SuperPC: A Single Diffusion Model for Point Cloud Completion, Upsampling, Denoising, and Colorization

Yi Du, Zhipeng Zhao, Shaoshu Su, Sharath Golluri, Haoze Zheng, Runmao Yao, Chen Wang
Spatial AI & Robotics (SAIR) Lab, University at Buffalo
{yid, chenw}@sairlab.org

Abstract

Point cloud (PC) processing tasks—such as completion, upsampling, denoising, and colorization—are crucial in applications like autonomous driving and 3D reconstruction. Despite substantial advancements, prior approaches often address each of these tasks independently, with separate models focused on individual issues. However, this isolated approach fails to account for the fact that defects like incompleteness, low resolution, noise, and lack of color frequently coexist, with each defect influencing and correlating with the others. Simply applying these models sequentially can lead to error accumulation from each model, along with increased computational costs. To address these challenges, we introduce SuperPC, the first unified diffusion model capable of concurrently handling all four tasks. Our approach employs a three-level-conditioned diffusion framework, enhanced by a novel spatial-mix-fusion strategy, to leverage the correlations among these four defects for simultaneous, efficient processing. We show that SuperPC outperforms the state-of-the-art specialized models as well as their combination on all four individual tasks. Project website: https://sairlab.org/superpc/.

1. Introduction

Point cloud (PC) processing [19, 26] has become a cornerstone technique in many fields such as object recognition [2, 81], mapping [56, 75], autonomous vehicle navigation [29, 43], and 3D modeling [13, 51]. This diverse applicability stems from the ability of point cloud to capture the geometric intricacies of the physical world with high fidelity, offering a detailed substrate from which nuanced computational analyses can be performed [5, 19, 26, 54].

Despite this critical need and importance, the prior PC processing methods, either geometry-based or data-driven, have typically concentrated on individual sub-tasks like denoising [35, 53], upsampling [21, 70], completion [72, 73], and colorization [32, 58], rather than a single comprehensive solution. While some studies [9, 35, 36, 38] addressed the partial combination of those tasks, there is still a lack of a single model that addresses all four tasks simultaneously, which could bring multiple benefits to this field. This is because an integrated model may be advantageous not just for its computational efficiency but also for preventing error accumulation and allowing them to implicitly improve each other due to the interconnectivity of the tasks. For instance, in the experiments, we found that the error of the completion model accumulated and affected the upsampling; the error pattern produced by an upsampling model often resulted in suboptimal denoising. As a result, we argue that a single unified model that is capable of concurrently handling the four point cloud processing tasks is necessary.

To this end, we take PC processing as generative tasks and resort to diffusion-based models [33, 33, 34, 36, 38, 55], which have shown significant advancements on PC generation. Despite this, using diffusion models to address all four tasks is challenging, and previous studies [33, 34, 36, 38] have been limited to tackling either individual tasks. The first limitation is their selective preservation of input information, whether global, local, or raw, using the input point cloud solely as a condition to guide the diffusion process. These approaches lacks the ability to generate point clouds that meet the requirements of all four tasks, as each task possesses distinct characteristics and demands. For instance, the approach [34] relies solely on the global shape latent features for guiding 3D shape generation, which results in a limited capacity to capture the local object details; "PC^2" [38] only employs the raw features of the input image as the condition, lacking the ability to preserve global-level information. Another limitation of previous methods is their restriction to a single modality. For instance, the approach by Luo et al. [34] is confined to processing point cloud inputs, while the method by Luke et al. [38] is limited to handling only image inputs. As a result, they only work for simple object-level PC like ShapeNet [7] but struggle on complex scene-level PC such as KITTI-360 [31].

To address the first limitation, we introduce SuperPC, which integrates a Three-Level-Conditioned (TLC) diffusion framework to simultaneously handle four tasks. For

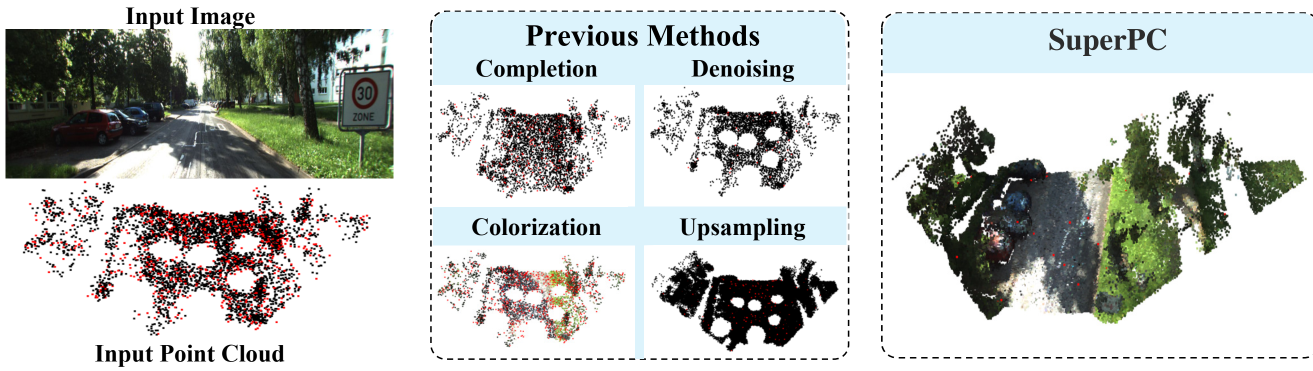


Figure 1. We propose SuperPC, a novel neural architecture that jointly solves inherent shortcomings in the raw point clouds, including noise, sparsity, incompleteness, and the absence of color. To the best of our knowledge, it is the **first** single diffusion model that can simultaneously tackle the **four** major challenges in the field of point cloud processing. Red points denote high noise for visualization.

tackling the second limitation, we propose a novel Spatial-Mix-Fusion (SMF) strategy to integrate images and point clouds. It loosely couples the early-fusion strategy and the deep-fusion strategy [24], fully leveraging the fused information as three-level conditions to guide PC generation. Additionally, we design three modules, namely, a raw, a local, and a global module as the three conditions to control the diffusion process. Specifically, (1) the raw module preserves the original information of the input point cloud and image via projection and interpolation, which can also be viewed as an early-fusion method to combine the two input modalities; (2) the local module extracts geometrical and texture details by fusing the local feature maps of the two modalities, which can also be viewed as a deep-fusion method; and (3) the global module guarantees high-level semantic consistency by condensing the fused local feature maps into a latent code. This enables SuperPC to generate high-quality, color-rich point clouds across diverse tasks and datasets while faithfully preserving local texture and geometrical information. To the best of our knowledge, SuperPC is the **first** single diffusion model fusing **two** modalities by utilizing **three**-level conditions to concurrently tackle the **four** PC processing tasks. Our contributions can be summarized as follows:

- We propose SuperPC, a three-level-conditioned diffusion model framework with a novel spatial-mix-fusion strategy, allowing a single model, for the first time, to effectively tackle all four major challenges in PC processing.
- Our SuperPC model not only demonstrates superior performance on combination task but also surpasses existing SOTA models across all four individual PC processing tasks on scene-level benchmarks while exhibiting impressive efficiency and generalization capability.
- We have designed three new benchmarks for PC processing tasks: one object-level and two scene-level benchmarks. Additionally, these benchmarks are organized into two evaluation tracks to assess generalization ability—Object-to-Scene and Sim-to-Real.

2. Related Works

2.1. PC processing Tasks

Point Cloud Upsampling, essential for enriching sparse 3D scanning outputs, has evolved from traditional methods like Alexa's work [3] to those deep learning-driven methods [15, 21, 28, 46, 47, 69, 70]. However, these methods struggle to upsample and preserve the details of scene-level point clouds. Unlike previous efforts, SuperPC addresses this challenge by preserving geometric details through the three-level-condition framework.

Point Cloud Completion is aiming to generate points for both unobserved areas and incomplete shapes, as initially proposed by pioneers in the field [40, 61]. Following research has employed deep learning techniques, resulting in substantial improvements in the quality of completed point clouds [4, 20, 25, 76]. Yu's methods [71, 72] exemplify this progress by converting the point cloud to a sequence of point proxies. However, they struggle in overly sparse conditions and unseen scenarios since they cannot extract dense structural and semantic information. In contrast, SuperPC can effectively fuse and utilize information from both the input point clouds and images, which contain dense semantic features via the novel spatial-mix-fusion strategy.

Point Cloud Denoising is critical for refining the point cloud by removing outliers. Previous efforts [27, 35, 48, 53] try to tackle it as a single task with traditional filter-based and novel learning-based methods. However, this process is heavily dependent on the point cloud upsampling and the completion tasks. For instance, denoising models may mistakenly remove a table leg because they recognize it for outliers when the table point cloud is too sparse or incomplete.

Point Cloud Colorization: Traditional projection methods have been supplemented by geometry-based studies [6, 8, 12], yet they struggle to color unseen and occluded points. Deep learning advancements [16, 23, 32, 58] have enabled colorization solely based on geometry without image input, but often at the cost of generating colors that diverge from reality. Specialized models like Takayuki's work [58] focus on niche applications such as Airborne LiDAR point clouds colorization and cannot be applied to everyday objects and scenes. In contrast, SuperPC can render both visible and occluded point clouds with realistic colors based on the input image and the point cloud geometric features for both object-level and scene-level point clouds on both simulation and real-world datasets.

2.2. Diffusion Models for Point Cloud

The diffusion model, as evidenced by several studies [33, 49, 55], has shown significant capabilities in image synthesis, prompting its application in 3D point cloud challenges. Luo [34] pioneered the use of a conditional diffusion model for 3D PC generation, guided by the shape latent condition of the input cloud. Building on this, Zhuo [80] and Zeng [74] explored unconditional shape generation and the hierarchical point diffusion model for point cloud synthesis in latent space. Besides the point cloud synthesis, some researchers [35, 36, 38, 46] also try to tackle the PC processing tasks mentioned in Section 2.1 by using different conditions to guide the PC generation. Despite previous efforts made some progress in leveraging diffusion models for point cloud completion [36], denoising [35], upsampling [46], and single-image 3D reconstruction [38], no work has yet to integrate two modalities into a structured three-level-condition framework to guide the diffusion model in addressing the four tasks simultaneously.

2.3. Image and Point Cloud Fusion

To effectively harness the inherent capabilities and potentials of both image and point cloud inputs, a robust fusion methodology becomes indispensable. The concept of image and point cloud fusion is rooted in the understanding that dense color or RGB information inherently carries 3D geometric details. As a result, images can be used as a reference to enhance the quality of the spatial point cloud. Stated by Huang et al. [24], those fusion methods can be divided into the early-fusion [59, 60, 66], the deep-fusion [62, 67, 78], and the late-fusion [18, 39, 44]. Previous research on image and point cloud fusion has predominantly concentrated on high-level tasks [30, 60, 62, 66, 79] such as object detection and specific conventional tasks like point cloud completion and upsampling [1, 41, 77]. However, these approaches are confined to a single task and limited scenarios. Hence, we propose SuperPC, a single diffusion model with a novel SMF strategy which combines the two modalities effectively and perform all the four tasks.

3. Background

3.1. Conditional Diffusion Probabilistic Models

SuperPC employs Conditional Diffusion Probabilistic Models (CDPMs) [33, 34, 36, 38, 55] to synthesize high-quality, dense point clouds, conditioned on the input images and point clouds. It consists of two Markov chains known as the forward processes and reverse processes [22].

Forward Process Also known as a Markov process that incrementally introduces Gaussian noises into the data distribution $\boldsymbol{x_0}$, until transitioning it towards a pure Gaussian noise $\boldsymbol{x_T}$. In other words, the point clouds progressively diffuse over time, finally forming purely chaotic points in a 3D space. This forward process from the target clean point cloud $\boldsymbol{x_0}$ to purely noisy point cloud $\boldsymbol{x_T}$ and and can be formulated as the following probability distribution:

$$\begin{aligned} q(\boldsymbol{x_{1:T}}|\boldsymbol{x_0}) &= \prod_{t=1}^{T} q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}), \\ q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}) &= \mathcal{N}\left(\boldsymbol{x_t}; \sqrt{1-\beta_t}\boldsymbol{x_{t-1}}, \beta_t \mathbf{I}\right). \end{aligned} \tag{1}$$

The noise level is controlled by β_t, and $\boldsymbol{x_t}$ is sampled using $\boldsymbol{x_t} = \sqrt{\bar{\alpha}_t}\boldsymbol{x_0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is standard Gaussian noise, $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$.

Reverse Process This process is conditioned on the input point cloud P and image I, and aims to remove the noise added by the forward process gradually as shown in the bottom Figure 2. This process is to predict the noise and reconstruct the point cloud data distribution $\boldsymbol{x_0}$ from $\boldsymbol{x_T}$ based on the input point cloud P and image I. Mathematically, it can be expressed as the probability distribution $p_{\boldsymbol{\theta}}$:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x_{1:T}}|\boldsymbol{x_T}, P, I) &= \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, P, I), \\ p_{\boldsymbol{\theta}}(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, P, I) &= \mathcal{N}(\boldsymbol{x_{t-1}}; \boldsymbol{\mu_\theta}(\boldsymbol{x_t}, P, I, t), \sigma_t^2\mathbf{I}), \end{aligned} \tag{2}$$

where $\boldsymbol{\mu_\theta}$ is the predicted mean of the Gaussian noise:

$$\boldsymbol{\mu_\theta}(\boldsymbol{x_t}, P, I, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, P, I, t)\right), \tag{3}$$

where $\boldsymbol{\epsilon_\theta}$ is the main noise prediction network of SuperPC, which predicts the noise added in the forward process, guided by the three-level conditions in Section 4.1.

Loss Function We aim to minimize the difference between the predicted and actual noise added during the forward process based on the input point cloud P and image I. This is quantified using a mean squared error loss function $\mathcal{L}(\boldsymbol{\theta})$:

$$\mathbb{E}_{\boldsymbol{x_0}, t, \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}(\sqrt{\bar{\alpha}_t}\boldsymbol{x_0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, P, I, t)\|^2\right], \tag{4}$$

where $\boldsymbol{\epsilon}$ is the actual Gaussian noise and $\boldsymbol{\epsilon_\theta}$ is the network predicted noise. This loss function ensures that the model refines its ability and knowledge to accurately generate the high-quality point cloud from the noised data while being guided by the three-level conditions.

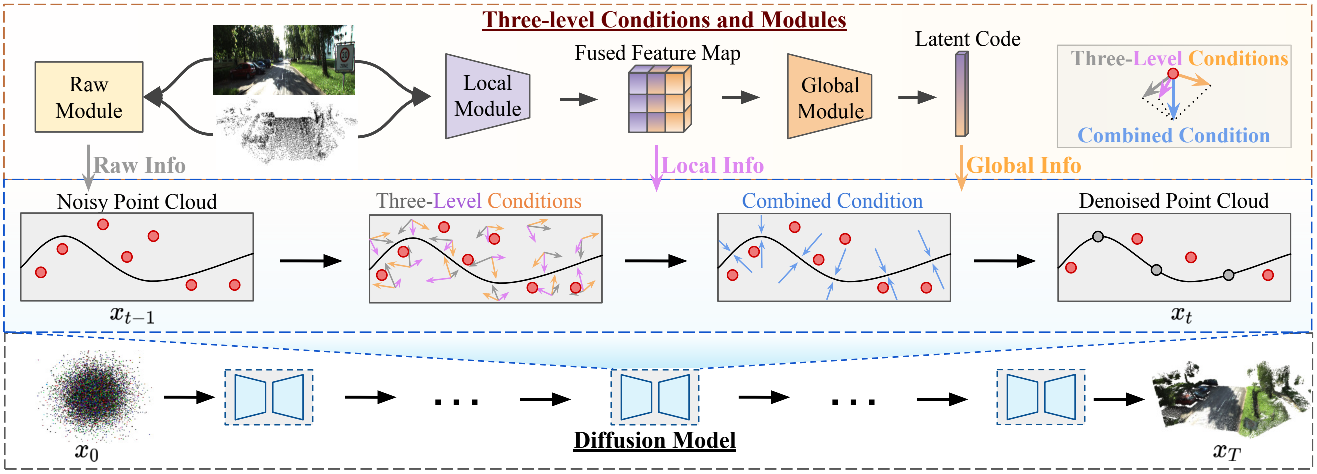


Figure 2. The architecture of the SuperPC model shown above integrates input images and point clouds to establish three-level conditions through innovative raw, local, and global modules. These conditions are seamlessly integrated into each step of the diffusion process, enabling SuperPC to utilize all levels of information from the two input modalities.

4. Methodology

As shown in Figure 2, we will explain the concept of the Three-Level-Conditioned (TLC) framework (middle of Figure 2, Section 4.1), and details of three novel modules that apply the Spatial-Mix-Fusion (SMF) strategy to fuse the two input modalities and generate three-level conditions (top of Figure 2, Section 4.2, Section 4.3, and Section 4.4).

4.1. TLC Framework & SMF Strategy

TLC Framework First, we discuss the TLC framework, which integrates raw, local, and global-level abstracted information as conditions to guide the diffusion model in simultaneously handling four point cloud processing tasks. Our rationale is based on the observation that relying solely on a single type of condition can easily trap the model in a local minimum or cause it to overfit to one particular task. For example, incorporating only global-level conditions into the diffusion model tends to produce basic shapes and lacks finer details, thereby hindering its capacity to generalize across more complex structures and expansive scenes [34]. To overcome this, we introduce three-level conditions into the diffusion model to enhance the model's performance and robustness to complex scenes and diverse tasks and get this new form of the loss function $\mathcal{L}(\boldsymbol{\theta})$:

$$\mathbb{E}_{\boldsymbol{x_0},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon_\theta}(\sqrt{\bar{\alpha}_t}\boldsymbol{x_0}+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \boldsymbol{c}_{\text{raw}}, \boldsymbol{c}_{\text{local}}, \boldsymbol{c}_{\text{global}}, t)\|^2\right], \quad (5)$$

where $\boldsymbol{c}_{\text{raw}}$, $\boldsymbol{c}_{\text{local}}$, and $\boldsymbol{c}_{\text{global}}$ represent the three-level conditions, collectively providing the combined instruction to guide the high-quality point clouds generating process.

SMF Strategy Previous works, whether employing early fusion [59, 66], deep fusion [67, 78], or late fusion [39, 44], have demonstrated strong performance in 3D object detection. However, we observed that these strategies are less effective for point cloud processing tasks. Specifically, late fusion in 3D object detection combines detection results from the image and point cloud branches at the final stage, making it unsuitable for point cloud processing tasks. Furthermore, as shown in our experiments (Appendix Table 3), both early and late fusion strategies show poor effectiveness across all three benchmarks. To address this challenge, we introduce the Spatial-Mix-Fusion (SMF) strategy, a loose coupling of early and deep fusion approaches. Specifically, we designed a **dual-spatial early fusion** and an **attention-based deep fusion** to integrate image and point cloud modalities into the diffusion conditions, both leveraging the spatial features of the partially denoised point cloud in the main diffusion network. We next delve into the specifics of constructing the TLC framework and the SMF strategy with the global, local, and raw modules.

4.2. Raw Module & Dual-Spatial Early Fusion

The raw module shown in Figure 3 (left) aims to preserve and fuse detailed texture and spatial information. It can be divided into two components: the image features projection and the point cloud features interpolation, which realize the dual-spatial early fusion. The two components are to incorporate and fuse the raw sensor information into each diffusion step to guide the target point cloud generating process, respectively. Specifically, the image branch utilizes the rasterization technique described in [38, 50] to project image features onto the partially denoised point cloud, while the point cloud branch uses spatial interpolation technique to align the input point cloud with the partially denoised point cloud in the main diffusion network.

Image Features Projection Before projecting the image to the point cloud, a normal 2D image network (like ResNet [58] or ViT [14]) is used to increase the image feature dimension from $I(H, W, 3)$ to $I'(H, W, C_1)$, providing richer raw RGB information. Then the computationally efficient `PointRasterizer` class of PyTorch3D is used to project the dense image features to the point cloud at the be-

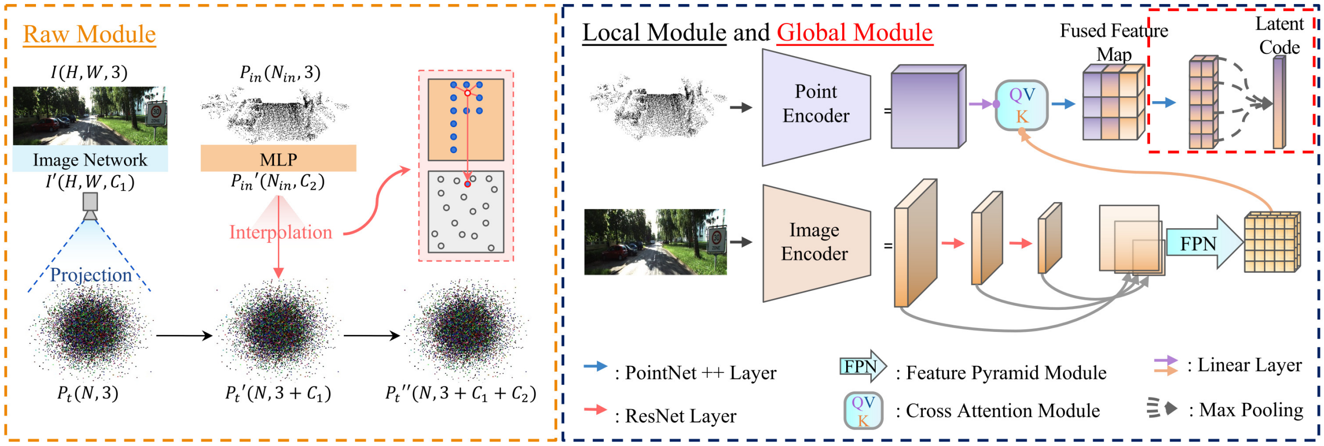


Figure 3. On the left, the raw module integrates the raw information of the image and point cloud into the target point cloud as the raw-level condition via the image projection and the point interpolation. On the right, the local module encodes the two inputs into feature maps, which are then fused using cross-attention to produce a local fused feature map as the local-level condition. Next, the global module condenses this feature map into a global latent code as the global-level condition.

ginning of each diffusion step. As a result, the point cloud dimension is increased from $P_t(N,3)$ to $P_t'(N,3+C_1)$. This step is crucial for roughly aligning the images and point clouds in terms of both raw spatial structural information and texture color information.

Point Cloud Features Interpolation Subsequently, we use a special spatial interpolation method to align the raw features of the input point cloud $P_{in}(N_{in},3)$ with the image-aligned features of $P_t'(N,3+C_1)$. Specifically, we first employ a multilayer perceptron (MLP) to increase the feature dimension of the input point cloud from $P_{in}(N_{in},3)$ to $P_{in}'(N_{in},C_2)$. Due to the disorder of the point cloud, we aggregated their features using the k-nearest neighbors based on the inverse-distance-weighted average algorithm [45]. Assuming the weighted features of the points in P_t' are $\boldsymbol{f}^{(j)}(j=1,2,...,N)$ and features of the k (here we use $k=4$) nearest points in P_{in}' are $\boldsymbol{f_i}^{(j)}(i=1,2,...,k)$, we can write the aggregated point features as:

$$\begin{aligned} &\boldsymbol{f}^{(j)} = \textstyle\sum_{i=1}^{k} w(\boldsymbol{x_j},\boldsymbol{x_i})\boldsymbol{f_i}^{(j)} / \textstyle\sum_{i=1}^{k} w(\boldsymbol{x_j},\boldsymbol{x_i}), \\ &w(\boldsymbol{x_j},\boldsymbol{x_i}) = 1/d(\boldsymbol{x_j},\boldsymbol{x_i})^2, \end{aligned} \tag{6}$$

where the $d(\boldsymbol{x_j},\boldsymbol{x_i})$ in the second equation indicates the Euclidean distance between points $\boldsymbol{x_j}$ and $\boldsymbol{x_i}$.

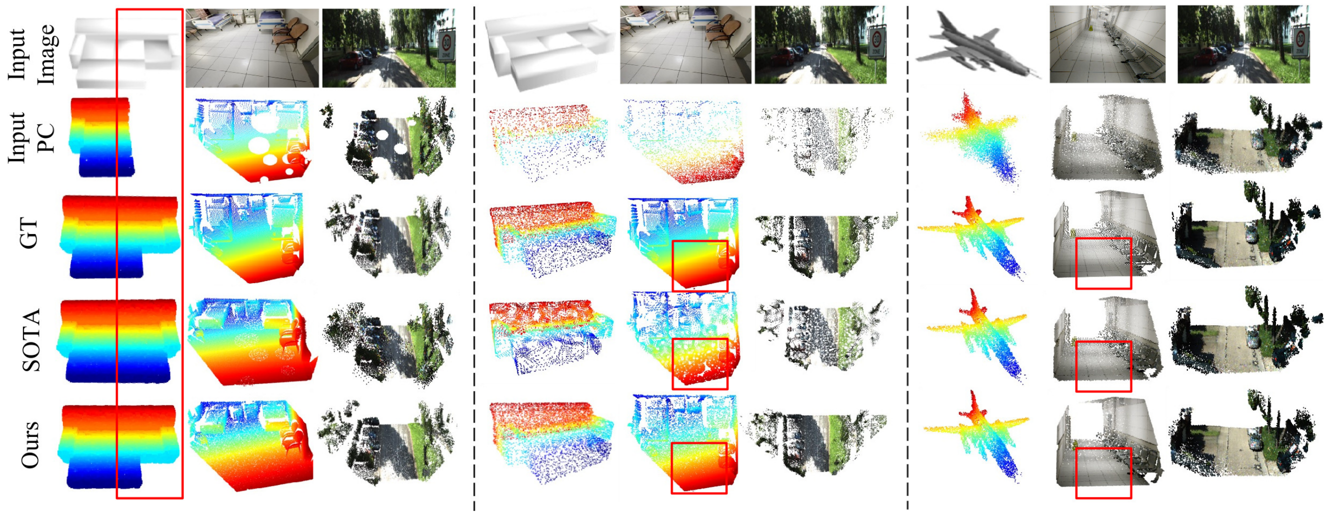
Dual-Spatial Early Fusion The two components described above result in a point cloud $P_t''(N,3+C_1+C_2)$ which successfully fuse the two modalities' raw-level information to the main diffusion network via the dual-spatial operations - (1) the spatial interpolation and (2) the point-rasterizer projection based on the spatial positions of the partially-denoised point cloud of the diffusion network.

4.3. Local Module & Attention-Based Deep Fusion

The local module shown in Figure 3 (right) is designed to introduce local object-level information to guide the diffusion model. Specifically, the local module merges two modalities into a fused feature map, which is then integrated into the main diffusion network as the local condition.

Inputs Encoding The local module uses the modality-specific encoders to extract features from the two inputs and fuse them into a local feature map based on cross-attention. Specifically, we use PointNet++ [45] and ResNet [58] as the point cloud and image encoders, respectively. We leverage the Feature Pyramid Networks (FPN) to map multi-scale spatial hierarchies to one comprehensive feature map for local representation. This is essential for capturing multi-scale semantic features, ensuring a detailed and multi-dimensional representation of the input visual information.

Attention-Based Deep Fusion Following the separate features extracting, the local module employs the cross-attention mechanism to fuse the images and point clouds. This fusion strategy is adapted from generative models [52, 57, 64], which are particularly beneficial for handling diverse types of conditioning inputs. The image features and point cloud features are transformed into a compatible feature space. The cross-attention mechanism is then applied to these intermediate representations, which can be formulated as: Attention$(Q,K,V) = \text{softmax}(QK/\sqrt{T}) \cdot V$, where Q represents the query tensor derived from image encoder's intermediate representation of the image feature map, while K and V are the key and value tensors obtained from the point encoder. These tensors are processed through learnable projection matrices $W_Q(f_{img})$, $W_K(f_{pc})$, and $W_V(f_{pc})$, which are passing the outputs of the image and the point encoders through three different linear layers to obtain Q, K and V, respectively. This cross-attention block aligns and integrates the point cloud spatial features with the multi-scale image semantic features, creating a unified fused feature map encoding both local spatial and visual features of the point cloud and the image.



(a) Completion. (b) Upsampling. (c) Denoising.

Figure 4. The qualitative results on the point cloud (a) completion, (b) upsampling, and (c) denoising tasks. For each subfigure, from left to right are the results for ShapeNet, TartanAir, and KITTI-360. Larger figures and more qualitative results are presented in Appendix E.

4.4. Global Module

Subsequently, a one-dimensional latent code representing global structural features is generated through a simple PointNet++ layer and a max pooling layer, a technique for emphasizing high-level features while reducing dimensionality. These two modules enable a seamless fusion of image and point cloud data, ensuring that the generated output effectively encapsulates the characteristics of both modalities. Consequently, they guide the diffusion process to extract and learn the structure details and the semantic consistency of the point clouds and images. We have included details of the implementation of the three modules in Appendix A.

5. Experiments

Benchmarks and Datasets Previous methods have mainly reported performance on the ShapeNet dataset [7]. However, it only contains simple 3D objects like chairs and planes without color, limiting its ability to evaluate like in scene-level scenarios and point cloud colorization. Therefore, We have designed three new benchmarks for point cloud processing tasks: one object-level (based on ShapeNet [7] dataset) and two scene-level benchmarks (based on TartanAir [63] and KITTI360 [31] datasets). Additionally, these benchmarks are organized into two evaluation tracks to assess generalization ability—Object-to-Scene and Sim-to-Real. The detailed setup of the three benchmarks is presented in Appendix B.

Evaluation Metrics For comprehensiveness, we select the most widely used metrics for point cloud processing including Earth Mover's Distance (EMD) and F1. Additionally, we also use Density-aware Chamfer Distance (DCD) [65] instead of the traditional Chamfer Distance (CD) as CD is sensitive to mismatched local density [65]. The specifics of the three metrics are detailed in Appendix C.

We next evaluate SuperPC against SOTA models across all four individual tasks and the combined task (Section 5.1), assess its generalization through object-to-scene and sim-to-real tests (Section 5.2).

5.1. Performance

We select the most well-known and recent point cloud processing models [21, 35, 37, 42, 47, 72] as well as their combined models to serve as our baselines (all the chosen models are the SOTAs for the individual tasks). All models were subjected to identical training-validation-testing settings.

As can be seen in Table 1, none of the baseline models can conduct all the tasks. Therefore, we integrate the best model for each task together to tackle the combination task and compare the integrated method with SuperPC. How we evaluate the best integration model and additional qualitative results can be found in Appendix D.4 and E. Remarkably, SuperPC surpasses all the models in all three tasks and their combination task across the three benchmarks. Besides the general quantitative results, we next delve into the details and qualitative results of each task.

Point Cloud Completion (PC) For the completion task, we follow the moderate sampling setting in [72] on the ShapeNet dataset, i.e., 50% of the points are used as the incompleted input point clouds. On the two scene-level benchmarks, we develop a simple random patch removal algorithm to generate point clouds with random blank areas as shown in Figure 4a and 7. Notably, as illustrated in the first row of Figure 4a, the best baseline method [72] inaccurately generates a completed symmetrical sofa, diverging from the actual asymmetrical shape of the ground truth. In contrast, SuperPC integrates information from the image input to accurately complete the asymmetrical shape.

Point Cloud Upsampling (PU) For this task, we set the in-

Table 1. Results compared with SOTA models on point cloud completion, upsampling, denoising, and combination tasks on the ShapeNet, TartanAir, and KITTI-360 benchmarks (from top to bottom). The cross mark "×" indicates that the model is incapable of performing the specified task. Different SOTA combinations, e.g., "[21]→[42]→[35]", are selected for different benchmarks because they are the top performers for that specific category. A more detailed comparison of these SOTA combinations can be found in Appendix D.4.

ShapeNet [7]	Completion			Upsampling			Denoising			Combination		
	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)
AdaPoinTr [72]	0.462	2.12	0.423	×	×	×	0.562	3.61	0.405	×	×	×
GradPU [21]	×	×	×	0.298	1.27	0.589	0.533	3.28	0.412	×	×	×
DDPMPU [47]	×	×	×	**0.281**	**1.13**	**0.659**	0.313	1.37	0.769	×	×	×
ScoreDenoise [35]	×	×	×	0.346	1.65	0.537	0.291	1.21	0.812	×	×	×
PD-LTS [37]	×	×	×	×	×	×	**0.280**	**1.12**	**0.856**	×	×	×
LiDiff [42]	0.484	2.43	0.401	0.312	1.38	0.526	×	×	×	×	×	×
[37]→[72]→[47]	0.462	2.12	0.423	0.281	1.13	0.659	0.280	1.12	0.856	0.509	2.71	0.382
SuperPC (ours)	**0.387**	**1.67**	**0.557**	0.293	1.25	0.631	0.285	1.16	0.837	**0.476**	**2.21**	**0.409**

TartanAir [63]	Completion			Upsampling			Denoising			Combination		
	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)
AdaPoinTr [72]	0.573	3.59	0.349	×	×	×	0.604	3.91	0.229	×	×	×
GradPU [21]	×	×	×	0.527	3.31	0.313	0.572	3.85	0.235	×	×	×
DDPMPU [47]	×	×	×	0.541	3.49	0.302	0.389	1.91	0.438	×	×	×
ScoreDenoise [35]	×	×	×	0.617	3.96	0.224	0.346	1.65	0.473	×	×	×
PD-LTS [37]	×	×	×	×	×	×	0.368	1.89	0.451	×	×	×
LiDiff [42]	0.559	3.51	0.356	0.562	3.62	0.248	×	×	×	×	×	×
[21]→[42]→[35]	0.559	3.51	0.356	0.527	3.31	0.313	0.346	1.65	0.473	0.583	3.64	0.369
SuperPC (ours)	**0.538**	**3.46**	**0.363**	**0.492**	**2.98**	**0.356**	**0.298**	**1.38**	**0.631**	**0.558**	**3.53**	**0.384**

KITTI-360 [31]	Completion			Upsampling			Denoising			Combination		
	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)
AdaPoinTr [72]	0.663	9.27	0.293	×	×	×	0.621	7.85	0.329	×	×	×
GradPU [21]	×	×	×	0.597	6.92	0.354	0.589	6.47	0.368	×	×	×
DDPMPU [47]	×	×	×	0.601	7.18	0.348	0.407	3.95	0.542	×	×	×
ScoreDenoise [35]	×	×	×	0.718	9.93	0.254	0.369	3.21	0.604	×	×	×
PD-LTS [37]	×	×	×	×	×	×	0.381	3.37	0.575	×	×	×
LiDiff [42]	0.649	9.03	0.307	0.609	7.35	0.341	×	×	×	×	×	×
[21]→[42]→[35]	0.649	9.03	0.307	0.597	6.92	0.354	0.369	3.21	0.604	0.725	10.06	0.347
SuperPC (ours)	**0.632**	**8.82**	**0.324**	**0.577**	**6.73**	**0.369**	**0.327**	**2.86**	**0.615**	**0.681**	**9.58**	**0.365**

teger testing upsampling rate as $\times 8$ for all three benchmarks to meet the inference limitation on the integer upsampling rate. Unlike the majority of previous upsampling methods [28, 69, 70], our model can take an arbitrary real number upsampling ratio like 1.3425. Moreover, Figure 4b reveals that the point clouds generated by the SOTA [21] exhibit an odd pattern of point concentration across all benchmarks. Noticeably, SuperPC surpasses the SOTA and baseline methods on both scene-level benchmarks, while DDPMPU [47] shows only marginally better performance on the object-level benchmark. Unlike previous works, SuperPC can produce not only precise but also evenly distributed point clouds with an arbitrary upsampling rate and shows dominating performance on the scene-level benchmarks.

Point Cloud Denoising (PD) In the denoising task, the accurate ground truth point clouds are perturbed by Gaussian noise with standard deviation from 0.5% to 2% of the bounding sphere radius to generate the input noisy data following the setting in [35]. The point clouds in the quantitative testing configuration are subjected to a noise level of 2%. As shown in the second and third rows of Figure 4c, the point cloud ground textures generated by SuperPC are noticeably less noisy compared to those produced by the SOTA point cloud denoising model - ScoreDenoise [35].

Combination In the combination task, we integrate the SOTA models [21, 35, 37, 42, 47, 72] for each individual task in sequence to tackle the challenging combination task. As illustrated in Figure 7, the combined approach with SOTA models falls short in achieving the final fully satisfactory map: despite successful upsampling, it leaves the rear window incomplete and introduces noise around the circled car. Conversely, SuperPC excels in producing more accurate point cloud maps, showcasing its enhanced capability to tackle complex combination tasks effectively.

Colorization The point cloud colorization task cannot be evaluated alongside the other three tasks for two main reasons: first, its metrics differ significantly from the other tasks and cannot be adjusted to align with their metrics; second, combining the colorization task with the others would expand the point cloud data space from three dimensions to

Tasks	Method	Object-to-Scene			Sim-to-Real		
		DCD (↓)	EMD (↓)	F1 (↑)	DCD (↓)	EMD (↓)	F1 (↑)
Completion	PC	0.901	29.8	0.122	0.873	19.39	0.156
	SuperPC (ours)	**0.857**	**24.2**	**0.159**	**0.789**	**13.93**	**0.225**
Upsampling	PU	**0.613**	**6.81**	**0.316**	0.693	10.42	0.279
	SuperPC (ours)	0.738	11.4	0.243	**0.648**	**8.96**	**0.336**
Denoising	PD	0.368	3.67	0.431	0.588	3.42	0.343
	SuperPC (ours)	**0.333**	**2.91**	**0.513**	**0.552**	**2.98**	**0.458**
Combination	PU→PC→PD	0.917	31.3	0.116	0.891	21.45	0.139
	SuperPC (ours)	**0.861**	**25.8**	**0.143**	**0.802**	**14.86**	**0.214**

Table 2. Generalization ability experiment on the four point cloud processing tasks. (5% data used for fine-tuning.)

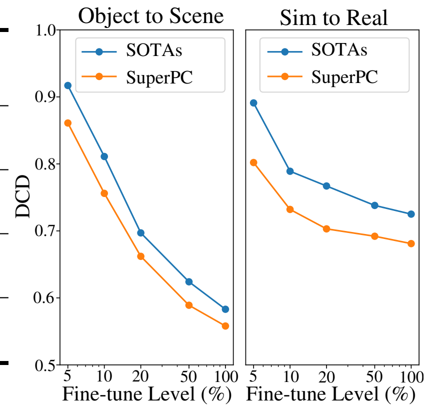


Figure 5. Combination task.

six, making it considerably more challenging for the diffusion model to learn the target data distribution. Therefore, we evaluate the colorization task separately using the same SuperPC model structure to demonstrate its effectiveness for this task. Details of the colorization task experiment can be found in the supplementary material Appendix D.5

5.2. Generalization Tests

In addition to performance assessments, we evaluated the generalization capabilities of SuperPC and current SOTA models across various point cloud processing tasks. These evaluations were conducted through two critical experiments: object-to-scene and sim-to-real. These experiments are pivotal for practical applications of all the four tasks.

Object-to-Scene Traditionally, research has predominantly concentrated on object-level datasets such as ShapeNet [7], often neglecting scene-level datasets. Yet, practical applications, such as indoor 3D reconstruction and urban digital twins, necessitate handling scene-scale point clouds. To address this, we trained SuperPC and various previous models on the object-level dataset (ShapeNet [7]) and subsequently fine-tuned and tested them on the scene-level dataset (TartanAir [63]). As shown in Table 2, SuperPC demonstrates superior generalization from object-scale to scene-scale date compared to most state-of-the-art models, particularly in the combination task.

Sim-to-Real Another significant challenge is the sim-to-real gap. Learning-based models frequently perform well in simulated environments but struggle when applied to real-world scenarios. Our study is the first to train point cloud processing models on a simulated dataset (TartanAir [63]) and fine-tune and test them on a real-world dataset (KITTI-360 [31]) to assess their capabilities in overcoming the sim-to-real gap. As illustrated in Table 2, SuperPC outshines all SOTA models across all the point cloud processing tasks.

Different Fine-tune Levels Furthermore, we fine-tuned both the integrated SOTA models and our SuperPC model using varying proportions of data from the TartanAir and KITTI-360 benchmarks to assess their generalization capabilities across both the object-to-scene and sim-to-real experiments. As depicted in Figure 5, increasing the fine-tuning levels enhances the performance of both SOTA models and SuperPC on the combination task. Notably, SuperPC consistently exhibits superior generalization ability across all levels of fine-tuning in both tracks.

5.3. Additional Experiments

In addition to the performance and generalization experiments, we present four more experiments in Appendix D to further support our work. These include (1) the ablation study, (2) the complexity analysis, (3) the comparison between single-model and multi-model approaches using the same SuperPC network design, and (4) an evaluation of different integration orders when combining SOTA methods.

6. Conclusion and Limitation

Conclusion We introduced SuperPC, a single diffusion model that excels in completion, upsampling, denoising, and colorization of point clouds. Leveraging the proposed TLC framework and SMF strategy, it outperforms specialized models in efficiency and effectiveness on complex scene-level datasets. SuperPC sets a new benchmark for generating high-quality point clouds, enhancing task interconnectivity, and paving the way for advanced 3D environmental analysis in future applications.

Limitation Nevertheless, SuperPC's efficiency can be further optimized by replacing the relatively heavy PointNet++ [45] backbone with a sparse-tensor-based backbone. Furthermore, establishing a real-world object-level benchmark for this task would be highly valuable and convincing. In future work, researchers could explore the performance of a unified model across varying degrees of incompleteness derived from realistic scanning patterns.

Acknowledgments This work was supported by the DARPA award HR00112490426. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. *Advances in Neural Information Processing Systems*, 35:37349–37362, 2022. 3

[2] Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robotics & Automation Magazine*, 19(3):80–91, 2012. 1

[3] Marc Alexa, Johannes Behr, Daniel Cohen-Or, Shachar Fleishman, David Levin, and Claudio T. Silva. Computing and rendering point set surfaces. *IEEE Transactions on visualization and computer graphics*, 9(1):3–15, 2003. 2

[4] Antonio Alliegro, Diego Valsesia, Giulia Fracastoro, Enrico Magli, and Tatiana Tommasi. Denoise and contrast for category agnostic shape completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4629–4638, 2021. 2

[5] YANG Bisheng, LIANG Fuxun, and HUANG Ronggang. Progress, challenges and perspectives of 3d lidar point cloud processing. *Acta Geodaetica et Cartographica Sinica*, 46 (10):1509, 2017. 1

[6] Ting On Chan, Hang Xiao, Lixin Liu, Yeran Sun, Tingting Chen, Wei Lang, and Ming Ho Li. A post-scan point cloud colorization method for cultural heritage documentation. *ISPRS International Journal of Geo-Information*, 10(11):737, 2021. 2

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 6, 7, 8, 3, 4

[8] Sunyoung Cho, Jizhou Yan, Yasuyuki Matsushita, and Hyeran Byun. Efficient colorization of large-scale point cloud using multi-pass z-ordering. In *2014 2nd International Conference on 3D Vision*, pages 689–696. IEEE, 2014. 2

[9] Jaesung Choe, Byeongin Joung, Francois Rameau, Jaesik Park, and In So Kweon. Deep point cloud reconstruction. *arXiv preprint arXiv:2111.11704*, 2021. 1

[10] 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. 3

[11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[12] Nathan Crombez, Guillaume Caron, and E Mouaddib. 3d point cloud model colorization by dense registration of digital images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40: 123–130, 2015. 2

[13] Lucía Díaz-Vilariño, Kourosh Khoshelham, Joaquín Martínez-Sánchez, and Pedro Arias. 3d modeling of building indoor spaces and closed doors from imagery and point clouds. *Sensors*, 15(2):3491–3512, 2015. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Wanquan Feng, Jin Li, Hongrui Cai, Xiaonan Luo, and Juyong Zhang. Neural points: Point cloud representation with neural fields for arbitrary upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18633–18642, 2022. 2

[16] Rongrong Gao, Tian-Zhu Xiang, Chenyang Lei, Jaesik Park, and Qifeng Chen. Scene-level point cloud colorization with semantics-and-geometry-aware networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2818–2824. IEEE, 2023. 3

[17] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018. 1

[18] Shuo Gu, Yigong Zhang, Jinhui Tang, Jian Yang, Jose M Alvarez, and Hui Kong. Integrating dense lidar-camera road detection maps by a multi-modal crf model. *IEEE Transactions on Vehicular Technology*, 68(12):11635–11645, 2019. 3

[19] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020. 1

[20] Zhizhong Han, Honglei Lu, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. 3d2seqviews: Aggregating sequential views for 3d global feature learning by cnn with hierarchical attention aggregation. *IEEE Transactions on Image Processing*, 28(8): 3986–3999, 2019. 2

[21] Yun He, Danhang Tang, Yinda Zhang, Xiangyang Xue, and Yanwei Fu. Grad-pu: Arbitrary-scale point cloud upsampling via gradient descent with learned distance functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2023. 1, 2, 6, 7, 3, 4, 5

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

[23] Jong-Uk Hou, Baoquan Zhao, Naushad Ansari, and Weisi Lin. Range image based point cloud colorization using conditional generative model. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 524–528. IEEE, 2019. 3

[24] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022. 2, 3

[25] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7662–7670, 2020. 2

[26] Franz Leberl, Arnold Irschara, Thomas Pock, Philipp Meixner, Michael Gruber, Set Scholz, and Alexander Wiechert. Point clouds. *Photogrammetric Engineering & Remote Sensing*, 76(10):1123–1134, 2010. 1

[27] In-Kwon Lee. Curve reconstruction from unorganized points. *Computer aided geometric design*, 17(2):161–177, 2000. 2

[28] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7203–7212, 2019. 2, 7

[29] You Li and Javier Ibanez-Guzman. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4):50–61, 2020. 1

[30] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 3

[31] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 1, 6, 7, 8, 2, 3, 4, 5

[32] Jitao Liu, Songmin Dai, and Xiaoqiang Li. Pccn: Point cloud colorization network. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3716–3720. IEEE, 2019. 1, 3, 4, 6

[33] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 1, 3

[34] 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. 1, 3, 4

[35] Shitong Luo and Wei Hu. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4583–4592, 2021. 1, 2, 3, 6, 7, 4, 5

[36] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 1, 3

[37] Aihua Mao, Biao Yan, Zijing Ma, and Ying He. Denoising point clouds in latent space via graph convolution and invertible neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5768–5777, 2024. 6, 7

[38] 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. 1, 3, 4

[39] Gledson Melotti, Cristiano Premebida, Nuno MM da S Gonçalves, Urbano JC Nunes, and Diego R Faria. Multimodal cnn pedestrian classification: a study on combining lidar and camera data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3138–3143. IEEE, 2018. 3, 4

[40] Duc Thanh Nguyen, Binh-Son Hua, Khoi Tran, Quang-Hieu Pham, and Sai-Kit Yeung. A field model for repairing 3d shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5684, 2016. 2

[41] Nhung Hong Thi Nguyen, Stuart Perry, Don Bone, Ha Le Thanh, Min Xu, and Thuy Thi Nguyen. Combination of images and point clouds in a generative adversarial network for upsampling crack point clouds. *IEEE Access*, 10: 67198–67209, 2022. 3

[42] Lucas Nunes, Rodrigo Marcuzzi, Benedikt Mersch, Jens Behley, and Cyrill Stachniss. Scaling diffusion models to real-world 3d lidar scene completion. *arXiv preprint arXiv:2403.13470*, 2024. 6, 7, 4, 5

[43] Teddy Ort, Liam Paull, and Daniela Rus. Autonomous vehicle navigation in rural environments without detailed prior maps. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2040–2047. IEEE, 2018. 1

[44] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10386–10393. IEEE, 2020. 3, 4

[45] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5, 8, 1, 3

[46] Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, and Liang Xiao. A conditional denoising diffusion probabilistic model for point cloud upsampling. *arXiv preprint arXiv:2312.02719*, 2023. 2, 3

[47] Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, and Liang Xiao. A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20786–20795, 2024. 2, 6, 7, 1, 3

[48] Marie-Julie Rakotosaona, Vittorio La Barbera, Paul Guerrero, Niloy J Mitra, and Maks Ovsjanikov. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer graphics forum*, pages 185–203. Wiley Online Library, 2020. 2

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[50] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 4

[51] Fabio Remondino. From point cloud to surface: the modeling and visualization problem. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34, 2003. 1

[52] 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. 5

[53] Riccardo Roveri, A Cengiz Öztireli, Ioana Pandele, and Markus Gross. Pointpronets: Consolidation of point clouds with convolutional neural networks. In *Computer Graphics Forum*, pages 87–99. Wiley Online Library, 2018. 1, 2

[54] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *2011 IEEE international conference on robotics and automation*, pages 1–4. IEEE, 2011. 1

[55] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 1, 3

[56] Brent Schwarz. Mapping the world in 3d. *Nature Photonics*, 4(7):429–430, 2010. 1

[57] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. 5

[58] Takayuki Shinohara, Haoyi Xiu, and Masashi Matsuoka. Point2color: 3d point cloud colorization using a conditional generative network and differentiable rendering for airborne lidar. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1062–1071, 2021. 1, 3, 4, 5

[59] Martin Simon, Karl Amende, Andrea Kraus, Jens Honer, Timo Samann, Hauke Kaulbersch, Stefan Milz, and Horst Michael Gross. Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 4

[60] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 808–816, 2016. 3

[61] Minhyuk Sung, Vladimir G Kim, Roland Angst, and Leonidas Guibas. Data-driven structural priors for shape completion. *ACM Transactions on Graphics (TOG)*, 34(6): 1–11, 2015. 2

[62] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 3

[63] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 6, 7, 8, 2, 3, 4, 5

[64] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 5

[65] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*, 2021. 6, 2

[66] Liang Xie, Chao Xiang, Zhengxu Yu, Guodong Xu, Zheng Yang, Deng Cai, and Xiaofei He. Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12460–12467, 2020. 3, 4

[67] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 244–253, 2018. 3, 4

[68] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019. 2

[69] Wang Yifan, Shihao Wu, Hui Huang, Daniel Cohen-Or, and Olga Sorkine-Hornung. Patch-based progressive 3d point set upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5958–5967, 2019. 2, 7

[70] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 1, 2, 7

[71] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021. 2

[72] Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers. *arXiv preprint arXiv:2301.04545*, 2023. 1, 2, 6, 7, 3, 4

[73] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018. 1, 5

[74] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3

[75] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, pages 1–9. Berkeley, CA, 2014. 1

[76] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1768–1777, 2021. 2

[77] Xuancheng Zhang, Yutong Feng, Siqi Li, Changqing Zou, Hai Wan, Xibin Zhao, Yandong Guo, and Yue Gao. View-guided point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15890–15899, 2021. 3

[78] Zehan Zhang, Yuxi Shen, Hao Li, Xian Zhao, Ming Yang, Wenming Tan, ShiLiang Pu, and Hui Mao. Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion. In *2022 IEEE 25th International conference on intelligent transportation systems (ITSC)*, pages 369–376. IEEE, 2022. 3, 4

[79] Zhipeng Zhao, Huai Yu, Chenwei Lyu, Wen Yang, and Sebastian Scherer. Attention-enhanced cross-modal localization between spherical images and point clouds. *IEEE Sensors Journal*, 2023. 3

[80] 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

[81] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1