal{L}_{\text{SA}}$
Take gradient step on $\nabla_{\theta,\phi} \mathcal{L}$
end if
until converged

For the sampling process, we follow the standard classifier-free guidance using the linear combination (Eq. (5)) as illustrated in Fig. 2. This approach combines conditional and unconditional noise predictions, effectively improving the fidelity of the generated lidar samples conditioning on semantic maps.

3.4. Lidar Translation Framework

The diffusion model naturally bridges different data domains by progressively destroying the original information until it converges to a Gaussian distribution. Building on this concept, we propose a novel lidar translation framework based on SG-LDM. This framework is visually demonstrated in Figure 3. Specifically, we apply the forward diffusion process (Eq. (1)) to degrade information in the synthetic dataset and then use the reverse diffusion process (Eq. (3)) to reconstruct lidar point clouds that conform to the target distribution from the intermediate states. This transformation can only be achieved with a semantic-to-lidar generative model. Unconditional models fail to retain the original semantic information, leading to the loss of semantic labels. By incorporating semantic labels as constraints, our approach ensures that the reverse diffusion process does not generate extraneous objects, thereby preserving the per-point labels essential for effective data augmentation.

4. Experiments

4.1. Experimental Setup

Our experiments consist of two stages: data generation and data augmentation. We utilize two datasets, SemanticKITTI [3] and SynLiDAR [67], in both sets of experiments. In the first stage, we evaluate the data generation performance of the proposed diffusion model and compare it with the state-of-the-art, where all the diffusion models are exclusively trained on the official training partition of SemanticKITTI, and tested on the validation partition of SemanticKITTI. We also evaluate the transferability of the models on the test partition of SynLiDAR. For fast sampling, we employ DDIM for both our SG-LDM and LiDM, reducing the original 1000 DDPM steps to 50 steps. The diffusion model is a standard 2D diffusion with the conventional convolution neural networks replaced by circular convolution [50].

In the second stage, we use MinkovUnet [8] as the baseline segmentation model to evaluate the performance of models trained with different data augmentation methods. These methods include traditional techniques such as jittering and random drop, synthetic data generated from virtual environments or generative models, and domain translation applied to the synthetic data produced in virtual environments. All experiments, encompassing both generative and segmentation models, are conducted using four Nvidia A100 GPUs.

4.2. Experiments on Data Generation

In this section, we evaluate the generated data against the ground truth. Following [45], we employ FRID, FSVD, and FPVD as perceptual metrics and JSD and MMD as statistical metrics. FRID, FSVD, and FPVD are analogous to the FID score [14] used in image generation. They are computed using different point cloud representation learning backbones, namely, RangeNet++ [33], MinkowskiNet [8], and SPVCNN [55]. JSD measures the diversity of the marginal distribution of two sets of point clouds, while MMD calculates the average distance between matched point clouds, indicating fidelity.

Table 1 presents the quantitative results, demonstrating that our method establishes a new state-of-the-art for semantic-to-lidar generation on most metrics. Specifically, when evaluated in the same domain (i.e., testing on SemanticKITTI), our approach outperforms the previous state-of-the-art method, LiDM [45], by $48.0\% \sim 80.8\%$ in the perceptual metrics and achieves a 59% improvement

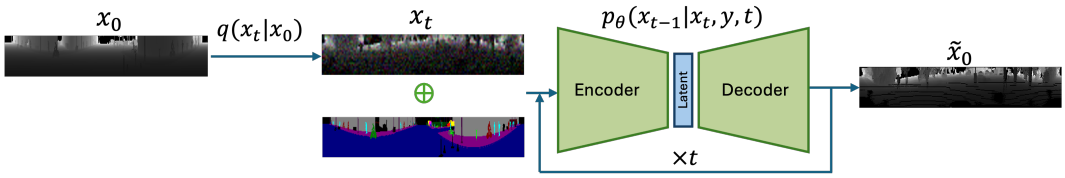

Figure 3. Lidar Translation Framework. Forward diffusion is applied to the source data x_0 to progressively degrade its information, while backward diffusion reconstructs the target data $\tilde{x}_0$ from an intermediate noisy state x_t. The timestep t represents the noise level, indicating the number of diffusion steps executed in the process. The range maps shown here are for demonstration only and do not represent actual experimental results.

Method	**SemanticKITTI [3] (Same Domain)**					**SynLiDAR [67] (Different Domain)**				
	FRID ↓	**FSVD** ↓	**FPVD** ↓	**JSD** ↓	**MMD** ($\times 10^{-4}$) ↓	**FRID** ↓	**FSVD** ↓	**FPVD** ↓	**JSD** ↓	**MMD** ($\times 10^{-4}$) ↓
LiDARGen [77]	42.5	31.7	30.1	0.130	5.18	-	-	-	-	-
Latent Diffusion [46]	24.0	21.3	20.3	0.088	3.73	-	-	-	-	-
LiDM [45]	22.9	20.2	17.7	**0.072**	3.16	184.1	147.9	144.3	0.277	**1.48**
SG-LDM (ours)	**4.4**	**10.5**	**7.9**	0.084	**1.31**	**118.7**	**78.0**	**80.3**	**0.145**	3.63

Table 1. Quantitative evaluation of semantic-to-lidar generation. All models are trained using the SemanticKITTI training partition. The results shown here are adapted from [45]. We employ the exact same evaluation toolkit to ensure consistency and comparability.

	Methods	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	other-g.	build.	fence	veget.	trunk	terr.	pole	traf.	mIoU	gain
	Source Only	95.7	25.0	57.0	62.1	46.4	63.4	77.3	0.0	93.0	47.9	80.5	2.2	89.7	58.6	89.5	66.5	78.0	64.6	50.1	60.3	+0.0
	Jittering [43]	95.7	27.8	56.2	66.0	45.8	65.3	82.8	0.0	93.0	48.2	79.9	2.5	89.7	**62.9**	88.9	64.0	77.0	64.8	**51.0**	61.2	+0.9
A	Dropout [54]	96.0	28.5	57.1	65.1	46.4	64.1	83.6	0.1	93.5	47.6	80.1	2.3	89.3	61.9	**90.1**	66.9	**78.8**	**65.8**	49.1	61.4	+1.1
	PointAug [27]	95.9	29.2	70.0	76.3	50.0	67.0	84.4	**2.4**	93.8	48.1	81.2	**4.6**	89.8	58.4	87.5	65.4	72.7	62.4	50.5	62.6	+2.3
	+SynLiDAR [67]	95.9	33.0	62.8	78.9	50.2	71.4	83.5	0.7	92.3	52.8	79.9	0.1	89.8	59.5	86.3	65.4	72.8	63.6	48.9	62.5	+2.2
B	+LiDM [45]	95.5	19.3	50.4	77.8	46.0	65.7	74.2	0.0	93.7	46.4	80.7	0.8	90.0	59.3	87.3	65.3	74.0	62.3	45.8	59.7	-0.6
	+SG-LDM (ours)	96.5	24.8	52.1	80.0	56.3	67.1	86.2	0.0	**94.4**	47.9	**81.6**	0.3	90.9	62.7	87.3	66.9	73.4	63.3	48.8	62.1	+1.8
	PCT [67]†	96.3	**38.7**	**73.4**	82.9	56.1	71.1	85.3	1.6	94.1	**54.3**	**81.6**	1.3	89.5	59.6	87.8	66.9	73.6	65.4	50.5	**64.7**	**+4.4**
C	DGT [71]	96.4	30.8	63.6	81.2	57.5	71.0	**86.5**	0.0	**94.4**	47.4	81.4	2.2	90.9	60.8	87.3	**67.9**	73.4	62.7	48.0	63.3	+3.0
	SG-LDM + LT (ours)	**97.3**	24.0	59.4	**91.5**	**73.6**	**71.7**	84.5	0.0	94.0	45.4	81.0	0.6	**91.0**	61.3	87.2	65.0	72.9	62.8	46.3	63.7	+3.4

Table 2. Data Augmentation Results. The baseline lidar segmentation model, MinkovUnet [8], is trained on the SemanticKITTI training partition and augmented using various methods. Group A applies augmentation directly to real data. Group B uses synthetic lidar point clouds generated either from a virtual environment (SynLiDAR) or via semantic-to-lidar generative models guided by SynLiDAR semantic labels. Group C combines SynLiDAR point clouds with a lidar translation (LT) method for domain adaptation. †: The PCT result is adapted from [67], DGT and our methods were evaluated under identical experimental conditions.

in MMD. Although our method sacrifices 17% performance in the JSD score due to its reliance on classifier-free guidance to balance fidelity and diversity, it still outperforms alternative approaches overall. The evaluation in a different domain (SynLiDAR) reveals that our method continues to outperform LiDM by $35.6\% \sim 47.3\%$ in the perceptual metrics. LiDM appears to perform slighlty better in terms of MMD. However, given that SynLiDAR exhibits a significantly different point range, LiDM produces a significant amount of noisy points (as shown in Figure 1) which is not captured by the statistical metrics.

Figure 4 further zooms in on object details and demonstrates the correspondence between semantic labels and the generated lidar point clouds. LiDM exhibits significant object collapse in the generated SemanticKITTI point clouds, whereas our method preserves the structural integrity of objects. As for SynLiDAR, LiDM suffers from considerable degradation in object and ground point cloud quality. This is attributed to the VAE used for latent diffusion being unable to generalize to lidar data sourced from a different domain.

4.3. Experiments on Data Augmentation

In this section, we leverage the official SemanticKITTI training and validation datasets to evaluate the performance of a baseline segmentation model under various data augmentation strategies, which we categorize into three groups. Group A comprises augmentation techniques applied directly to the original real data. Group B augments the dataset using synthetic lidar point clouds generated either by a virtual environment (SynLiDAR) or by semantic-to-

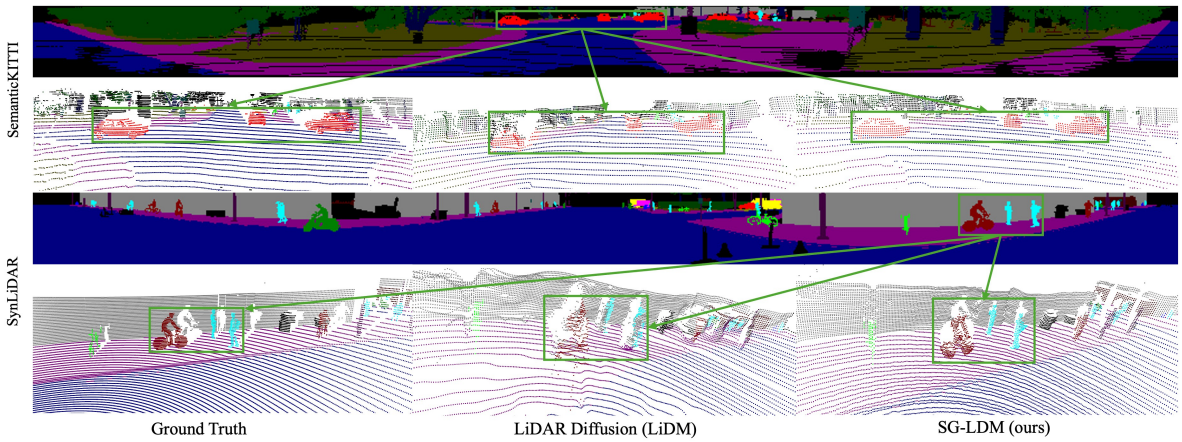

Figure 4. Qualitative comparison of generated lidar point clouds between the LiDM and our SG-LDM.

lidar generative models, with the latter leveraging semantic labels from the SynLiDAR dataset as guidance. Group C employs SynLiDAR point clouds in combination with a lidar translation method to facilitate domain adaptation.

Table 2 presents the results from all groups. In Group B, the data generated by SG-LDM demonstrates a clear data augmentation effect. When using SG-LDM–generated data, the performance is only 0.4% lower than with SynLiDAR, demonstrating that relying solely on semantic information can yield comparable results. In contrast, due to the poor quality of the data generated by LiDM, incorporating it as additional training data actually degrades model performance.

In Group C, we evaluate the data augmentation performance using synthetic data from SynLiDAR with various lidar transation methods. As a density translation method, our approach shows strong potential by outperforming the state-of-the-art density guided translator (DGT) [71]. However, since our method focuses solely on point cloud density, there remains some gap between our approach and PCT [67], which considers both the density and appearance of point clouds.

5. Ablation Study

5.1. Semantic Alignment

In this section, we compare three variations of the proposed conditional diffusion model: one without CFG, one with regular CFG, and one with CFG combined with semantic alignment (SA). Both experiments with CFG are conducted with $P_{uncon} = 0.2$ and a CFG scale of 2. Table 3 presents the quantitative results for these three models. We observe a clear drop in performance when CFG is applied to a conditional model without the semantic alignment module. However, when semantic alignment is incorporated, classifier-free guidance maintains its intended behavior, which successfully trading off fidelity, as indicated by the perceptual metrics and MMD, with only a minimal cost in diversity as measured by JSD.

Method	**FRID** ↓	**FSVD** ↓	**FPVD** ↓	**JSD** ↓	**MMD** ($\times 10^{-4}$) ↓
No CFG	9.4	10.8	9.3	**0.078**	1.43
CFG w/o SA	10.4	32.1	24.7	0.222	1.75
CFG w/ SA	**4.4**	**10.5**	**7.9**	0.084	**1.31**

Table 3. Evaluation of the different components of SG-LDM on SemanticKITTI. The CFG scale is set as 2, $\mathbf{P}_{uncon}$ is set as 20%.

5.2. Classifier-Free Guidance

In this section, we present an analysis of two key parameters related to classifier-free guidance: $\mathbf{P}_{\text{uncon}}$ and the CFG scale. Table 4 presents the quantitative analysis of the performance of our SG-LDM under different P_{uncon} values. The results are quite stable across various settings, except for a notable drop in the FRID score when P_{uncon} is set to 0.5. We attribute this decline to the unconditional generation component beginning to dominate the training process, which leads to a degradation in conditional generation performance. When P_{uncon} is too high, the model's focus shifts away from the conditioning information, thereby compromising the quality of the generated data.

P_{uncon}	**FRID** ↓	**FSVD** ↓	**FPVD** ↓	**JSD** ↓	**MMD** ($\times 10^{-4}$) ↓
0.1	4.5	11.9	9.27	0.085	**1.26**
0.2	**4.4**	**10.5**	7.9	**0.084**	1.31
0.5	14.5	9.5	**7.8**	0.099	1.75

Table 4. Effect of $\mathbf{P}_{uncon}$ on semantic-to-lidar generation using SemanticKITTI. The CFG scale is set as 2.

Figure 5 illustrates the trade-off between fidelity and diversity for CFG scales ranging from 1.1 to 4.0. We use

FRID as the fidelity metric because our point clouds undergo a range conversion process, ensuring that only points which can be projected to a range image and re-projected back are evaluated. As shown in the figure, we successfully balance fidelity and diversity, with FRID achieving optimal performance at a CFG scale of 2.

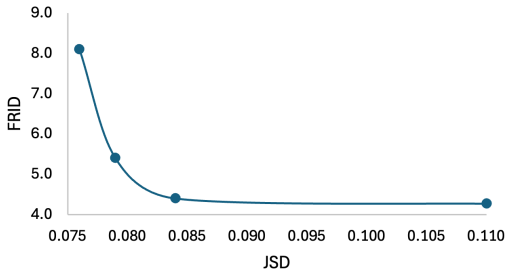

Figure 5. Trade-off between FRID (fidelity) and JSD (diversity). Data points correspond to CFG scales of 1.1, 1.5, 2.0, and 4.0, respectively.

5.3. Noising Percentage in Lidar Translation

We conduct an ablation study on the noising percentage applied to the source data in the lidar translation framework. Here, 0% represents using the raw data from SynLiDAR without any translation, while 100% indicates that the data is fully generated by SG-LDM. As shown in Figure 6, the best performance is achieved at 50%. This observation suggests that a moderate amount of noise introduced during the generation process helps the model produce augmented data that is both realistic and diverse. At 50%, the process strikes an effective balance by preserving the basic geometric structure from the synthetic dataset while also incorporating the translation benefits learned by SG-LDM from real data.

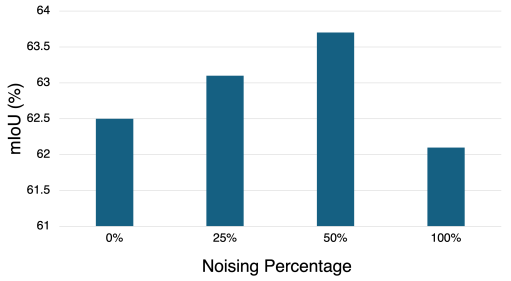

Figure 6. Quantitative Comparison of the data augmentation performance under different noising percentage in the lidar translation framework. If we have 1000 DDPM steps, 50% denoising means applying forward diffusion for 500 steps on the source data and then performing 500 backward diffusion steps using SG-LDM.

5.4. Efficiency

One of the main challenges of traditional diffusion models is their inference-time efficiency. In our work, we benchmark SG-LDM against a latent diffusion model, LiDM [45], using a single Nvidia A100 GPU. As shown in Table 5, while LiDM is 170% faster at inference, its performance on the FRID metric is 420.5% inferior compared to SG-LDM. These results suggest that the additional inference cost of SG-LDM is a worthwhile trade-off for its significant performance improvements.

Method	Latent Diffusion	Throughput ↑	Infer. Speed ↑
LiDM [45]	✓	**1.711**	**85.53**
SG-LDM (ours)	✗	0.634	31.68

Table 5. Comparison on efficiency of LiDM and our SG-LDM using the semanticKITTI validation set. We test both models on one NVIDIA A100 with different batch sizes to ensure full utilization of 80GB GPU memory. Both applied DDIM at the inference stage. Throughput is defined as the number of generated samples per second, while inference speed refers to the number of diffusion steps executed per second.

6. Limitations and Future Work

The primary limitation of our lidar translation framework is that we can only transfer density, not appearance. This constraint arises because the translation process is guided by a semantic label that is generated from a virtual environment. As a result, our method does not outperform the state-of-the-art method PCT [67], which translate both the density and appearance using conditional GAN. Future work should explore ways to relax this requirement, either by using an alternative guidance mechanism or by first translating the semantic map to better match the appearance of lidar point clouds. Additionally, our approach relies on the standard DDPM architecture, which is less efficient than LiDM [45]. Future research should focus on developing methods for training VAEs that are robust to the domain variability in lidar point clouds, thereby laying the foundation for an effective latent diffusion model for lidar data.

7. Conclusion

In this paper we proposed SG-LDM, a novel lidar diffusion model specifically designed for the semantic-to-lidar task. We evaluated our approach under two settings: by comparing the quality of the generated data and by using the generated data to augment a semantic segmentation model. In both cases, our method clearly outperforms the previous state of the art. Furthermore, we introduce a simple yet effective semantic-guided lidar translation framework that achieves comparable performance and a more stable alternative to GAN-based approaches, surpassing synthetic data directly generated from virtual environments.

Acknowledgments

The first two authors acknowledge the financial support from The University of Melbourne through the Melbourne Research Scholarship. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

[1] Debaditya Acharya, Christopher James Tatli, and Kourosh Khoshelham. Synthetic-real image domain adaptation for indoor camera pose regression using a 3d model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:405–421, 2023. 3

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 3, 1

[3] 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. 1, 2, 5, 6

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 3

[5] Lucas Caccia, Herke Van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of lidar data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5034–5040. IEEE, 2019. 2, 3, 1

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 5, 6, 1

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023. 3

[13] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 1

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 3

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 3

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[20] Qianjiang Hu, Zhimin Zhang, and Wei Hu. Rangeldm: Fast realistic lidar point cloud generation. In *European Conference on Computer Vision*, pages 115–135. Springer, 2024. 2, 3, 1

[21] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[23] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023. 1

[24] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17545–17555, 2023. 1

[25] Yixing Lao, Tao Tang, Xiaoyang Wu, Peng Chen, Kaicheng Yu, and Hengshuang Zhao. Lit: Unifying lidar" languages" with lidar translator. *Advances in Neural Information Processing Systems*, 37:93767–93789, 2025. 3

[26] Alex Levering, Martin Tomko, Devis Tuia, and Kourosh Khoshelham. Detecting unsigned physical road incidents from driver-view images. *IEEE Transactions on Intelligent Vehicles*, 6(1):24–33, 2020. 2

[27] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Pointaugment: an auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6378–6387, 2020. 6

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 3

[29] Haifeng Luo, Kourosh Khoshelham, Lina Fang, and Chongcheng Chen. Unsupervised scene adaptation for semantic segmentation of urban mobile laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:253–267, 2020. 2

[30] 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. 3

[31] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2507–2516, 2019. 2

[32] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 3

[33] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4213–4220. IEEE, 2019. 3, 5, 1

[34] 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. 3

[35] Kazuto Nakashima and Ryo Kurazume. Learning to drop points for lidar scan synthesis. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 222–229. IEEE, 2021. 3

[36] 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, 3

[37] Kazuto Nakashima, Yumi Iwashita, and Ryo Kurazume. Generative range imaging for learning scene priors of 3d lidar data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1256–1266, 2023. 2, 3

[38] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[39] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 3

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[41] Duo Peng, Ping Hu, Qiuhong Ke, and Jun Liu. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 808–820, 2023. 3

[42] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[43] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 6

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[45] 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. 1, 2, 3, 5, 6, 8

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 6

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour,

Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[49] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022. 2

[50] Stefan Schubert, Peer Neubert, Johannes Pöschmann, and Peter Protzel. Circular convolutional neural networks for panoramic images and laser data. In *2019 IEEE intelligent vehicles symposium (IV)*, pages 653–660. IEEE, 2019. 5, 1

[51] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018. 2

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 1

[54] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 6

[55] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 5, 1

[56] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2

[57] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 3

[58] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 2

[59] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020. 2

[60] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 3

[61] Lemeng Wu, Dilin Wang, Chengyue Gong, Xingchao Liu, Yunyang Xiong, Rakesh Ranjan, Raghuraman Krishnamoorthi, Vikas Chandra, and Qiang Liu. Fast point cloud generation with straight flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9445–9454, 2023. 3

[62] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 1

[63] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024.

[64] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19551–19562, 2024. 1

[65] Yang Wu, Kaihua Zhang, Jianjun Qian, Jin Xie, and Jian Yang. Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer. In *European Conference on Computer Vision*, pages 291–310. Springer, 2024. 2, 3

[66] Zhengkang Xiang, Zexian Huang, and Kourosh Khoshelham. Synthetic lidar point cloud generation using deep generative models for improved driving scene object recognition. *Image and Vision Computing*, 150:105207, 2024. 2, 3, 1

[67] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2795–2803, 2022. 2, 3, 5, 6, 7, 8

[68] 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. 3

[69] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. 3

[70] Zhimin Yuan, Ming Cheng, Wankang Zeng, Yanfei Su, Weiquan Liu, Shangshu Yu, and Cheng Wang. Prototype-guided multitask adversarial network for cross-domain lidar point clouds semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2

[71] Zhimin Yuan, Wankang Zeng, Yanfei Su, Weiquan Liu, Ming Cheng, Yulan Guo, and Cheng Wang. Density-guided translator boosts synthetic-to-real unsupervised do-

main adaptive segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23303–23312, 2024. 2, 3, 6, 7

[72] Maciej Zamorski, Maciej Zięba, Piotr Klukowski, Rafał Nowak, Karol Kurach, Wojciech Stokowiec, and Tomasz Trzciński. Adversarial autoencoders for compact representations of 3d point clouds. *Computer Vision and Image Understanding*, 193:102921, 2020. 3

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[74] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 1

[75] 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. 3

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[77] 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, 3, 6, 1