TouchDream: 3D Object Completion through Imagined Touch

Yuanbo Wang[1] Xining Wang[1] Zhaoxuan Zhang[2] Changlong Wang[1]
Qianchen Xia[3] Xiaopeng Wei[1*] Xin Yang[1*]
[1] Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[2] Nanjing University of Posts and Telecommunications [3] Tsinghua University
wangyuanbo@mail.dlut.edu.cn xinning_w@163.com zhangzx@njupt.edu.cn
changlongwang@mail.dlut.edu.cn qianchenxia@tsinghua.edu.cn {xpwei, xinyang}@dlut.edu.cn

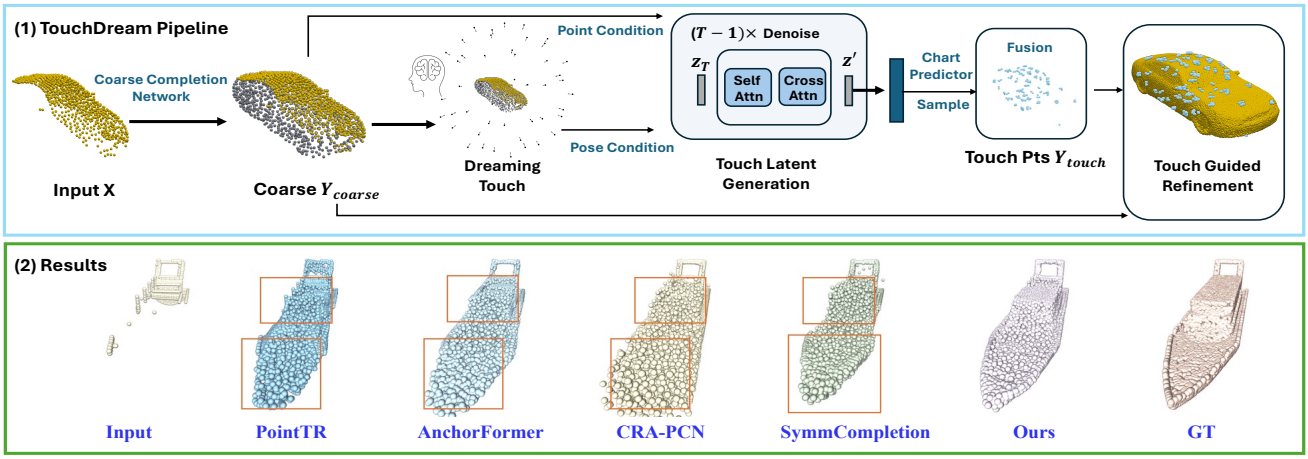

Figure 1. (1) TouchDream Workflow: generating imagined touch for coarse points refinement. (2) Completion Results Comparison.

Abstract

Point cloud completion is crucial for robust 3D perception but remains challenging. Coarse-to-fine methods can lead to unconstrained local guesses in the absence of key structures, whereas diffusion-based approaches may introduce geometric inconsistencies. To overcome these limitations, we present TouchDream, a novel framework that leverages a diffusion model to 'dream' of tactile sensing on object surfaces, which reformulates the sensing process as a learnable generative modeling task. Unlike visual cues, tactile data provides rich local geometry that can be directly converted into 3D space for point fusion, offering a powerful guide for detail-aware completion. Specifically, our approach generate compact tactile latent representations conditioned on coarse points and sampled touch poses. These generated touches are then used to optimize the coarse geometry. Extensive experiments show that our TouchDream model achieves the state-of-the-art performance.

*Corresponding authors.

1. Introduction

3D perception is crucial for a wide range of applications, from robotic manipulation and autonomous navigation to augmented reality. As a compact and efficient representation of 3D objects, point clouds are extensively used. However, in real-world scenarios, point clouds acquired by sensors are often incomplete due to occlusion, limited sensor resolution, and adverse viewing angles. Point cloud completion, the task of inferring a complete and dense geometry from a partial scan, is therefore a fundamental step for downstream 3D understanding tasks.

Nevertheless, this task is inherently challenging and ill-posed, as it requires not only the reconstruction of the overall global structure but also the generation of plausible and detailed geometry for the missing regions. Current mainstream learning-based point cloud completion methods can be broadly categorized into two paradigms. The first paradigm operates in a coarse-to-fine manner, where an initial coarse shape is first generated. This is followed by a refinement stage that leverages multi-granularity features

from both the coarse and fine point clouds [15, 27, 30, 40], or utilizes geometric symmetries [34, 42], to guide the detailed reconstruction of the missing regions. However, the absence of key structures in the input point cloud can lead to an unconstrained local guess, which has been documented in prior work [29]. The second category employs generative models, such as diffusion models, which either directly generate the completed point cloud through a probabilistic denoising process [3, 10, 14, 48] or produce auxiliary cues to promote the reconstruction of missing areas [10, 29, 49, 50]. These studies confirm that effective fusion of external signals substantially boosts completion quality, while potentially introducing geometric consistency challenges.

In this work, we enhance point cloud completion by generating tactile information. Specifically, we use a diffusion model conditioned on the coarse points and sampled touch poses to generate tactile information. This tactile information subsequently guides the refinement stage to produce a complete and detailed point cloud. Our approach is motivated by two key advantages of tactile over multi-view visual data. First, as established in prior work [5, 22, 23, 28], tactile sensing provides high-fidelity local 3D shape information (such as spatial contact points and fine surface details) which is critical for reconstructing local geometry. Second, the local point clouds decoded from tactile signals can be directly fused with the coarse point cloud, enabling more straightforward and effective point cloud completion and optimization. Previous work [5, 23, 28] required multiple tactile data sensing steps towards the target object. Particularly in real-world scenarios, such concact could introduce additional risks of damage and safety hazards, often necessitating the object to be secured in place to prevent movement. To this end, we propose TouchDream, a novel framework that leverages a diffusion model to 'dream' of tactile sensing on object surfaces. By 'dreaming' of plausible tactile signals, we can provide extra guidance to the completion network. This approach shifts the completion paradigm from purely extrapolating global visual patterns to incorporating imagined local tactile information.

However, generating tactile data based on coarse points is a non-trivial problem. The primary challenge lies in the extreme variability and complexity of 3D shapes. The tactile distribution across an object's surface is highly heterogeneous and is intimately tied to its local curvature properties. To circumvent the difficulty of direct 3D generation or tactile image generation, we address this problem in a latent space. We employ a diffusion model to generate compact tactile latent vectors from coarse points and touch poses. The generated latent vectors are then decoded to predict local 3D geometry, which is subsequently transformed into the world coordinate system. The sampled touch points is finally used to refine the coarse points to achieve final completion (Figure 1).

Extensive experiments validate the effectiveness of our approach, demonstrating significant performance improvements on point cloud completion task. The main contributions are as follows: (1) We propose TouchDream, the first method to refine coarse point clouds by generating tactile information. (2) We introduce a diffusion model conditioned on a coarse point cloud to generate tactile data. This model learns latent vectors of tactile representations based on the sampled touch poses. (3) We further leverages the generated tactile information to refine the point cloud and achieves state-of-the-art completed results on multiple benchmark datasets.

2. Related Work

Partial point cloud-based shape completion. Point cloud completion [7, 11, 17, 21, 43] has evolved from geometry-driven methods based on geometric priors to data-driven learning approaches. Among these, PCN [40] established the coarse-to-fine framework, inspiring a series of two-stage successors [15, 27, 30]. Subsequent efforts have focused on enhancing fine-grained detail recovery through specialized architectures. CRA-PCN [20] and AnchorFormer [2] leverage cross-resolution interactions and anchor-based transformers to improve multi-scale and region-aware feature representation. SeedFormer [47] and SnowflakeNet [32] propose novel upsampling strategies via patch seeds and hierarchical splitting, respectively. PoinTr [37], in contrast, reformulates point cloud completion as a set-to-set translation task. SymmCompletion [34] incorporates symmetry priors to mitigate geometric inconsistencies, while LAKe-Net [26] addresses topological challenges using skeletal structure guidance.

Recently, diffusion models [8, 24] have become a powerful paradigm in 3D vision [4, 28, 38, 45]. Beginning with PVD's [48] hybrid point-voxel representation, subsequent works have expanded its application, including NSDS [3] for zero-shot completion, 3DQD [14] for part-discretized priors, and SDS-Complete [10] for text-guided generalization. However, the performance of these methods remains fundamentally limited by a lack of supplementary information, resulting in poor reconstructions for severely incomplete inputs and an inability to recover fine local details.

Partial point completion with additional data. To address the inherent ambiguity and severe geometric sparsity of partial point clouds, recent works increasingly integrate auxiliary data to enhance completion performance. In single-view visual fusion, MVCN [9] employs depth images rendered from 3D shapes to perform 2D-domain completion via a conditional GAN. ViP [41] fuses a single-view RGB image with the partial point cloud to extract global structural priors, which are then integrated with local 3D details and camera pose information to improve reconstruction fidelity. Similarly, CSDN [31] leverages RGB images

to capture fine-grained intrinsic shape features and refine the completed geometry. For multi-view settings, SVDFormer [49] utilizes multi-view depth images of the partial input to generate global shape hypotheses, synthesizing new points through learned shape priors and geometric constraints. PointSea [50] interprets incomplete shapes via self-projected multi-view cues and explores self-structure augmentation to enable global-to-local completion. More recently, PCDreamer [29] encodes global–local shape information using a multi-view diffusion prior (which generates a set of multi-view images) and employs a shape fusion module to suppress diffusion-induced noise, significantly enhancing detail recovery.

Beyond visual cues, incorporating high-level semantic or physical priors has also proven effective. For instance, Kasten *et al*. [10] use textual descriptions of objects together with partial point clouds to guide the reconstruction process. Yang *et al*. [36] inject category-level semantics via a semantic segmentation branch, while Ren *et al*. [19] exploit intra-class geometric regularities, such as symmetry and shape similarity among vehicles, as supplementary cues. Collectively, these studies demonstrate that the effective fusion of external signals with partial inputs substantially improves completion quality, particularly for complex objects whose observed geometry alone is insufficient.

Touch-based shape reconstruction. Touch-based reconstruction leverages tactile signals to recover object geometry in visually challenging scenarios. Tactile data is used to map into 3D patches or combined with other modalities to provide auxiliary information. For example, Li *et al*. [13] integrated force-torque data to reconstruct curved surfaces. TouchRoller [1] reconstructs object surface geometry through continuous tactile scanning. Touch-GS [25] combines monocular depth and tactile information to produce a single set of training images that combine touch and vision for 3D Gaussian Splatting. TAPCNet [16] fuses tactile points and incomplete point cloud information in the feature domain for point completion. Zhong *et al*. [46] employ NeRF-rendered RGB-D images as inputs to a conditional GAN for generating tactile images at target orientations, but require extensive RGB data for NeRF training. In contrast, ours needs only a coarse point cloud to generate touch latent vectors for sampled contact poses.

Recent advances in deep learning have facilitated 3D object reconstruction from sparse tactile observations. Existing methods typically employ high-resolution visuo-tactile sensors, such as Gelsight [39] and Digit [12], to capture tactile images either from randomly sampled surface locations [5, 22] or via learned exploration policies that strategically select contact points [23, 28]. The acquired tactile data supports accurate recovery of local geometry, thereby refining fine-grained structural details. These methods require physical interaction in either simulated or real-world environments to collect tactile images. In contrast, we accomplish this by "dreaming" of tactile sensing, which reformulates the sensing process as a learnable generative modeling task. This allows us to optimize a coarse point cloud using generated tactile information from poses sampled on coarse point clouds.

3. Method

As shown in Figure 1, our method completes 3D objects by 'dreaming' of tactile sensing on objects to refine coarse points. At the inference stage, we first produce a coarse point cloud Y_{coarse} from the partial input X using a pre-trained LSTNet from SymmCompletion [34]. We then sample tactile poses, and for each pose, conditionally generate a touch latent code, using both the pose and coarse point features, and decode it into local points. These points generated from different poses are merged and sampled to form Y_{touch} for coarse points refinement.

Figure 2 illustrates the architecture of our touch generation model, which comprises three key components: sampling touch poses from coarse points (Figure 2a and Section 3.1), a touch encoder paired with a chart predictor to encode touch signals into latent vectors and decode them into local tactile points (Figure 2b and Section 3.1), and a conditional diffusion model that generates touch latents from the poses and coarse features (Figure 2c and Section 3.2). The generated tactile points are merged and utilized to guide the coarse point completion (Section 3.3). In the following sections, we describe each of these components in detail.

3.1. Touch Latents

Generating tactile data from coarse points presents two main challenges. First, the generation process is complex due to the fact that tactile interactions sample from the entire 3D space. Tactile information is inherently tied to local curvature properties and points generated from different touch poses exhibit significant coordinate variations. Second, our objective involves generating tactile data on the completed surface using poses sampled from coarse points. This creates a training paradigm where the model uses poses derived from the coarse point cloud, while the corresponding tactile signals originate from ground-truth shapes, necessitating compensation for pose discrepancies.

To address the first challenge, we map touches to 1D latent vectors and subsequently utilize a diffusion model to effectively learn and sample from the distribution of latents. Specifically, we train a touch encoder to extract latent representations and then employ a local chart prediction network [22, 23] for local shape inference. The local chart, comprising mesh surface elements, is initialized from a base triangular mesh with 25 vertices and faces. This base mesh serves as a canonical template, with the network trained to predict deformed 3D positions for all vertices while main-

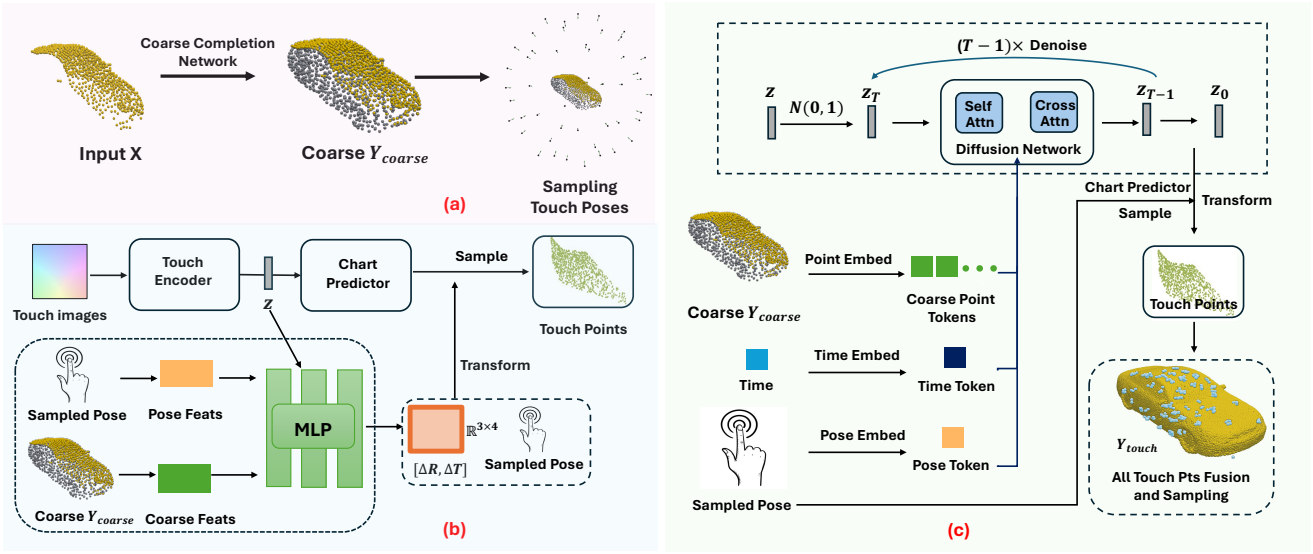

Figure 2. Our framework (a) leverages a pre-trained coarse completion network to initialize a point cloud for touch pose sampling, (c) trains a conditional diffusion model to generate corresponding touch latents, (b) decodes them into local geometries via a pre-trained predictor, and finally refines the output through integrated and sampled touch points.

taining fixed face connectivity. The network processes tactile images by encoding them into a latent vector z, then decodes this vector to predict 3D coordinates of local chart vertices. These coordinates are subsequently transformed into the world coordinate system using the rotation and translation from the samples poses. For pose sampling, we follow the framework used in [23, 28]. First, we define the interaction action space by uniformly sampling 50 discrete points on a spherical surface centered at the object's centroid. Then, the corresponding contact points are identified as the intersection points between the approach rays and the object's convex hull. Notably, not all sampled poses yield valid interaction outcomes: approach rays that fail to intersect the convex hull generate no usable tactile signals. We utilize only valid poses for both training and inference.

For the second challenge, we introduce an additional network that predicts the transformation matrix to align point clouds generated from coarse-point-sampled poses with ground-truth completed surfaces. As illustrated in Figure 2b, we extract features from both coarse points and sampled poses. These features are concatenated with the encoded latent vector z and processed through MLPs to predict rotation and translation parameters. The resulting transformation is applied to the predicted chart coordinates to achieve alignment with ground-truth surfaces. Local points are sampled on the chart surface.

During training, we construct a dataset as follows: for each object, we render tactile images at every valid touch pose using the object's mesh. Additionally, we sample corresponding touch poses from the coarse point cloud. For each such pose, the touch image is encoded, and the local chart prediction network generates a local point cloud, which is then transformed into world coordinates using the predicted transformation. The resulting points are compared with the ground-truth surface points via the Chamfer Distance (CD) to supervise the training of both the transformation network and the touch encoder-decoder.

3.2. Conditioned Diffusion

Next, as illustrated in Figure 2c, we utilize the sampled latent vectors z as sample space for probabilistic model Ω. During training, Gaussian noise is incrementally added to the latent vectors across random timesteps, and the model learns to reconstruct the original vectors conditioned on both the sampled touch poses and the initial coarse point cloud Y_{coarse}. To enable this conditioning, we implement two feature extractors: a point embedding module π_Y that extracts features from Y_{coarse} and MLP π_p that encodes features from each valid touch pose p. These conditioned features guide the diffusion model's denoising process. The training procedure involves sampling a timestep t and noise to obtain z_t from input latent vector z_0. The model is then trained to reconstruct z_0 using the following loss function:

$$L_{diff} = \|\Omega(z_t, \gamma(t)|\pi_Y(Y_{\text{coarse}}), \pi_p(p)) - z_0\|_2, \quad (1)$$

where $\gamma(t)$ is the tiemstamps embeddings, $\|\cdot\|_2$ is MSE loss. Point embedding module first partitions the point cloud into local patches via Farthest Point Sampling (FPS) and K-Nearest Neighbors (KNN), then aggregates each patch into a compact feature vector through an MLP and max pooling.

The diffusion model Ω builds upon the DiffusionSDF [4] framework, incorporating DALLE-2 [18] architecture with additional cross-attention layers in each block. At inference time, the conditioned diffusion model performs iterative generation starting from Gaussian noise $z_T \sim \mathcal{N}(0, 1)$:

$$z' = (g \circ \ldots \circ g)(z_T, T, \pi_Y(Y_{\text{coarse}}), \pi_p(p)), \quad (2)$$

where the denoising function g is defined as:

$$g(x_t, t, \pi_Y(Y_{\text{coarse}}), \pi_p(p)) = \Omega(x_t, \gamma(t)|\pi_Y(Y_{\text{coarse}}), \pi_p(p)) + \sigma_t \epsilon, \quad (3)$$

with σ_t being the fixed standard deviation at timestep t and $\epsilon \sim \mathcal{N}(0, 1)$.

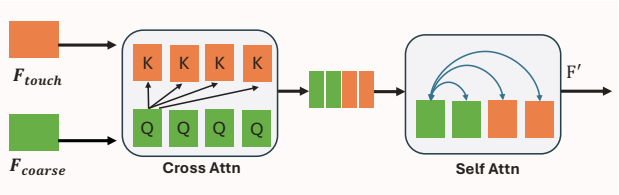

Figure 3. Feature Fusion for Tactile and Coarse Point Cloud.

3.3. Touch-Guided Refinement

In this stage, our goal is to refine the initial coarse point cloud Y_{coarse} using the generated touch point clouds Y_{touch}. The reason is that tactile signals provide valuable local 3D shape information (including spatial contact points and detailed geometric features) that enhances the reconstruction of fine-grained local details, which has been demonstrated in previous work [22, 23, 28].

For each sampled pose, we perform inference with the diffusion model by denoising z_T over T steps to obtain the latent code z_0. Using the pre-trained Chart Predictor and MLPs (Figure 2b), we then generate a local chart and predict its transformation parameters. Each pose yields one chart, which is transformed into the world coordinate system according to its pose. All transformed charts are subsequently merged, and points are randomly sampled to form a final point cloud composed of scattered tactile points.

To guide coarse point refinement, we adopt a network architecture similar to SymmCompletion [34] and leverage tactile features as guidance. We first use multiple MLPs and point transformers [44] to extract features from the tactile point cloud. Cross-attention and self-attention mechanisms are then employed to effectively incorporate tactile cues into the refinement process, aggregating geometric features from the coarse points with the imagined tactile features. The fused representations are iteratively fed into an upsampling network composed of self-attention layers and a point-shuffle module, which predicts per-point offsets to generate a denser and more detailed point cloud.

The loss function for point completion is total Chamfer Distance between the initial and each finer output:

$$L = \sum_{i=1}^{n} L_{CD}(Y_{gt}, Y_i) \quad (4)$$

where Y_{gt} is the ground truth point cloud, Y_i is the i-th finer output and n is the the number of Touch Guidance Blocks, respectively.

4. Experiment

In this section, we begin by presenting the implementation details and experimental setup. We then perform experiments on several datasets to evaluate our method with previous approaches, assessing both quantitative and qualitative performance in point completion task. Finally, we perform experiments specifically designed to validate whether our proposed approach is indeed necessary.

4.1. Experimental Settings

Implementation details. In the latent touch training stage, for each target object in the dataset, we used up to 7 tactile image and point cloud pairs, trained for 500 epochs with the Adam optimizer and a initial learning rate of 1e-5. For the diffusion model training, we conducted 500 epochs using the Adam optimizer with a initial learning rate of 2e-4. In the touch-guided refinement stage, we employed the AdamW optimizer with a initial learning rate of 2e-4 for 420 epochs. All training procedures were performed on a single RTX 4090 GPU with a batch size of 16. For touch points fusion and sampling, we aggregated points generated from valid poses among 50 predefined locations and randomly sampled 256 points.

Datasets. We evaluate our method on the standard PCN and ShapeNet55/34 benchmarks for point cloud completion. The PCN dataset [40], derived from ShapeNet core, is an established point cloud completion benchmark containing 30,974 models across 8 categories with partial-complete point cloud pairs. In comparison, the ShapeNet55 dataset offers a broader testbed with approximately 55,000 models spanning 55 categories, while the ShapeNet34 dataset [37] provides 34 seen categories and 21 unseen categories to validate the model's generalization capabilities. For touch data rendering, we utilize the object mesh and employ the simulation environment from [23] and simulate tactile interactions via poking to render tactile data of the target object at 50 predefined positions. The objects are normalized with their centers aligned at the origin. During the collection of tactile data and input into the diffusion model, they are scaled down by a factor of 3.1, and will be rescaled back to their original size after generating the tactile point cloud.

Method	Plane	Cabinet	Car	Chair	Lamp	Couch	Table	Boat	CD-Avg (↓)	F1 (↑)
PCN [40]	5.50	22.70	10.63	8.70	11.00	11.34	11.68	8.59	9.64	0.70
PoinTr [37]	4.75	10.47	8.68	9.39	7.75	10.93	7.78	7.29	8.38	0.75
SnowflakeNet [32]	4.29	9.16	8.08	7.89	6.07	9.23	6.55	6.40	7.21	0.80
FBNet [35]	3.99	9.05	7.90	7.38	5.82	8.85	6.35	6.18	6.94	-
SeedFormer [47]	3.85	9.05	8.06	7.06	5.21	8.85	6.05	5.58	6.74	0.82
SVDFormer [49]	3.62	8.79	7.46	6.91	5.33	8.49	5.90	5.83	6.54	0.84
AnchorFormer [2]	3.70	8.94	7.57	7.05	5.21	8.40	6.03	5.81	6.59	0.83
CRA-PCN [20]	3.59	8.70	7.50	6.70	5.06	8.24	5.72	5.64	6.39	-
PCDreamer [29]	3.52	8.72	6.89	6.71	5.64	8.32	6.24	5.84	6.49	0.86
PointSea [50]	3.52	8.54	7.33	6.58	5.21	8.24	5.75	5.62	6.35	0.86
SymmCompletion [34]	3.53	8.49	7.30	6.52	5.06	8.23	5.64	**5.49**	6.28	0.85
Ours	**3.39**	**7.93**	**7.16**	**6.34**	**4.91**	**7.83**	**5.43**	5.53	**6.05**	**0.87**

Table 1. Quantitative results on the PCN dataset (L1 CD $\times 10^3$, and F1-Score@1%). We show the Chamfer Distance for each category while reporting the average on the metrics on the right.

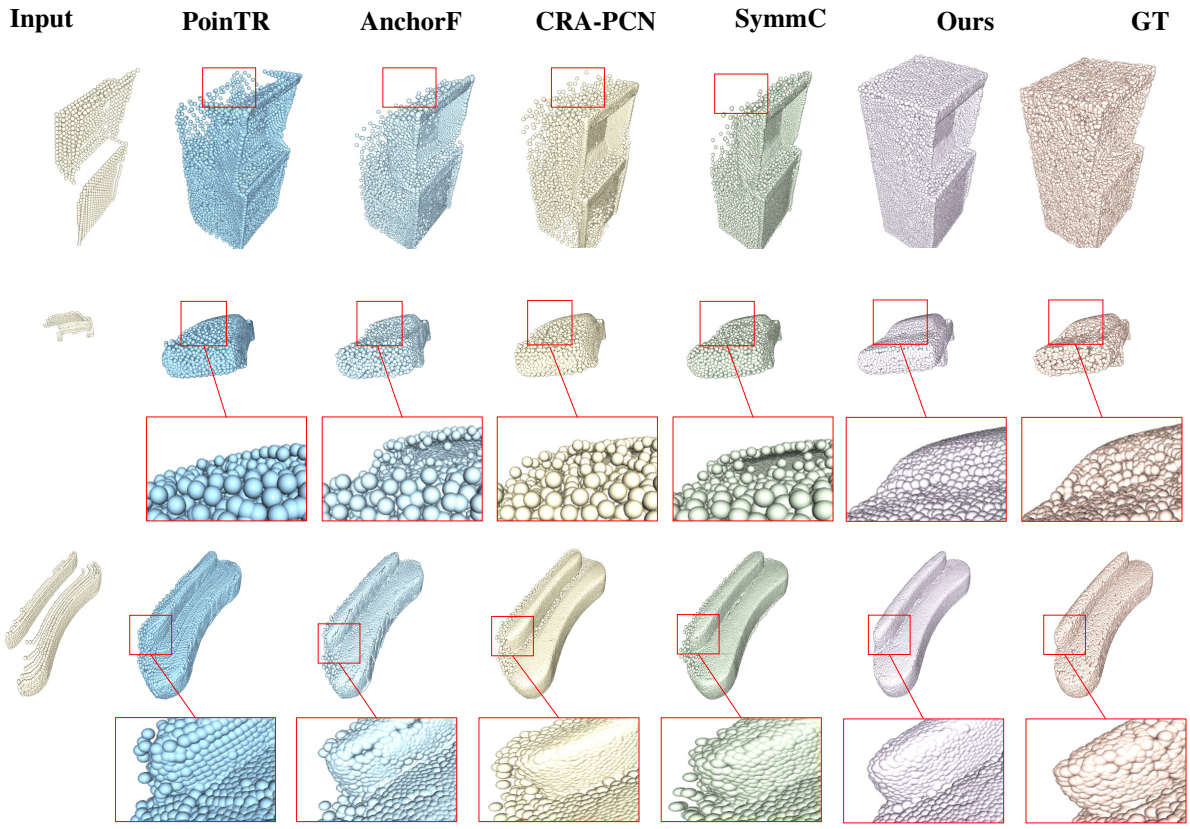

Figure 4. Visualization of results on the PCN dataset. Our model outperforms previous methods in completion fidelity, consistency, and local geometry recovery. AnchorF: AnchorFormer, SymmC: SymmCompletion.

Note that not all positions yield valid tactile information. Additionally, to rigorously test generalization, we follow [37] to validate our model (trained on ShapeNetCar) on the real-world KITTI dataset [6], challenging it with sparse LiDAR data from driving scenarios.

4.2. Comparisons with State-of-the-Art Methods

We present a comprehensive evaluation of our method against state-of-the-art approaches across multiple datasets.

Evaluation on PCN. Table 1 summarizes the quantitative comparison on the PCN dataset using L1 CD and F1-Score metrics. Our method achieves the best overall performance, obtaining the lowest average CD values and the highest F-Score across most object categories. This demonstrates our approach's effectiveness in generating geometrically accurate and complete point clouds. The visual quality of our completed point clouds is further demonstrated through qualitative comparisons in Figures 4. The results

Method	55 categories						34 seen categories	21 unseen categories
	Table	Chair	Plane	Car	Sofa	CD-Avg (↓)	CD-Avg (↓)	CD-Avg (↓)
PCN [40]	2.13	2.29	1.02	1.85	2.06	2.66	2.22	3.85
GRNet [33]	1.63	1.88	1.02	1.64	1.72	1.97	1.74	2.99
PoinTr [37]	0.81	0.95	0.44	0.91	0.79	1.09	1.23	2.05
SeedFormer [47]	0.72	0.81	0.40	0.89	0.71	0.92	0.83	1.34
AnchorFormer [2]	0.58	0.67	0.33	0.69	0.58	0.76	0.70	1.19
CRA-PCN [20]	0.66	0.74	0.37	0.85	0.66	0.85	0.76	1.24
PCDreamer [29]	0.98	0.71	0.40	0.96	0.85	0.90	0.76	1.09
SymmCompletion [34]	0.50	0.60	0.29	**0.64**	**0.51**	0.69	0.60	0.97
Ours	**0.42**	**0.58**	**0.28**	0.68	**0.51**	**0.65**	**0.59**	**0.66**

Table 2. Results of L2 CD $\times 10^3$ on the ShapeNet55/34 dataset. We show the Chamfer Distance for some categories while reporting the average on the metrics on the right.

confirm that our method is superior at preserving the underlying geometry and reconstructing consistent local details. For instance, as shown in the first example, prior methods tend to yield incomplete structures, while in the second and third examples, they fail to recover fine details. In contrast, our approach generates more complete and detailed outputs. We attribute this enhancement to the guidance from the generated tactile signals, which provides effective cues for refining the object's local structures.

Evaluation on ShapeNet-55/34. We extend our evaluation to the more comprehensive ShapeNet-55 dataset, with results detailed in Table 2. Furthermore, we investigate the generalization capability by evaluating on both the 34 seen categories and 21 unseen categories from ShapeNet-34. Our method outperforms other methods, and it achieves a significant improvement particularly on unseen categories. The results reveal that our method exhibits superior generalization performance, suggesting strong learning of fundamental shape priors.

The key to our method's consistent improvements across these datasets lies in its unique capability to generate the local geometry from imagined tactile information. This local geometry enhances the fine-grained details of the final output, as will be further validated in the experimental analysis (Section 4.3).

Metric	PointTR [37]	SVDF [49]	SymmC [34]	Ours
FD ↓	**0.0**	11.3	2.54	1.43
MMD ↓	8.21	0.97	1.72	**0.71**

Table 3. Generalization performance on KITTI dataset. SVDF: SVDFormer, SymmC: SymmCompletion.

Evaluation on Real-World Generalization To assess the practical applicability of our approach, we evaluate its generalization performance on the KITTI dataset using models trained on the ShapeNet Car category. This setup tests sim-to-real transfer capability under challenging real-world conditions, including severe point sparsity, sensor noise, and unstructured environments. We report the Fidelity Distance (FD) and Minimal Matching Distance (MMD) in Table 3.

As noted by SymmCompletion [34], the domain and scale gap between the KITTI dataset and the PCN dataset may affect the fairness and accuracy of quantitative comparisons. Although their method does not achieve the best results, it demonstrates better visual completeness. Figure 5 presents the completion results from two viewpoints for two representative cases. The qualitative results demonstrate that our method produces more reasonable and coherent completions on real-world data, with particularly superior handling of local geometric details. These qualitative outcomes, combined with quantitative results on the ShapeNet 21 unseen categories, strongly support the generalization ability of our approach.

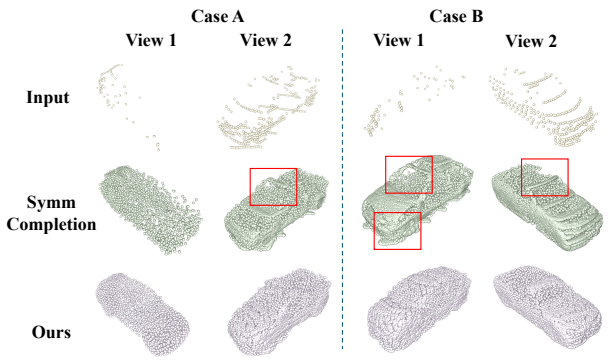

Figure 5. Visualization of results on the KITTI dataset.

4.3. Analysis of TouchDream

To validate the necessity of our approach, we conducted further experimental analysis and report results on the PCN dataset in the following.

Does generated touch help? We compare three guidance strategies for point cloud refinement: generated touch (T-GEN), ground truth touch (T-GT), and skeleton points from the ground truth shape (S-GT). Results in Table 4 and Figure 6 show that the ground truth tactile pointss obtained

are superior to skeleton points, confirming the role of tactile information in point completion tasks. Moreover, the tactile generation method proposed in this paper achieves results comparable to those obtained from simulated tactile point clouds, illustrating the effectiveness of imagined touch.

Metric	T-GEN	S-GT	T-GT
CD-Avg ↓	6.05	6.04	5.93

Table 4. Comparison of guidance point sources.

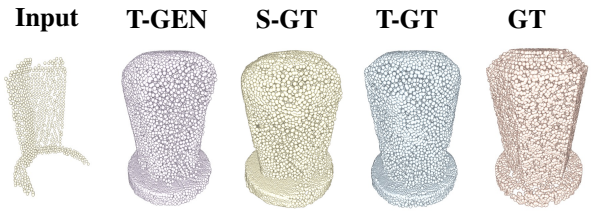

Figure 6. Visualized comparison of guidance point sources.

Metric	Symmetry Guide	Touch Guide	Combine
CD-Avg ↓	6.28	6.05	6.02

Table 5. Comparison of different Guidances.

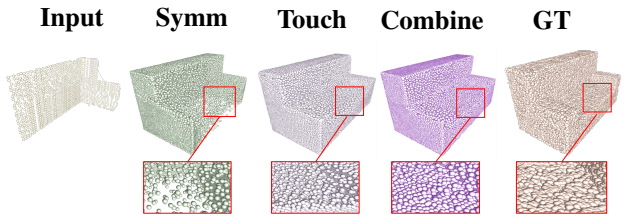

Figure 7. Visualized comparison of different guidance.

R / T	0 / 0	0.005 / 0.002	0.01 / 0.005	0.02 / 0.01
w/o DA	6.053	6.061	6.085	6.183
DA	6.060	6.064	6.069	6.087

Table 6. Evaluation of the touch pose estimation affect.

Comparison of symmetry guidance with touch guidance. In this test, we use either the only symmetry guidance from SymmCompletion[34], only touch guidance, or their combination to guide the point refinement. The quantitative and qualitative results are summarized in Table 5 and Figure 7 (Symm: SymmCompletion [34]). Our results show that generated touch guidance yields superior performance to symmetry guidance and also achieves competitive performance compared to their combination, as it robustly recovers finer details and avoids the poor local results that symmetry guidance produces with highly incomplete inputs.

Affect of Touch Pose Estimation We conducted experiments to assess how touch pose estimation affects final results. As shown in the table 6: The first row lists the maximum values of uniformly random noise applied to the rotation euler angles and translation respectively. The second row reports the corresponding Chamfer Distance (↓) of our model at test time. To mitigate pose sensitivity, we also applied uniform random noises to poses as data augmentation during training and presented test results in the third row.

Limitations. Although our model achieves state-of-the-art performance, its effectiveness is contingent upon the initial coarse point cloud. When input data is limited, leading to a coarse prediction that significantly deviates from the ground truth, the subsequent tactile point cloud may fail to capture the object's true geometry, thereby compromising the final output. Furthermore, we need to use the object mesh to render tactile images as supervision for the network. Another key limitation pertains to computational efficiency. Since our approach generates a tactile latent code for every sampled touch pose and then performs decoding, transformation, and sampling to obtain the local geometry, the process is inherently more time-consuming. Notably, a full evaluation on the 1,200 shapes from the PCN dataset required 6.7 hours on a single NVIDIA GTX 4090 GPU.

5. Conclusion

In this work, we present TouchDream, a novel framework that integrates a diffusion model to imagine touch on object surfaces. A key component of our framework is a diffusion model trained to generate tactile latent vectors, where the generation is conditioned on the coarse point cloud and a series of sampled touch poses. This collection of local points is then merged and sampled to serve as the fuel for the point completion and refinement. Extensive experiments validate the effectiveness of our method, demonstrating significant improvements in point completion performance, particularly in recovering fine-grained details.

There are some future directions. First, a learnable tactile pose sampling strategy, trained via reinforcement learning as in [28] and [23], could be developed to enable active tactile generation. Second, integrating tactile and visual generation into a unified multimodal framework may further enhance completion quality through cross-modal fusion. Finally, the local geometry generated from imagined touch could be leveraged to refine 3D Gaussians predicted from single (or few) images.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.U25 B2044/62502234), the Key Research and Development Program of Liaoning Province of China (No.2023JH2 6/10200014), and the Ningbo Major Research and Development Plan Project of China (Grant No.2023Z225).

References

[1] Guanqun Cao, Jiaqi Jiang, Chen Lu, Daniel Fernandes Gomes, and Shan Luo. Touchroller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas. *Sensors*, 23(5):2661, 2023. 3

[2] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Anchorformer: Point cloud completion from discriminative nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13581–13590, 2023. 2, 6, 7

[3] Xinhua Cheng, Nan Zhang, Jiwen Yu, Yinhuai Wang, Ge Li, and Jian Zhang. Null-space diffusion sampling for zero-shot point cloud completion. In *International Joint Conference on Artificial Intelligence*, pages 618–626, 2023. 2

[4] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2272, 2023. 2, 5

[5] Mauro Comi, Yijiong Lin, Alex Church, Alessio Tonioni, Laurence Aitchison, and Nathan F Lepora. Touchsdf: A deepsdf approach for 3d shape reconstruction using vision-based tactile sensing. *IEEE Robotics and Automation Letters*, 2024. 2, 3

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[7] Xiaoguang Han, Zhaoxuan Zhang, Dong Du, Mingdai Yang, Jingming Yu, Pan Pan, Xin Yang, Ligang Liu, Zixiang Xiong, and Shuguang Cui. Deep reinforcement learning of volume-guided progressive view inpainting for 3d point scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2019. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[9] Tao Hu, Zhizhong Han, Abhinav Shrivastava, and Matthias Zwicker. Render4completion: Synthesizing multi-view depth maps for 3d shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2

[10] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 12171–12191, 2023. 2, 3

[11] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Symmetry descriptors and 3d shape matching. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, pages 115–123, 2004. 2

[12] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 3

[13] Chen Li, Hao Wang, Lingxi Xie, Aude Yu, ille, and Yu-Wing Tai. Force-guided tactile reconstruction of curved 3d surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14342–14352, 2021. 3

[14] Yuhan Li, Yishun Dou, Xuanhong Chen, Bingbing Ni, Yilin Sun, Yutian Liu, and Fuzhen Wang. 3dqd: Generalized deep 3d shape prior via part-discretized diffusion process. *arXiv preprint arXiv:2303.10406*, 2023. 2

[15] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11596–11603, 2020. 2

[16] Yangrong Liu, Jian Li, Huaiyu Wang, Ming Lu, Haorao Shen, and Qin Wang. Tapcnet: Tactile-assisted point cloud completion network via iterative fusion strategy. *IET Computer Vision*, 19(1):e70012, 2025. 3

[17] Xuehan Ma, Xueyan Li, and Junfeng Song. Point cloud completion network applied to vehicle data. *Sensors*, 22(19): 7346, 2022. 2

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5

[19] Yiming Ren, Peishan Cong, Xinge Zhu, and Yuexin Ma. Self-supervised point cloud completion on real traffic scenes via scene-concerned bottom-up mechanism. In *2022 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022. 3

[20] Yi Rong, Haoran Zhou, Lixin Yuan, Cheng Mei, Jiahao Wang, and Tong Lu. Cra-pcn: Point cloud completion with intra-and inter-level cross-resolution transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4676–4685, 2024. 2, 6, 7

[21] Taras Rumezhak, Oles Dobosevych, Rostyslav Hryniv, Vladyslav Selotkin, Volodymyr Karpiv, and Mykola Maksymenko. Towards realistic symmetry-based completion of previously unseen point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2542–2550, 2021. 2

[22] Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 33: 14193–14206, 2020. 2, 3, 5

[23] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 34:16064–16078, 2021. 2, 3, 4, 5, 8

[24] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. 2

[25] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy. Touchgs: Visual-tactile supervised 3d gaussian splatting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10511–10518, 2024. 3

[26] Junshu Tang, Zhijun Gong, Ran Yi, Yuan Xie, and Lizhuang Ma. Lake-net: Topology-aware point cloud completion by localizing aligned keypoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1726–1735, 2022. 2

[27] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 790–799, 2020. 2

[28] Yuanbo Wang, Zhaoxuan Zhang, Jiajin Qiu, Dilong Sun, Zhengyu Meng, Xiaopeng Wei, and Xin Yang. Touch2shape: Touch-conditioned 3d diffusion for shape exploration and reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5656–5665, 2025. 2, 3, 4, 5, 8

[29] Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. Pcdreamer: Point cloud completion through multi-view diffusion priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27243–27253, 2025. 2, 3, 6, 7

[30] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Pmp-net++: Point cloud completion by transformer-enhanced multi-step point moving paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):852–867, 2022. 2

[31] Lintai Wu, Qijian Zhang, Junhui Hou, and Yong Xu. Leveraging single-view images for unsupervised 3d point cloud completion. *IEEE Transactions on Multimedia*, 2023. 2

[32] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5499–5509, 2021. 2, 6

[33] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *European Conference on Computer Vision*, pages 365–381, 2020. 7

[34] Hongyu Yan, Zijun Li, Kunming Luo, Li Lu, and Ping Tan. Symmcompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9094–9102, 2025. 2, 3, 5, 6, 7, 8, 1

[35] Xuejun Yan, Hongyu Yan, Jingjing Wang, Hang Du, Zhihong Wu, Di Xie, Shiliang Pu, and Li Lu. Fbnet: Feedback network for point cloud completion. In *European Conference on Computer Vision*, pages 676–693, 2022. 6

[36] Xuemeng Yang, Hao Zou, Xin Kong, Tianxin Huang, Yong Liu, Wanlong Li, Feng Wen, and Hongbo Zhang. Semantic segmentation-assisted scene completion for lidar point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3555–3562, 2021. 3

[37] 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, 5, 6, 7

[38] Liang Yuan, Dingkun Yan, Suguru Saito, and Issei Fujishiro. Diffmat: Latent diffusion models for image-guided material generation. *Visual Informatics*, 8(1):6–14, 2024. 2

[39] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 3

[40] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *International Conference on 3D Vision*, pages 728–737, 2018. 2, 5, 6, 7

[41] 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. 2

[42] Zhaoxuan Zhang, Bo Dong, Tong Li, Felix Heide, Pieter Peers, Baocai Yin, and Xin Yang. Single depth-image 3d reflection symmetry and shape prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8896–8906, 2023. 2

[43] Zhaoxuan Zhang, Xiaoguang Han, Bo Dong, Tong Li, Baocai Yin, and Xin Yang. Point cloud scene completion with joint color and semantic estimation from single rgb-d image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11079–11095, 2023. 2

[44] 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. 5

[45] Hongliang Zhong, Can Wang, Jingbo Zhang, and Jing Liao. Generative object insertion in gaussian splatting with a multi-view diffusion model. *Visual Informatics*, 9(2): 100238, 2025. 2

[46] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pages 1618–1628, 2023. 3

[47] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. In *European Conference on Computer Vision*, pages 416–432, 2022. 2, 6, 7

[48] 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. 2

[49] Zhe Zhu, Honghua Chen, Xing He, Weiming Wang, Jing Qin, and Mingqiang Wei. Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14508–14518, 2023. 2, 3, 6, 7

[50] Zhe Zhu, Honghua Chen, Xing He, and Mingqiang Wei. Pointsea: Point cloud completion via self-structure augmentation. *International Journal of Computer Vision*, 133(7): 4770–4794, 2025. 2, 3, 6