# DiffComplete: Diffusion-based Generative 3D Shape Completion

**Ruihang Chu**[1] **Enze Xie** [2] **Shentong Mo**[3]
**Zhenguo Li**[2] **Matthias Nießner**[4] **Chi-Wing Fu**[1] **Jiaya Jia**[1,5]

[1]The Chinese University of Hong Kong [2]Huawei Noah's Ark Lab
[3]MBZUAI [4]Technical University of Munich [5]SmartMore
https://ruihangchu.com/diffcomplete.html

## Abstract

We introduce a new diffusion-based approach for shape completion on 3D range scans. Compared with prior deterministic and probabilistic methods, we strike a balance between realism, multi-modality, and high fidelity. We propose DiffComplete by casting shape completion as a generative task conditioned on the incomplete shape. Our key designs are two-fold. First, we devise a hierarchical feature aggregation mechanism to inject conditional features in a spatially-consistent manner. So, we can capture both local details and broader contexts of the conditional inputs to control the shape completion. Second, we propose an occupancy-aware fusion strategy in our model to enable the completion of multiple partial shapes and introduce higher flexibility on the input conditions. DiffComplete sets a new SOTA performance (*e.g.*, 40% decrease on $l_1$ error) on two large-scale 3D shape completion benchmarks. Our completed shapes not only have a realistic outlook compared with the deterministic methods but also exhibit high similarity to the ground truths compared with the probabilistic alternatives. Further, DiffComplete has strong generalizability on objects of entirely unseen classes for both synthetic and real data, eliminating the need for model re-training in various applications.


Figure 1: Our method is able to (a) produce realistic completed shapes from partial scans, (b) incorporate multiple incomplete scans (denoted by the plus symbol) to improve the completion accuracy, and (c) directly generalize to work on real objects of unseen classes without finetuning.

## 1 Introduction

The advent of affordable range sensors, such as the Microsoft Kinect and Intel RealSense, has spurred remarkable progress in 3D reconstruction [1], facilitating applications in content creation, mixed reality, robot navigation, and more. Despite the improved reconstruction quality [2–5], typical scanning sessions cannot cover every aspect of the scene objects. Hence, the reconstructed 3D models often have incomplete geometry, thereby limiting their practical usage in downstream applications.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).


https://doi.org/10.52202/075280-3318

To fully unlock the potential of 3D reconstruction for supporting assorted applications, it is essential to address the challenges of completing incomplete shapes.

Effectively, a shape completion should produce shapes that are *realistic*, *probabilistic*, and *high-fidelity*. First, the produced shapes should look plausible without visual artifacts. Second, considering the under-determined nature of completion, it is desirable for the model to generate diverse, multi-modal outputs when filling in the missing regions, to improve the likelihood of obtaining a superior shape and enable creative use. Third, while multiple outputs encourage the model's generative ability, maintaining effective control over the completion results is crucial to ensure coherent reconstructions that closely resemble the ground-truth shapes.

Current approaches to 3D shape completion fall into deterministic and probabilistic paradigms. The former class excels at aligning predictions with ground truths, owing to their determined mapping functions and full supervision. However, this can expose the models to a higher risk of over-fitting, leading to undesirable artifacts, such as rugged surfaces, especially on unseen objects. On the other hand, probabilistic approaches formulate the shape completion as a conditional generation task to produce plausible results, paired with techniques like autoregressive model [6], adversarial training [7–9], or diffusion models [10–13]. These approaches mainly focus on cases that prioritize completion diversity, for instance, filling in large missing regions cropped out by 3D boxes, where the algorithm has more freedom to explore various plausible shapes. However, when completing shapes obtained from range scans/reconstructions, whose geometries are often contaminated with noise, having varying degrees of incompleteness (see row 1 in Fig. 1), a promising goal is to recover the ground-truth scanned objects as accurately as possible. Therefore, relying solely on previous probabilistic approaches may compromise the completion accuracy and leads to low-fidelity results.

In this work, we focus on completing shapes acquired from range scans, aiming to produce *realistic* and *high-fidelity* completed shapes, while considering also the *probabilistic* uncertainty. We approach this task as a conditional generation with the proposed diffusion model named DiffComplete. To address the accuracy challenge typically in probabilistic models, we introduce a key design that incorporates range scan features in a hierarchical and spatially-consistent manner to facilitate effective controls. Additionally, we propose a novel occupancy-aware fusion strategy that allows taking multiple partial shapes as input for more precise and controllable completions.

Specifically, we formulate the diffusion process on a truncated distance field in a volumetric space. This representation allows us to encode both complete and incomplete shapes into multi-resolution structured feature volumes, thereby enabling their interactions at various network levels. At each level, features are aggregated based on voxel-to-voxel correspondence, enabling precise correlation of the difference in partialness between the shapes. By leveraging multi-level aggregation, the completion process can be controlled by both local details and broader contexts of the conditional inputs. It effectively improves the completion accuracy and network generalizability, as illusrated in Sec. 4.5.

To condition the completion on multiple partial shapes, our occupancy-aware fusion strategy adopts the occupancy mask as a weight to adaptively fuse all observed geometry features from various incomplete shapes, and then employ the fused features to control the completion process. For robustness, such an operation is performed in a multi-scale feature space. With our novel design, switching between single and multiple conditions can be efficiently achieved by finetuning an MLP layer. As Fig. 1(b) and 6 show, using multiple conditions for completion progressively refines the final shapes towards the ground truths.

DiffComplete sets a new state-of-the-art performance on two large-scale shape completion benchmarks [14, 15] in terms of completion accuracy and visual quality. Compared to prior deterministic and probabilistic approaches, we not only generate multimodal plausible shapes with minimal artifacts but also present high shape fidelity relative to the ground truths. Further, DiffComplete can directly generalize to work on unseen object categories. Without specific zero-shot designs, it surpasses all existing approaches for both synthetic and real-world data, allowing us to eliminate the need for model re-training in various applications. To sum up, our contributions are listed as follows:

- We introduce a diffusion model to complete 3D shapes on range scans, enabling realistic, probabilistic, and high-fidelity 3D geometry.
- We enhance the completion accuracy with two designs. For a single condition, we propose a hierarchical feature aggregation strategy to control the outputs. For multiple conditions, we introduce an occupancy-aware fusion strategy to incorporate more shape details.

https://doi.org/10.52202/075280-3318

- We show SOTA performance on shape completion on both novel instances and entirely unseen object categories, along with an in-depth analysis on a range of model characteristics.

## 2 Related Work

**RGB-D reconstruction.** Traditional methods rely on geometric approaches for 3D reconstruction [16–20]. A pioneering method [21] proposes a volumetric fusion strategy to integrate multiple range images into a unified 3D model on truncated signed distance fields (TSDF), forming the basis for many modern techniques like KinectFusion [22, 23], VoxelHashing [24], and BundleFusion [2]. Recent learning-based approaches further improve the reconstruction quality with fewer artifacts [25–28, 3], yet the intrinsic occlusions and measurement noise of 3D scans constrain the completeness of 3D reconstructions, making them still less refined than manually-created assets.

**3D shape completion** is a common post-processing step to fill in missing parts in the reconstructed shapes. Classic methods mainly handle small holes and geometry primitives via Laplacian hole filling [29–31] or Poisson surface reconstruction [32, 33]. Another line exploits the structural regularities of 3D shapes, such as the symmetries, to predict the unobserved data [34–38].

The availability of large 3D data has sparked retrieval-based methods [39–42] and learning-based fitting methods [43, 44, 14, 45–49]. The former retrieves the shapes from a database that best match the incomplete inputs, whereas the latter minimizes the difference between the network-predicted shapes and ground truths. 3D-EPN [14], for instance, proposes a 3D encoder-decoder architecture to predict the complete shape from partial volumetric data. Scan2Mesh [50] converts range scans into 3D meshes with a direct optimization on the mesh surface. PatchComplete [15] further leverages local structural priors for completing shapes of unseen categories.

Generative methods, e.g., GANs [51, 9, 52, 7, 8] and AutoEncoders [53], offer an alternative to shape completion. While some generative models can generate diverse global shapes given a partial input, they potentially allow for a high generation freedom and overlook the completion accuracy. IMLE [54], for instance, adopts an Implicit Maximum Likelihood Estimation technique to specially enhance structural variance among the generated shapes. Distinctively, we formulate a diffusion model for shape completion. Our method mainly prioritizes fidelity relative to ground truths while preserving output diversity. Compared to both generative and fitting-based paradigms, our method also effectively reduces surface artifacts, producing more realistic and natural 3D shapes. In addition, we show superior generalization ability on completing objects of novel classes over SOTAs.

**Diffusion models for 3D generation.** Diffusion models [55–62] have shown superior performance in various generation tasks, outperforming GANs' sample quality while preserving the likelihood evaluation property of VAEs. When adopted in 3D domain, a range of works [12, 63–65] focus on point cloud generation. For more complex surface generation, some works [66, 13, 10, 67, 68, 11] adopt latent diffusion models to learn implicit neural representations and form final shapes via a decoder. For conditional shape completion, both DiffRF [10] and Diffusion-SDF [69] adopt a masked diffusion strategy to fill in missing regions cropped out by 3D boxes. However, their training processes do not leverage a paired incomplete-to-complete ground truth, which may prevent them from accurately learning the completion rules. Contrarily, our method explicitly uses the scan pairs for conditional training, yielding outputs closely resembling the actual objects scanned. Also, we apply the diffusion process in explicit volumetric TSDFs, preserving more geometrical structures during the completing process.

## 3 Method

### 3.1 Formulation

To prepare the training data, we generate an incomplete 3D scan from depth frames using volumetric fusion [21] and represent the scan as a truncated signed distance field (TSDF) in a volumetric grid. Yet, to accurately represent a ground-truth shape, we choose to use a volumetric truncated unsigned distance field (TUDF) instead. This is because retrieving the sign bit from arbitrary 3D CAD models (*e.g.*, some are not closed) is non-trivial. By using TUDF, we can robustly capture the geometric features of different objects without being limited by the topology.


https://doi.org/10.52202/075280-3318

Given such a volume pair, *i.e.*, the incomplete scan $c$ and complete 3D shape $x_0$, we formulate the shape completion as a generation task that produces $x_0$ conditioned on $c$. We employ the probabilistic diffusion model as our generative model. In particular, it contains (i) a *forward process* (denoted as $q(x_{0:T})$) that gradually adds Gaussian noise to corrupt the ground-truth shape $x_0$ into a random noise volume $x_T$, where $T$ is the total number of time steps; and (ii) a *backward process* that employs a shape completion network, with learned parameters $\theta$, to iteratively remove the noise from $x_T$ conditioned on the partial scan $c$ and produce the final shape, denoted as $p_\theta(x_{0:T}, c)$. As both the *forward* and *backward* processes are governed by a discrete-time Markov chain with time steps $\{0, ..., T\}$, their Gaussian transition probabilities can be formulated as

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) := \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \tag{1}$$

$$\text{and } p_\theta(x_{0:T}, c) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c), \quad p_\theta(x_{t-1}|x_t) := \mathcal{N}(\mu_\theta(x_t, t, c), \sigma_t^2 \mathbf{I}). \tag{2}$$

In the *forward* process (Eq. (1)), the scalars $\beta_t \in [0, 1]$ control a variance schedule that defines the amount of noise added in each step $t$. In the *backward* process (Eq. (2)), $p(x_T)$ is a Gaussian prior in time step $t$, $\mu_\theta$ represents the mean predicted from our network and $\sigma_t^2$ is the variance. As suggested in DDPM [60], predicting $\mu_\theta(x_t, t, c)$ can be simplified to alternatively predicting $\epsilon_\theta(x_t, t, c)$, which should approximate the noise used to corrupt $x_{t-1}$ in the *forward* process, and $\sigma_t$ can be replaced by the pre-defined $\beta_t$. With these modifications, we can optimize the network parameter $\theta$ with a mean squared error loss to maximize the generation probability $p_\theta(x_0)$. The training objective is

$$\arg\min_\theta E_{t,x_0,\epsilon,c}[||\epsilon - \epsilon_\theta(x_t, t, c)||^2], \quad \epsilon \in \mathcal{N}(0, \mathbf{I}) \tag{3}$$

where $\epsilon$ is the applied noise to corrupt $x_0$ into $x_t$ and $\mathcal{N}(0, \mathbf{I})$ denotes a unit Gaussian distribution. We define all the diffusion processes in a volume space of a resolution $S^3$, *i.e.*, $x_{0:T}, c, \epsilon \in \mathbb{R}^{S \times S \times S}$, where each voxel stores a scalar TSDF/TUDF value; $S$=32 in our experiments. Compared to previous latent diffusion models [11, 13, 66, 10] that require shape embedding first, we directly manipulate the shape with a better preservation of geometric features. Doing so naturally enables our hierarchical feature aggregation strategy (see Sec. 3.2). Next, we will introduce how to predict $\epsilon_\theta(x_t, t, c)$.

### 3.2 Shape Completion Pipeline

**Overview.** To enhance the completion accuracy, we encourage the incomplete scans to control completion behaviors. Inspired from the recent ControlNet [70], which shows great control ability given 2D conditions, we adopt a similar principle to train an additional control branch. To predict the noise $\epsilon_\theta(x_t, t, c)$ in Eq. (3), we encode the corrupted ground-truth shape $x_t$ by a main branch and the partial shape $c$ by a control branch, where two branches have the same network structure without parameter sharing. Owing to the compact shape representation in 3D volume space, both complete and incomplete shapes are encoded into multi-resolution feature volumes with preserved spatial structures. Then, we hierarchically aggregate their features at each network level, as the sizes of two feature volumes are always aligned. In this work, we simply add up the two feature volumes to allow for a spatially-consistent feature aggregation, *i.e.*, only features at the same 3D location are combined. Compared with frequently-used cross-attention technique [11, 13, 57], doing so can significantly reduce computation costs. By multi-scale feature interaction, the network can effectively correlate the difference in partialness between two shapes, both locally and globally, to learn the completion regularities. The final integrated features are then used to predict the noise volume $\epsilon_\theta$.

**Network architecture.** Fig. 2 provides an overview of our network architecture, which has a main branch and a control branch to respectively handle complete and incomplete shapes. The main branch (upper) is a 3D U-Net modified from a 2D version [59]. It takes as input an corrupted complete shape $x_t$, including (i) a pre-processing block $\varepsilon_x(\cdot)$ of two convolution layers to project $x_t$ into a high-dimension space, (ii) $N$ subsequent encoder blocks, denoted as $\{F_x^i(\cdot)\}_{i=1}^N$, to progressively encode and downsample the corrupted shape $x_t$ into a collection of multi-scale features, (iii) a middle block $M_x(\cdot)$ with a self-attention layer to incorporate non-local information into the encoded feature volume, and (iv) $N$ decoder blocks $\{D_x^i(\cdot)\}_{i=1}^N$ that sequentially upsample features through an inversion convolution to produce a feature volume of the same size as the input $x_t$. The term

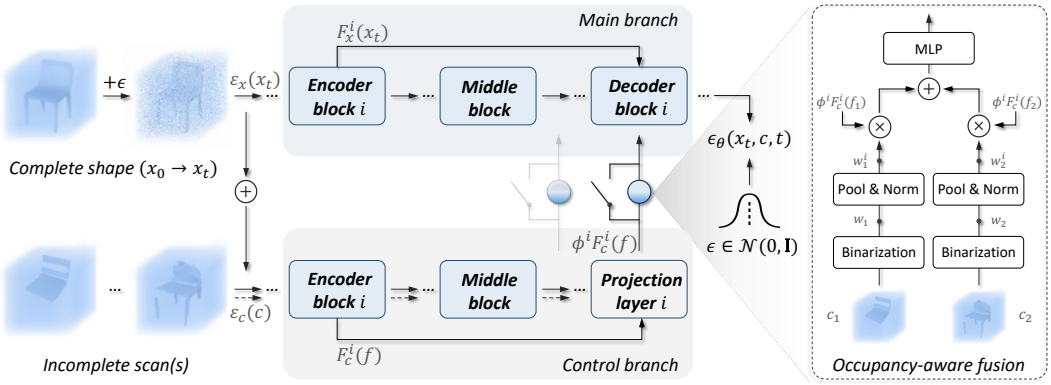


Figure 2: Given a corrupted complete shape $x_t$ (diffused from $x_0$) and an incomplete scan $c$, we first process them into $\varepsilon_x(x_t)$ and $\varepsilon_c(c)$ to align the distributions. We employ a main branch to forward $\varepsilon_x(x_t)$, and a control branch to propagate their fused features $f$ into deep layers. Multi-level features of $f$ are aggregated into the main branch for hierarchical control in predicting the diffusion noise. To support multiple partial scans as condition, *e.g.*, two scans $\{c_1, c_2\}$, we switch on occupancy-aware fusion (Sec. 3.3). This strategy utilizes the occupancy masks to enable a weighted feature fusion for $c_1$ and $c_2$ by considering their geometry reliability before feeding them into the main branch.

"e(d)ncoder/middle block" denotes a group of neural layers commonly adopted as a unit in neural networks, *e.g.*, a "resnet+downsample" block.

We use a control branch (bottom) to extract the features of incomplete scan(s). Likewise, we employ a pre-processing block, $N$ encoder blocks, and a middle block to extract multi-scale features from the conditional input, denoted as $\varepsilon_c(\cdot)$, $\{F_c^i(\cdot)\}_{i=1}^N$, and $M_c$, respectively. They mirror the architecture of the corresponding blocks in the upper branch, yet operating with non-shared parameters, considering the different input data regularities (*e.g.*, complete and sparse). Unlike the upper branch, there are no decoders for computation efficiency. Instead, we append a projection layer after each encoder/middle block $F_c^i/M_c$ to forward multi-scale features and feed them into the decoder blocks of the main branch. As diffusion models are time-conditioned, we also convert the time step $t$ into an embedding via two MLPs and add it to the volume features in each network block (the blue box in Fig. 2).

**Hierarchical feature aggregation.** To integrate the features of the complete and incomplete shapes for conditional control, we fuse them across multiple network levels. In this process, we first consider a *single* incomplete shape (denoted as $c$) as the condition. Given the significant disparity between $c$ and $x_t$ (incomplete v.s. complete, TSDF v.s. TUDF), we employ the pre-processing layers ($\varepsilon_x(\cdot)$ and $\varepsilon_c(\cdot)$) to first empirically project the two different fields into more compatible feature space. Then we fuse them into feature volume $f$, *i.e.*, $f = \varepsilon_x(x_t) + \varepsilon_c(c)$. After that, $f$ is passed through the control branch to better propagate the useful information to the deeper layers. In turn, we use multi-scale condition features from the control branch to guide each level of decoder blocks $\{D_x^i(\cdot)\}_{i=1}^N$ in the main branch. The features that enter the $i$-th decoder block are denoted as

$$d^i = [D_x^{i-1}(x_t), F_x^i(x_t) + \phi^i(F_c^i(f))], \quad f = \varepsilon_x(x_t) + \varepsilon_c(c) \tag{4}$$

where $[\cdot, \cdot]$ is the concatenation operation and $\phi^i$ is one $1 \times 1$ convolution layer for feature projection. For simplicity, we use $F_x^i(x_0)$ to denote the features of the $x_0$ after the $i$-th encoder block $F_x^i(\cdot)$, as $x_0$ is not directly processed by $F_x^i(\cdot)$; the same notation is used for $F_c^i(f)$ and $D_x^{i-1}(x_t)$. The feature after the final decoder block is the network output $\epsilon_\theta(x_t, t, c)$.

By such a design, we can hierarchically incorporate conditional features, leveraging both their local and broader contexts to optimize the network outputs. Interestingly, we observe that adjusting the network level for feature aggregation can alter a trade-off between completion accuracy and diversity. Specifically, if we only aggregate features at the low-resolution network layers, we might miss finer geometry details present in the higher-resolution layers. This could reduce completion accuracy, as the model has less contextual information to work with. In contrast, if aggregating features at all levels, the effective control might make completion results closely resemble the ground truths, yet lowering the diversity. Further ablation study is provided in Sec. 4.5.


https://doi.org/10.52202/075280-3318

### 3.3 Occupancy-aware Fusion

Our framework can also take multiple incomplete scans as inputs. This option not only provides richer information to constrain the completed shape geometry but also enhances the approach's practicality, particularly in scenarios that a single-pass scan may not fully capture the entire object. Our critical design is to effectively fuse multiple partial shape features. As averaging the original TSDF volumes is sensitive to registration noise, we first register multiple partial shapes (*e.g.*, using Fast-Robust-ICP [71]) and propose an occupancy-aware approach to fuse them in the feature space.

Given a set of incomplete shapes of the same object, denoted as $\{c_1, ..., c_M\}$, we individually feed them into the control branch to produce feature volumes $\{F_c^i(f_1), ..., F_c^i(f_M)\}$ after the $i$-th encoder block. Before feeding them into the decoder $D_x^i$, we refer to their occupancy masks to perform a weighted feature average. Concretely, we first compute the original occupancy mask for each partial shape based on the TSDF values, *i.e.*, $w_j = (abs(c_j) < \tau) \in \mathbf{B}^{S\times S\times S}$ for the $j$-th partial shape. $\tau$ is a pre-defined threshold that assigns the volumes near the object surface as occupied and the rest as unoccupied; $\mathbf{B}$ is a binary tensor. Then, we perform a pooling operation to resize $w_j$ into $w_j^i$ to align the resolution with feature volume $F_c^i(f_j)$ and normalize $w_j^i$ by $w_j^i = w_j^i/(w_0^i + .. + w_M^i)$. For the voxels with zero values across all occupancy masks, we uniformly assign them a 1e-2 value to avoid the division-by-zero issue. We rewrite Eq. (4), which is for single partial shape condition, as

$$d^i = [D_x^{i-1}(x_t), F_x^i(x_t) + \psi(\sum_j w_j^i \phi^i(F_c^i(f_j)))], \quad f_j = \varepsilon_x(x_t) + \varepsilon_c(c_j) \tag{5}$$

where $\psi$ is an MLP layer to refine the fused condition features, aiming to mitigate the discrepancies among different partial shape features, as validated in Sec. 4.4. The right part of Fig. 2 illustrates the above process with two incomplete scans as the inputs.

### 3.4 Training and Inference

We first train the network with a single incomplete shape as the conditional input. In this phase, all the network parameters, except the MLP layer $\psi$ for occupancy-aware fusion (in Eq. (5)), are trained with the objective in Eq. (3). When the network converges, we lock the optimized network parameters and efficiently finetune the MLP layer $\psi$ with multiple incomplete shapes as input.

At the inference stage, we randomize a 3D noise volume as $x_T$ from the standard Gaussian distribution. The trained completion networks are then employed for $T$ iterations to produce $x_0$ from $x_T$ conditioned on partial shape(s) $c$. The occupancy-aware fusion is activated only for multi-condition completion. To accelerate the inference process, we adopt a technique from [61] to sub-sample a set of time steps from [1,...,$T$/10] during inference. After obtaining the generated shape volume $x_0$, we extract an explicit 3D mesh using the marching cube algorithm [72].

## 4 Experiment

### 4.1 Experimental Setup

**Benchmarks.** We evaluate on two large-scale shape completion benchmarks: 3D-EPN [14] and PatchComplete [15]. 3D-EPN comprises 25,590 object instances of eight classes in ShapeNet [73]. For each instance, six partial scans of varying completeness are created in the $32^3$ TSDF volumes by virtual scanning; the ground-truth counterpart, represented by $32^3$ TUDF, is obtained by a distance field transform on a 3D scanline method [74]. While using a similar data generation pipeline, PatchComplete emphasizes completing objects of unseen categories. It includes both the synthetic data from ShapeNet [73] and the challenging real data from ScanNet [75]. For a fair comparison, we follow their data splits and evaluation metrics, *i.e.*, mean $l_1$ error on the TUDF predictions across all voxels on 3D-EPN, and $l_1$ Chamfer Distance (CD) and Intersection over Union (IoU) between the predicted and ground-truth shapes on PatchComplete. As these metrics only measure the completion accuracy, we introduce other metrics in Sec. 4.3 to compare multimodal completion characteristics.

**Implementation details.** We first train our network using a single partial scan as input by 200k iterations on four RTX3090 GPUs, taking around two days. If multiple conditions are needed, we finetune project layers $\psi$ for additional 50k iterations. Adam optimizer [76] is employed with a learning rate of $1e^{-4}$ and the batch size is 32. On the 3D-EPN benchmark, we train a specific

Table 1: Quantitative shape completion results on objects of known categories [14].

| $l_1$-err. (↓) | Avg. (↓) | Chair | Table | Sofa | Lamp | Plane | Car | Dresser | Boat |
|---|---|---|---|---|---|---|---|---|---|
| 3D-EPN [14] | 0.374 | 0.418 | 0.377 | 0.392 | 0.388 | 0.421 | 0.259 | 0.381 | 0.356 |
| SDF-StyleGAN [51] | 0.278 | 0.321 | 0.256 | 0.289 | 0.280 | 0.295 | 0.224 | 0.273 | 0.282 |
| RePaint-3D [78] | 0.266 | 0.289 | 0.264 | 0.266 | 0.268 | 0.302 | 0.214 | 0.285 | 0.243 |
| ConvONet [26] | 0.220 | 0.210 | 0.247 | 0.254 | 0.234 | 0.185 | 0.195 | 0.250 | 0.184 |
| AutoSDF [6] | 0.217 | 0.201 | 0.258 | 0.226 | 0.275 | 0.184 | 0.187 | 0.248 | 0.157 |
| cGCA [79] | 0.185 | 0.174 | 0.212 | 0.179 | 0.239 | 0.170 | 0.161 | 0.204 | 0.143 |
| ShapeFormer [80] | 0.141 | 0.104 | 0.175 | 0.133 | 0.176 | 0.136 | 0.127 | 0.157 | 0.119 |
| PVD [12] | 0.114 | 0.097 | 0.122 | 0.154 | 0.128 | 0.093 | 0.087 | 0.127 | 0.107 |
| PatchComplete [15] | 0.088 | 0.134 | 0.095 | 0.084 | 0.087 | 0.061 | 0.053 | 0.134 | 0.058 |
| DiffComplete (Ours) | **0.053** | **0.070** | **0.073** | **0.061** | **0.059** | **0.015** | **0.025** | **0.086** | **0.031** |

Table 2: Shape completion results on synthetic objects [73] of unseen categories. $\cdot/\cdot$ means CD/IoU.

| CD(↓)/IoU(↑) | 3D-EPN [14] | Few-Shot [81] | IF-Nets [49] | Auto-SDF [6] | ConvONet [26] | PatchComplete [15] | Ours |
|---|---|---|---|---|---|---|---|
| Bag | 5.01 / 73.8 | 8.00 / 56.1 | 4.77 / 69.8 | 5.81 / 56.3 | 5.10 / 70.8 | 3.94 / 77.6 | **3.86 / 78.3** |
| Lamp | 8.07 / 47.2 | 15.1 / 25.4 | 5.70 / 50.8 | 6.57 / 39.1 | 5.42 / 52.6 | **4.68** / 56.4 | 4.80 / **57.9** |
| Bathtub | 4.21 / 57.9 | 7.05 / 45.7 | 4.72 / 55.0 | 5.17 / 41.0 | 4.96 / 60.4 | 3.78 / 66.3 | **3.52 / 68.9** |
| Bed | 5.84 / 58.4 | 10.0 / 39.6 | 5.34 / 60.7 | 6.01 / 44.6 | 5.42 / 63.2 | 4.49 / 66.8 | **4.16 / 67.1** |
| Basket | 7.90 / 54.0 | 8.72 / 40.6 | **4.44** / 50.2 | 6.70 / 39.8 | 6.16 / 54.6 | 5.15 / 61.0 | 4.94 / **65.5** |
| Printer | 5.15 / 73.6 | 9.26 / 56.7 | 5.83 / 70.5 | 7.52 / 49.9 | 5.56 / 72.1 | 4.63 / **77.6** | **4.40** / 76.8 |
| Laptop | 3.90 / 62.0 | 10.4 / 31.3 | 6.47 / 58.3 | 4.81 / 51.1 | 4.78 / 57.3 | 3.77 / 63.8 | **3.52 / 67.4** |
| Bench | 4.54 / 48.3 | 8.11 / 27.2 | 5.03 / 49.7 | 4.31 / 39.5 | 4.69 / 49.6 | 3.70 / 53.9 | **3.56 / 58.2** |
| Avg. | 5.58 / 59.4 | 9.58 / 40.3 | 5.29 / 58.1 | 5.86 / 45.2 | 5.26 / 60.1 | 4.27 / 65.4 | **4.10 / 67.5** |

Table 3: Shape completion results on real-world objects [75] of unseen categories. $\cdot/\cdot$ means CD/IoU.

| CD(↓)/IoU(↑) | 3D-EPN [14] | Few-Shot [81] | IF-Nets [49] | Auto-SDF [6] | ConvONet [26] | PatchComplete [15] | Ours |
|---|---|---|---|---|---|---|---|
| Bag | 8.83 / 53.7 | 9.10 / 44.9 | 8.96 / 44.2 | 9.30 / 48.7 | 9.12 / 52.5 | 8.23 / **58.3** | **7.05** / 48.5 |
| Lamp | 14.3 / 20.7 | 11.9 / 19.6 | 10.2 / 24.9 | 11.2 / 24.4 | 9.83 / 20.3 | 9.42 / 28.4 | **6.84 / 30.5** |
| Bathtub | 7.56 / 41.0 | 7.77 / 38.2 | 7.19 / 39.5 | 7.84 / 36.6 | 7.93 / 41.2 | **6.77** / 48.0 | 8.22 / **48.5** |
| Bed | 7.76 / 47.8 | 9.07 / 34.9 | 8.24 / 44.9 | 7.91 / 38.0 | 8.14 / 41.6 | 7.24 / **48.4** | **7.20** / 46.6 |
| Basket | 7.74 / 36.5 | 8.02 / 34.3 | 6.74 / 42.7 | 7.54 / 36.1 | 7.39 / 37.0 | **6.60** / 45.5 | 7.42 / **59.2** |
| Printer | 8.36 / 63.0 | 8.30 / 62.2 | 8.28 / 60.7 | 9.66 / 49.9 | 7.62 / 64.9 | 6.84 / 70.5 | **6.36 / 74.5** |
| Avg. | 9.09 / 44.0 | 9.02 / 38.6 | 8.26 / 42.6 | 8.90 / 38.9 | 8.34 / 42.9 | 7.52 / 49.8 | **7.18 / 51.3** |

model for completing shapes of each known category; while on PatchComplete, we merge all object categories to optimize one model for promoting general completion learning. Due to the unknown class IDs at test time, no classifier-guided [56] or classifier-free [77] sampling techniques are used in our diffusion model. On both two benchmarks, all training data are the same across the compared methods for fairness. The truncation distance in TSDF/TUDFs is set as 3 voxel units. More details about network architecture and experiments are available in the supplementary file. Unless otherwise specified, we report the results on *single* partial shape completion.

### 4.2 Main Results

**Completion on known object categories.** On the 3D-EPN benchmark, we compare DiffComplete against SOTA deterministic [14, 15, 51] and probabilistic [6, 78] methods in terms of completion accuracy (*i.e.*, $l_1$ errors). For probabilistic methods, we use the average results from five inferences, each with random initialization, to account for multimodal outcomes. As shown in Table 1 and Fig. 3, DiffComplete improves over state of the arts by 40% on $l_1$ error (0.053 v.s. 0.088), as well as producing more realistic and high-fidelity shapes. Deterministic methods include 3D-EPN [14], ConvONet [26], and PatchComplete [15]. Unlike these learning a one-step map function for shape completion, we iteratively refine the generated shape, thus significantly mitigating the surface artifacts; see visualization comparisons in Fig. 3. Compared to GAN-based SDF-StyleGAN [51] and Autoregressive-based AutoSDF [6], our diffusion-based generative model offers superior mode coverage and sampling quality.


https://doi.org/10.52202/075280-3318

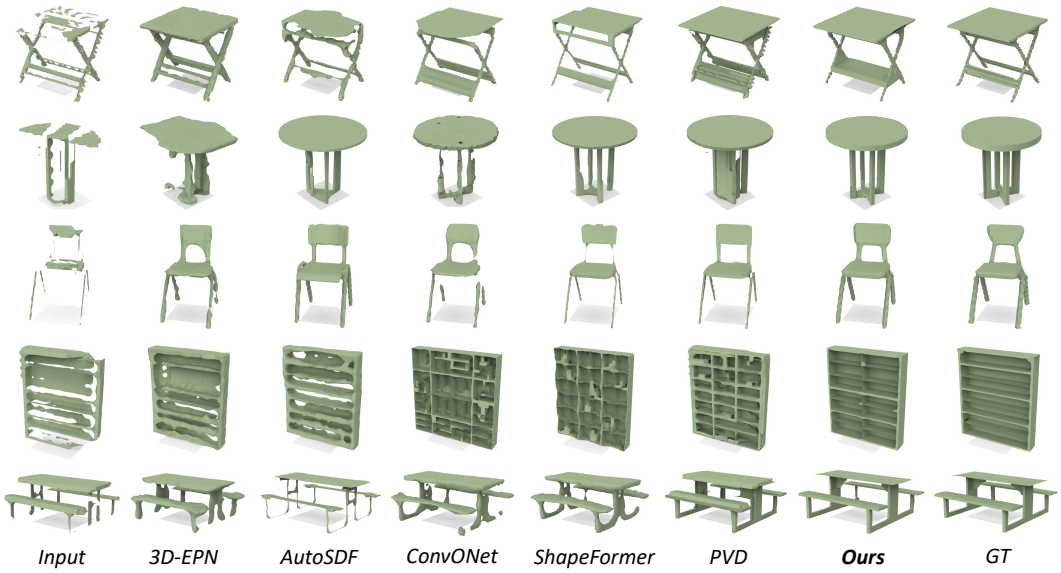


Figure 3: Shape Completion on various known object classes. We achieve the best completion quality.

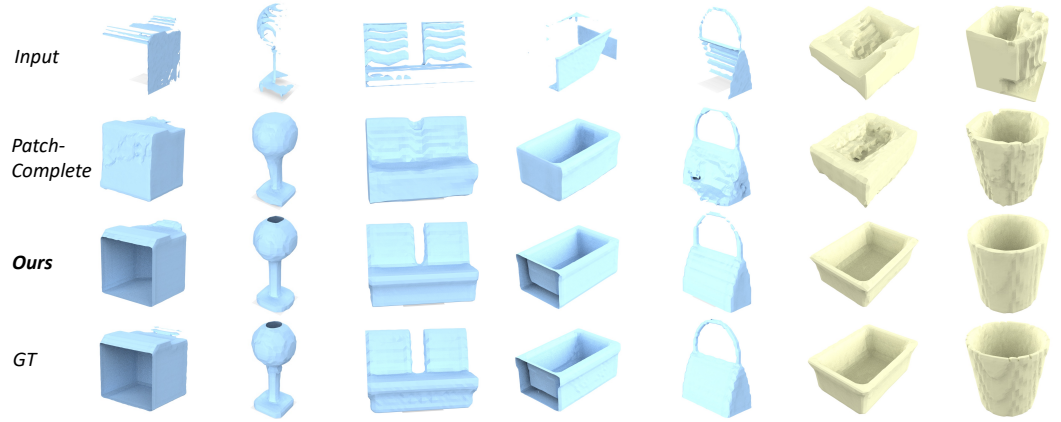


Figure 4: Shape completion on synthetic (blue) and real (yellow) objects of *entirely unseen* classes. Our method produces the completed shapes in superior quality given both synthetic and real data.

Regarding diffusion-based approaches, RePaint-3D is adapted from Repaint [78], a 2D diffusion-based inpainting method that only involves partial shape conditions during the inference process. In contrast, our DiffComplete explicitly matches each partial shape with a complete counterpart at the training stage, thereby improving the output consistency with the ground truths. PVD [12] is originally a point cloud diffusion model. Here, we adapt it to perform TSDF (TUDF) diffusion. A limitation of PVD is that it retains noises of the partial input. Hence, when it is mixed with the generated missing part, noise severely affects the final completion quality. Instead, we design two branches to separately process the partial and complete shapes. By doing so, the model can learn a diffusion process from noise to clean shapes.

**Completion on unseen object categories.** In two datasets of the PatchComplete benchmark, we compare the generalizability of DiffComplete against the state of the arts, including approaches particularly designed for unseen-class completion [81, 15]. As summarized in Table 2, our method exhibits the best completion quality on average for eight unseen object categories in the synthetic ShapeNet data, despite lacking zero-shot designs. The previous SOTA PatchComplete, mainly leverages the multi-scale structural information to improve the completion robustness. Our method inherently embraces this concept within the diffusion models. With our hierarchical feature aggregation, the network learns multi-scale local completion patterns, which could generalize to various object classes,

https://doi.org/10.52202/075280-3318

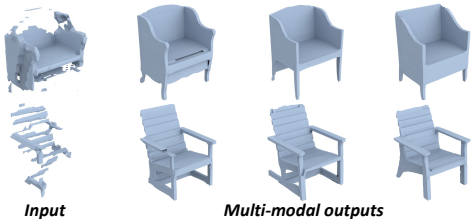


Figure 5: Our method produces multimodal plausible results given the same partial shape.

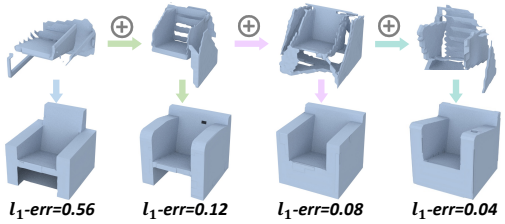


Figure 6: Our method incorporates multiple partial shapes to refine completion results ($l_1$-err. ↓).

Table 4: Multimodal capacity.

| Method | MMD ↓ | TMD ↑ | UHD ↓ |
|---|---|---|---|
| AutoSDF [6] | 0.008 | 0.028 | 0.061 |
| ShapeFormer [80] | 0.007 | 0.024 | 0.055 |
| PVD [12] | 0.007 | 0.027 | 0.042 |
| RePaint-3D [78] | 0.007 | **0.029** | 0.053 |
| Ours | **0.002** | 0.025 | **0.032** |

Figure 7: TMD curve.

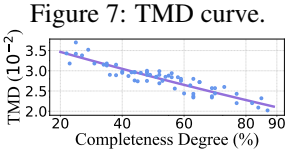


Table 5: Multi-condition results.

| Cond. Num. | $l_1$-err. ↓ | MMD ↓ | TMD ↑ |
|---|---|---|---|
| one | 0.07 | 0.002 | **0.025** |
| two | 0.05 | 0.001 | 0.021 |
| three | **0.04** | **0.001** | 0.019 |

Table 6: Effects of fusion choices for multiple conditional inputs.

| Strategy | $l_1$-err. ↓ | CD ↓ | IoU ↑ |
|---|---|---|---|
| simple average | 0.13 | 4.56 | 63.3 |
| w/o MLP $\psi$ | 0.23 | 5.88 | 57.4 |
| occ-aware (Ours) | **0.05** | **3.97** | **68.3** |

Table 7: Feat. aggregation levels.

| 1/8 | 1/4 | 1/2 | 1 | MMD ↓ | TMD ↑ | CD ↓ |
|---|---|---|---|---|---|---|
| ✓ | | | | 0.005 | **0.031** | 4.8 |
| ✓ | ✓ | | | 0.004 | 0.028 | 4.4 |
| ✓ | ✓ | ✓ | | 0.003 | 0.026 | 4.2 |
| ✓ | ✓ | ✓ | ✓ | **0.002** | 0.025 | **4.1** |

Table 8: Ablation on different feature aggregation mechanisms.

| Operation | $l_1$-err. ↓ | MMD ↓ | TMD ↑ |
|---|---|---|---|
| cross-attn. | 0.12 | 0.005 | **0.027** |
| concat. | 0.07 | 0.002 | 0.024 |
| addition (Ours) | **0.07** | **0.002** | 0.025 |

as their local structures are often shared. Our ablation study in Sec. 4.5 further validates this benefit. Table 3 demonstrates our method's superior performance with real-world scans, which are often cluttered and noisy. As showcased in Fig. 4, the 3D shapes produced by DiffComplete stand out for their impressive global coherence and local details.

### 4.3 Multimodal Completion Characteristics

**Quantitative evaluations.** Probabilistic methods have the intriguing ability to produce multiple plausible completions on the same partial shape, known as multimodal completion. For the 3D-EPN chair class, we generate ten results per partial shape with randomized initial noise $x_T$, and employ metrics from prior works [7, 6], *i.e.*, (i) MMD measures the completion accuracy against the ground truths, (ii) TMD for completion diversity, and (iii) UHD for completion fidelity to a partial input. Table 4 shows that our method attains much better completion accuracy and fidelity, while exhibiting moderate diversity. This aligns with our design choice of leveraging the control mechanism to prioritize completion accuracy over diversity. Yet, we can adjust this trade-off to improve shape diversity, as discussed in Sec. 4.5. Fig. 5 presents our multimodal outputs, all showing great realism.

**Effects of input completeness degree.** To verify the influence of input completeness degree on diversity, we select ten chair instances from ShapeNet, due to chair's large structural variation. For each instance, we create six virtual scans of varying completeness, and for each scan, we generate ten diverse completions to compute their TMD. Fig. 7 presents a negative correlation between the shape diversity (reflected by TMD) and input completeness degree, which is measured by the occupied voxel's ratio between the partial and complete GT shapes.

### 4.4 Multiple Conditional Inputs

**Quantitative evaluations.** As indicated in Table 5, DiffComplete consistently improves completion accuracy (lower $l_1$ error) and reduces diversity (lower TMD) when more conditional inputs are added. Fig. 1(b) and 6 showcase the progression of completion results when we gradually introduce more partial shapes of the same object (denoted by the plus symbol). Our model incorporates the local structures of all partial inputs and harmonizes them into a coherent final shape with the lower $l_1$ error.

**Effects of occupancy-aware fusion.** In Table 6, we compare our design with alternatives on the 3D-EPN chair class and PatchComplete benchmark using two conditional inputs. Averaging fea-


https://doi.org/10.52202/075280-3318

tures without occupancy masks largely lowers the completion accuracy due to disturbances from non-informative free-space features. Removing the learnable MLP layer $\psi$ hinders the network's adaptation from single to multiple conditions, also worsening the results. Instead, we adaptively aggregate multi-condition features and refine them to mitigate their discrepancy for reliable completions. Note that directly fusing original scans in TSDF (TUDF) space might be vulnerable to registration errors, as compared with our strategy in the supplementary file.

### 4.5 Ablation Study

**Effects of hierarchical feature aggregation.** Table 7 shows the effects of aggregating features of the complete and incomplete shapes at different decoder levels (see Eq. (4)). First, increasing feature aggregation layers (from single to hierarchical) consistently boosts the completion accuracy (lower MMD), while decreasing it yields better diversity (higher TMD). If connecting features only at the network layer with 1/8 resolution, we achieve the best TMD that surpasses other methods (see Table 4). Thus, the accuracy-diversity trade-off can be adjusted by altering the level of feature aggregation. Second, hierarchical feature aggregation facilitates unseen-class completion (lower CD, tested on PatchComplete benchmark). This improvement suggests that leveraging multi-scale structural information from partial inputs enhances the completion robustness.

**Effects of feature aggregation mechanism.** We ablate on the way to aggregate $F_x^i(x_t)$ and $\phi^i(F_c^i(f)$ in Eq. (4). As Table 8 shows, direct addition achieves the best completion accuracy (the lowest $l_1$-err and MMD), as it only combines features at the same 3D location to precisely correlate their difference for completion. Cross-attention disrupts this spatial consistency and yields less accurate results. Concatenation has a similar performance with addition, while the latter is more efficient.

## 5 Conclusion and Discussion

We presented DiffComplete, a new diffusion-based approach to enable multimodal, realistic, and high-fidelity 3D shape completion, surpassing prior approaches on completion accuracy and quality. This success is attributed to two key designs: a hierarchical feature aggregation mechanism for effective conditional control and an occupancy-aware fusion strategy to seamlessly incorporate additional inputs for geometry refinement. Also, DiffComplete exhibits robust generalization to unseen object classes for both synthetic and real data, and allows an adjustable balance between completion diversity and accuracy to suit specific needs. These features position DiffComplete as a powerful tool in various applications in 3D perception and content creation. Yet, as with most diffusion models, DiffComplete requires additional computation due to its multi-step inference. A comprehensive discussion on the limitation and broader impact is provided in the supplementary file.

https://doi.org/10.52202/075280-3318

## Acknowledgement

This work is partially supported by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R, Shenzhen Science and Technology Program KQTD20210811090149095, and the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14206320).

## References

[1] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum*, 37:625–652, 05 2018. 1

[2] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3):24:1–24:18, 2017. 1, 3

[3] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 3

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

[5] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 1

[6] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2, 7, 9

[7] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 281–296. Springer, 2020. 2, 3, 9

[8] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pages 87–96. PMLR, 2017. 3

[9] 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, 3

[10] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 2, 3, 4

[11] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. *arXiv preprint arXiv:2212.04493*, 2022. 3, 4

[12] 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, 7, 8, 9

[13] 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, 3, 4

 https://doi.org/10.52202/075280-3318

[14] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 2, 3, 6, 7, 17, 18

[15] Yuchen Rao, Yinyu Nie, and Angela Dai. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *arXiv preprint arXiv:2206.04916*, 2022. 2, 3, 6, 7, 8, 18, 23, 24

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 834–849. Springer, 2014. 3

[17] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[18] Thomas Whelan, Stefan Leutenegger, Renato F. Salas-Moreno, Ben Glocker, and Andrew J. Davison. Elasticfusion: Dense SLAM without A pose graph. In *Robotics: Science and Systems XI, Sapienza University of Rome, Rome, Italy, July 13-17, 2015*, 2015.

[19] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner. Intrinsic3d: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[20] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (TOG)*, 2015. 3

[21] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312, 1996. 3

[22] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard A. Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew J. Davison, and Andrew W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 559–568, 2011. 3

[23] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, Basel, Switzerland, October 26-29, 2011*, pages 127–136, 2011. 3

[24] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 3

[25] Silvan Weder, Johannes Schonberger, Marc Pollefeys, and Martin R Oswald. Routedfusion: Learning real-time depth map fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4887–4897, 2020. 3

[26] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 2020. 7, 19

[27] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2020.

[28] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1747–1756, 2021. 3

[29] Olga Sorkine and Daniel Cohen-Or. Least-squares meshes. In *Proceedings Shape Modeling Applications, 2004.*, pages 191–199. IEEE, 2004. 3

[30] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 381–389, 2006.

[31] Wei Zhao, Shuming Gao, and Hongwei Lin. A robust hole-filling algorithm for triangular mesh. *The Visual Computer*, 23:987–997, 2007. 3

[32] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing, Cagliari, Sardinia, Italy, June 26-28, 2006*, pages 61–70, 2006. 3

[33] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 3

[34] Sebastian Thrun and Ben Wegbreit. Shape from symmetry. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1824–1831. IEEE, 2005. 3

[35] Niloy J Mitra, Leonidas J Guibas, and Mark Pauly. Partial and approximate symmetry detection for 3d geometry. *ACM Transactions on Graphics (ToG)*, 25(3):560–568, 2006.

[36] Mark Pauly, Niloy J Mitra, Johannes Wallner, Helmut Pottmann, and Leonidas J Guibas. Discovering structural regularity in 3d geometry. In *ACM SIGGRAPH 2008 papers*. 2008.

[37] Ivan Sipiran, Robert Gregor, and Tobias Schreck. Approximate symmetry detection in partial 3d meshes. In *Computer Graphics Forum*, 2014.

[38] Pablo Speciale, Martin R Oswald, Andrea Cohen, and Marc Pollefeys. A symmetry prior for convex variational 3d reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 313–328. Springer, 2016. 3

[39] 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. 3

[40] Yangyan Li, Angela Dai, Leonidas Guibas, and Matthias Nießner. Database-assisted object retrieval for real-time 3d reconstruction. In *Computer graphics forum*, 2015.

[41] Liangliang Nan, Ke Xie, and Andrei Sharf. A search-classify approach for cluttered indoor scene understanding. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012.

[42] Young Min Kim, Niloy J Mitra, Dong-Ming Yan, and Leonidas Guibas. Acquiring 3d indoor environments with variability and repetition. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012. 3

[43] 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. 3

[44] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5431–5440, 2016. 3

[45] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 3

[46] 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.

 https://doi.org/10.52202/075280-3318

[47] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE international conference on computer vision*, pages 85–93, 2017.

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

[49] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 3, 7

[50] Angela Dai and Matthias Nießner. Scan2mesh: From unstructured range scans to 3d meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5574–5583, 2019. 3

[51] X Zheng, Yang Liu, P Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, 2022. 3, 7

[52] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired point cloud completion on real scans using adversarial training. *arXiv preprint arXiv:1904.00069*, 2019. 3

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

[54] Himanshu Arora, Saurabh Mishra, Shichong Peng, Ke Li, and Ali Mahdavi-Amiri. Multimodal shape completion via implicit maximum likelihood estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2958–2967, 2022. 3

[55] 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. PMLR, 2015. 3

[56] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 7

[57] 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. 4

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

[59] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 4, 22

[60] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4, 22

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6, 22

[62] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021. 3

[63] 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

[64] 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.

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

[66] Gimin Nam, Mariem Khlifi, Andrew Rodriguez, Alberto Tono, Linqi Zhou, and Paul Guerrero. 3d-ldm: Neural implicit 3d shape generation with latent diffusion models. *arXiv preprint arXiv:2212.00842*, 2022. 3, 4

[67] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023. 3

[68] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022. 3

[69] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12642–12651, 2023. 3, 22

[70] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 4, 17, 21

[71] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[72] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6

[73] 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. 6, 7, 19

[74] John Amanatides, Andrew Woo, et al. A fast voxel traversal algorithm for ray tracing. In *Eurographics*, 1987. 6

[75] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 7

[76] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[77] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7

[78] 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. 7, 8, 9

[79] Dongsu Zhang, Changwoon Choi, Inbum Park, and Young Min Kim. Probabilistic implicit scene completion. In *International Conference on Learning Representations*, 2022. 7

[80] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7, 9

 https://doi.org/10.52202/075280-3318

[81] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3818–3827, 2019. 7, 8

[82] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflake point deconvolution for point cloud completion and generation with skip-transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 19

[83] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. *arXiv preprint arXiv:2211.13449*, 2022. 22

[84] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *CVPR*, 2018. 22

[85] Tao Hu, Xiaogang Xu, Ruihang Chu, and Jiaya Jia. Trivol: Point cloud rendering via triple volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 22

[86] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 22