IPoD: Implicit Field Learning with Point Diffusion for Generalizable 3D Object Reconstruction from Single RGB-D Images

Yushuang Wu[1,2*]   Luyue Shi[1,2]   Junhao Cai[4]   Weihao Yuan[3]   Lingteng Qiu[1,2]
Zilong Dong[3]   Liefeng Bo[3]   Shuguang Cui[2,1]   Xiaoguang Han[2,1†]

[1]SSE, CUHKSZ   [2]FNii, CUHKSZ   [3]Alibaba Group   [4]HKUST

Abstract

Generalizable 3D object reconstruction from single-view RGB-D images remains a challenging task, particularly with real-world data. Current state-of-the-art methods develop Transformer-based implicit field learning, necessitating an intensive learning paradigm that requires dense query-supervision uniformly sampled throughout the entire space. We propose a novel approach, IPoD, which harmonizes implicit field learning with point diffusion. This approach treats the query points for implicit field learning as a noisy point cloud for iterative denoising, allowing for their dynamic adaptation to the target object shape. Such adaptive query points harness diffusion learning's capability for coarse shape recovery and also enhances the implicit representation's ability to delineate finer details. Besides, an additional self-conditioning mechanism is designed to use implicit predictions as the guidance of diffusion learning, leading to a cooperative system. Experiments conducted on the CO3D-v2 dataset affirm the superiority of IPoD, achieving 7.8% improvement in F-score and 28.6% in Chamfer distance over existing methods. The generalizability of IPoD is also demonstrated on the MVImgNet dataset. Our project page is at https://yushuang-wu.github.io/IPoD.

1. Introduction

3D reconstruction from a single-view image is a challenging problem that with widespread implications in fields such as robotics, autonomous driving, and AR/VR. Recent efforts have been directed towards developing a generalizable model for object reconstruction from real RGB-D data [28, 61]. It aims to learn a category-agnostic network to recover an accurate and complete shape from a single-view object RGB-D image but with only imperfect ground-truth (GT) supervision, which are usually point clouds reconstructed from multiple views inevitably containing noise and incompleteness as in real 3D datasets [46].
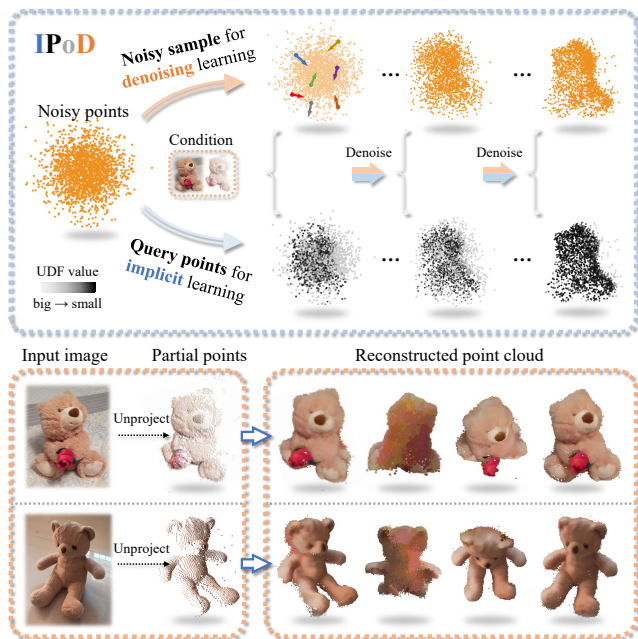

Figure 1. Our work focuses on the task of generalizable 3D object reconstruction from a single RGB-D image. The proposed method conducts implicit field learning with point diffusion that iteratively denoises a point cloud as adaptive query points for better implicit field learning, which leads to high reconstruction quality on both the global shape and fine details.

To tackle this problem, the state-of-the-art methods MCC [61] and NU-MCC [28] develop Transformer-based networks to learn an implicit field for reconstruction. These methods involve an intensive learning paradigm that demands dense query-supervision sampled uniformly in a predefined bounded space. Simultaneously, denoising diffusion models, notable for their emergent role in generative modeling, have achieved remarkable outcomes in a variety of 2D and 3D tasks [12, 19, 23, 27, 29, 45, 48, 55]. The diffusion models prove powerful in generation especially given large amounts of data for training. Consequently, with the development of large-scale, realistic 3D datasets [46, 65], the integration of diffusion models into

---

* Work done during internship supervised by Weihao Yuan at Alibaba.
† Corresponding author: hanxiaoguang@cuhk.edu.cn.

this field is anticipated to further enhance problem-solving capabilities.

A subset of methods directly use 2D diffusion models for 3D reconstruction, which first pre-train an auto-encoder to represent object shapes into the latent space and learn to denosie a random noise of latent code (or tri-plane features) under specific conditions such as images or text [2, 7, 10, 18, 53, 66, 69]. The denoised latent code is subsequently utilized by the well-trained decoder to generate the reconstructed shape. Conversely, another approach simplifies this process by extending 2D diffusion into 3D [32, 33, 35, 67, 68]. It works on iteratively denoising a noisy 3D point cloud or voxels under given conditions to conduct reconstructing that has achieved remarkably generation results. Based on this, our work explores introducing the point diffusion into the target task.

In this paper, we propose to integrate **I**mplicit field learning with **Po**int **D**iffusion to address the target task, named as **IPoD**. The proposed integration approach is novel, simple, yet effective: we perceive the query points as a noisy point cloud to denoise. Specifically, a bunch of noisy points are sampled not only for denoising as in diffusion learning, but also as spatial queries for implicit field learning, so that our model iteratively conducts denoising and implicit predicting to recover the target shape under the RGB-D condition, as shown in Fig. 1. In this process, the sampled points can gradually get close to the true object shape through denoising, which can provide adaptive query positions according to the object shape for more effective implicit field learning. In comparison, a pure implicit field learning method queries all possible positions aimlessly and views them equally, while our method can more effectively conduct implicit field learning by attending more valuable local regions near the object surface, thus also capturing the fine shape details more easily. Further, we propose a novel self-conditioning mechanism [4], which leverages the predicted implicit values to reversely assist the diffusion learning and thus forges a cooperative system. Specifically, we predict the unsigned distances of all noisy points away from the true object surface and employ them as the condition of denoising at the next time step. The implicit predictions actually indicate the final target shape that is useful to guide every one-step denoising. The proposed method actually leads to a simple framework that conducts point diffusion learning and implicit field learning concurrently but well combines the advantages of both: the diffusion model for recovering the global coarse shape and the implicit field learning for giving accurate predictions on local fine details.

We conduct experiments on the CO3D-v2 [46] dataset and demonstrate the superiority of the proposed approach, which surpasses the state-of-the-art results by $\sim$7.8% of F-score and $\sim$28.6% of Chamfer distance in average. In addition, we clean 100k point clouds reconstructed in MVImgNet [65] and show that (i) trained on CO3D-v2,

our method is generalizable to not only unseen categories in CO3D-v2 but also other various categories in MVImgNet; (ii) using the cleaned MVImgNet point clouds for training can help further improve the generalizability.

In summary, our key contributions are as follows:

- We propose IPoD that conducts implicit field learning with point diffusion for generalizable 3D object reconstruction from single RGB-D images, where the diffusion model provides adaptive queries for a more effective implicit field learning.
- We design a novel self-conditioning mechanism that leverages the implicit predictions to reversely assist the denoising thus leading to a mutually beneficial system.
- We conduct extensive experiments to show the superiority of IPoD and the effectiveness of each component. IPoD achieves state-of-the-art reconstruction results on the CO3D-v2 dataset [46].

2. Related Work

Single-view 3D reconstruction One line of methods on this task train neural networks supervised by CAD [16, 60], voxels [15, 62], point clouds [14, 37], or meshes [25, 63]. However, they mainly focus on simplistic synthetic data with perfect supervision [59, 64], *e.g.* from the ShapeNet dataset [1] or work on one or several categories [5, 17, 21, 30] as on the Pix3D dataset [57]. Our work follows two more recent works, MCC [61] and NU-MCC [28], that tackle the problem of category-agnostic reconstruction from a single-view RGB-D image on real datasets [46, 65]. MCC first introduces a large Transformer-based network for occupancy field learning, and NU-MCC further proposes a repulsive unsigned distance field (Rep-UDF) for finer reconstruction and a neighborhood decoder that speeds up the inference. Different in methodology, MCC-style methods take pure implicit learning for reconstruction, while our method integrates diffusion models for better reconstruction quality on both coarse shapes and fine details.

Diffusion models Since proposed in [19, 41, 55], diffusion models have rapidly emerged as a popular family of generative models. Methods based on diffusion technology have shown impressive quality, diversity, and expressiveness in various tasks such as image synthesis [12, 19, 20, 41], super-resolution [26, 48], image editing [6, 36, 47, 54], and 3D vision tasks like shape completion [11, 33, 68], text/image-to-3d [3, 18, 40, 45, 67], and 3D reconstruction [7, 10, 22, 35, 52]. Although applied on the 3D tasks, most methods develop 1D or 2D diffusion models by first pre-training an auto-encoder to encode the input into the 1D or 2D latent space and conduct diffusion learning on the latent space [2, 7, 10, 18, 53, 66, 69]. Zhou *et al.* [68] first extend diffusion models into 3D and propose point-voxel diffusion (PVD) to iteratively denoise a noisy point cloud for 3D shape generation and reconstruction. The 3D diffusion

model is attracting increasing attention [32, 33, 35, 67, 68]. Closed to our target task, PC^2 [35] applies PVD on real-world reconstruction from one image with a known camera pose, which projects image features to the point cloud as the diffusion condition. Different from our work, PC^2 focuses on reconstruction from posed images without depth and conducts category-specific learning. In methodology, our method is also based on point cloud diffusion models, but differently, ours integrates implicit predictions as the denoising guidance for more accurate noise estimation.

Implicit field learning   Implicit field learning is a powerful approach that learns an implicit field that assigns a value to each position as a representation. It has been widely used for 3D shape representation using occupancy field [42, 44, 49, 61], signed/unsigned distance field (SD-F/UDF) [9, 28, 43], implicit feature field [8], radiance field [34, 38], etc. Although impressive results have been achieved, implicit field learning suffers from the large training cost because it usually demands dense position querying and implicit value supervising to well fit the field. Such an intensive learning paradigm is also used in the SOTA methods [28, 61] on the target task. However, our method integrates implicit field learning with diffusion learning, which can provide adaptive query positions via iteratively denoising the query points. In this way, the implicit field learning can more effectively extract the valuable information in data, which especially eases the recovery of fine details.

3. Method

In this section, we first formulate the target problem and introduce the solutions based on implicit field learning and point diffusion models. Then we give the formulation of the proposed IPoD that integrates the two solutions. Finally, we introduce the design of our self-conditioning mechanism.

3.1. Preliminary

Problem Formulation   The task of this work aims to recover a 3D point cloud $X \in \mathbb{R}^{N \times 3}$ from a RGB-D input, which is usually processed into an image $I \in [0, 255]^{H \times W \times 3}$ of size H and W and a partial point cloud $P \in \mathbb{R}^{M \times 3}$ unprojected from I with depth information, where N and M denote the point numbers and $M = HW$ if without filtering or down-sampling. All point clouds are normalized with zero-mean and unit-variance as the CO3D [46] coordinate system.

Implicit Field Learning   This solution aims to learn an implicit field: $f : (x, y, z) \rightarrow v$, where (x, y, z) denotes any position in the 3D space, v is an implicit value, *e.g.* a UDF value $udf \in \mathbb{R}^+$ representing the distance to the object surface. In training, a bunch of N query points $Q = \{(x, y, z)\} \subseteq \text{U}(-b, b)^3$ are uniformly sampled from a bounded space of scale $2b$, and their corresponding implicit values $\nu \in \mathbb{R}^{N \times 1}$ are computed as supervision. Thus given

P and I as references, the implicit field learning network f_θ aims to learn:

$$f_\theta(Q \mid P, I) \rightarrow \nu, \tag{1}$$

so that the target shape can be obtained via densely sampling query points and preserving those with desired implicit values to derive an output point cloud. The objective function for training is usually to minimize an $L1$ distance:

$$\mathcal{L}_{\text{imp}} = \left\| f_\theta(Q \mid P, I) - \nu \right\|_1, \tag{2}$$

Point Diffusion Models   Diffusion denoising probabilistic models are inspired by a thermodynamic diffusion process. It works by iteratively adding noise to a sample $X_0 \sim q(X_0)$ from the target data distribution $q(X_0)$ and finally into purely random noise, and the generative model is formed by reversing the Markovian noising process. The noising stepsize in the diffusion process is defined by a variance schedule $\{\beta_t\}_{t=0}^T$:

$$q(X_t | X_{t-1}) = \mathcal{N}(X_t; \sqrt{1 - \beta_t} X_{t-1}, \beta_t \mathbf{I}), \tag{3}$$

where $q(X_t | X_{t-1})$ is a normal distribution, so that $q(X_t | X_0)$ can be modeled by a reparameterization trick:

$$q(X_t | X_0) = \sqrt{\bar{\alpha}_t} X_0 + \epsilon \sqrt{1 - \bar{\alpha}_t}, \tag{4}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = 1 - \prod_{s=0}^t \alpha_s$, and the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. A generative network g_θ is learned to approximate $q(X_{t-1} | X_t)$, so that a sample $X_0 \sim q(X_0)$ can be generated by starting from a random sample $X_T \sim \mathcal{N}(0, \mathbf{I})$ and then iteratively sampling from the estimated $q(X_{t-1} | X_t)$ for denoising.

The target task can be tackled by a conditional point cloud diffusion model, where a network g_θ is learned to denoise a point cloud $X_T \sim \mathcal{N}(0, 1)^{N \times 3}$ sampled from a spherical Gaussian ball into the original object $X_0 = X$ given condition P and I. At each denoising step t, g_θ is required to predict the noise $\epsilon \sim \mathcal{N}(0, 1)^{N \times 3}$ added in the most recent time step in X_t:

$$g_\theta(X_t, t \mid P, I) \rightarrow \epsilon, \tag{5}$$

where both X_t and t act as the input of g_θ. At the inference stage, we recover the mean value of the approximated $q(X_t | X_{t-1})$ from the prediction $g_\theta(X_t, t \mid P, I)$, and a sample from the distribution $q(X_t | X_{t-1})$ can be obtained with this mean to update X_t. When the time step gets sufficiently small, the denoised X_t can well approximate the shape of X. The objective function for optimizing the parameters in a diffusion model g_θ is usually to minimize an $L2$ distance:

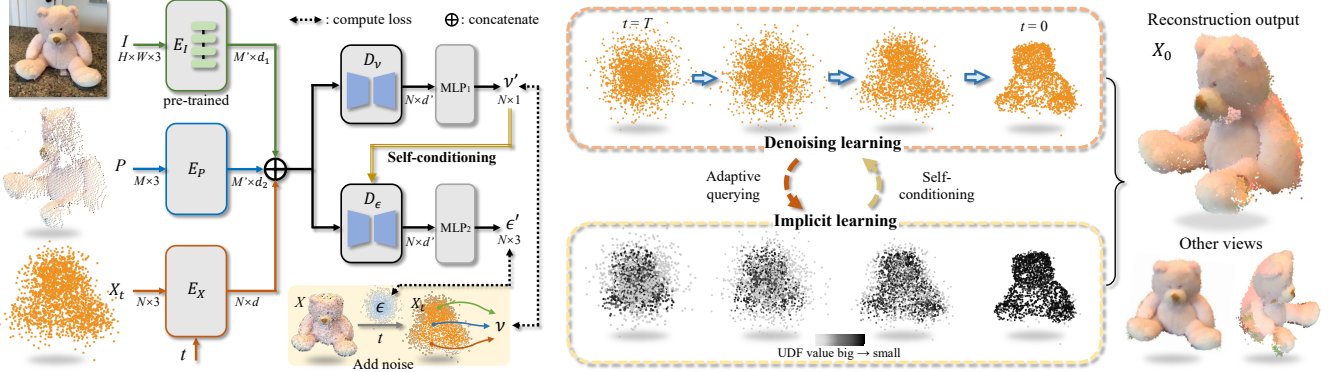$$\mathcal{L}_{\text{diff}} = \left\| g_\theta(X_t, t \mid P, I) - \epsilon \right\|_2, \tag{6}$$

Figure 2. Overview of the proposed method. The network takes a single-view image and a partial point cloud unprojected from the image according to the depth information as the input. A bunch of points are sampled both as a noisy point cloud for point diffusion learning and also as query points for implicit field learning. The proposed self-conditioning mechanism leverages the implicit predictions to reversely assist denoising. The reconstruction result can be obtained via iteratively conducting implicit predicting and denoising concurrently.

3.2. Implicit Field Learning with Point Diffusion

We propose to integrate the implicit field learning and point diffusion into one framework, where a bunch of points are sampled not only as spatial queries for implicit field learning but also as a noisy sample for denoising to approximate the object shape.

See Fig. 2 for an overview of the proposed IPoD. The spatial queries start from a randomly sampled noise that can be viewed as a point cloud $X_T \sim \mathcal{N}(0,1)^{N\times3}$. As in the inference stage of diffusion learning, it gradually gets close to the target shape via an iterative denoising process $t = T, T-1, \cdots, 0$ under the condition of I and P. A network conducts implicit value predicting on these adaptive queries X_t concurrently and produces the final shape when $t = 0$. Denoting the new network as h_θ, at any time step t, we estimate the noise ϵ in X_t and the UDF value ν at the position of each point in X_t:

$$h_\theta(X_t, t \mid P, I) \to (\epsilon, \nu). \qquad (7)$$

For the training of h_θ, we randomly sample a noise $\epsilon \sim \mathcal{N}(0,1)^{N\times3}$ and $t \in [1,T]$ at each iteration, and the noise ϵ is added into the GT point cloud X to produce X_t according to Eq. 4. We also compute the distance of each point in X_t to the nearest one in X as the supervision ν. Then the network h_θ takes X_t and t as input to estimate ϵ and ν conditioned by P and I as in Eq. 7. We optimize the parameters in h_θ by jointly minimizing the losses in Eq. 2 and Eq. 6:

$$\mathcal{L}_{\text{uni}} = \left\| \nu' - \nu \right\|_1 + \lambda \left\| \epsilon' - \epsilon \right\|_2, \qquad (8)$$

where ν', ϵ' are the prediction of h_θ, and $\lambda \in \mathbb{R}^+$ is a weighting factor. Note that our model can also predict the target object color by extending the implicit value ν with RGB values $\nu = \{udf, rgb\} \in \mathbb{R}^{N\times4}$, whose supervision can be obtained from the color of the nearest point in X within a pre-defined small distance ρ.

Compared with pure implicit field learning, ours exploits the generation power of point diffusion models to conduct more effective querying, where the query points can adaptively get close to the target shape rather than being sampled aimlessly. With the coarse shape indicated by the denoised query points, the network can better capture local fine details via implicit prediction.

Implementation In the proposed framework, the concrete implementation of h_θ follows the work of [28, 35, 61]. We provide two versions of implementation based on PVCNN [31] and Transformer [58], respectively. As illustrated in Fig. 3, the condition image I is first fed into a Vision-Transformer [13] (ViT) encoder E_I (well pre-trained and frozen), where a patch embedding is adopted to down-sample and serialize the image input and several Transformer layers then extract features in the shape of $M' \times d_1$, of which M' is the sequence length and d_1 is the feature dimension. The encoding of P is similar, an encoder E_P extracts the features of P in the shape of $M' \times d_2$. In the encoder E_X for encoding X_t, the time step t is first embedded into a vector of length e' and a linear layer is adopted to project it into two values $\{scale, shift\}$, which then serve as the affine factors to transform the embedding of X_t. The final feature of X_t is finally obtained through a linear layer. In the Transformer-based implementation, we employ the similar anchor prediction operation following NU-MCC [28], which further encode the features of I and P into M'' anchors with positions and features. Note that our method is independent to this operation. In the PVCNN-based implementation, we follow the projection manner in PC^2 to project the image features onto all positions in P and X_t and extend their features via concatenation into $(M+N) \times d$ with $d = d_1 + d_2$. In the decoding stage, we use two decoders with the same architecture except the input and output dimension for the UDF ν' and noise ϵ' prediction, respectively. The UDF prediction ν' is first computed and sent into the the other decoder,

which is concatenated with the encoded feature of X_t as the self-condition for the noise prediction. More details can be found in Sec. 4 and supplementary materials.

3.3. Self-conditioning

The adaptive queries provided by denoising benefit the implicit field learning to place greater importance on recovering the fine shape. We further propose a self-conditioning technique to exploit the implicit field learning to reciprocally assist the diffusion learning, thus forming a mutually beneficial system.

The typical self-conditioning technique is first proposed in [4], where the model is directly conditioned by the previously generated variable during the sampling process, *e.g.* $\widetilde{X_0}$ namely the approximated X_0 at each time step as used in [4]. It is proposed to leverage the internal approximation to improve the prediction accuracy of the diffusion model.

We propose a novel self-conditioning method by taking the predicted implicit value ν' as the self-condition. Compared with using $\widetilde{X_0}$, ours can provide more accurate self-condition information, because the errors in $\widetilde{X_0}$ are usually significant when t is large at both the training and inference stage, while the approximated implicit values ν are relatively independent of the time step variable.

The self-conditioning mechanism works as illustrated in Fig. 3. The UDF prediction ν' is first computed and sent into the the other decoder. As $\nu' \in \mathbb{R}^{N \times 1}$ provides point-wise information, we simply concatenate it with the embedding of X_t in the feature dimension. The concatenated features then go through the decoder and an MLP for the noise prediction. At the inference stage, the self-condition is initialized with a vector with all negative values (*e.g.*, -1) and updated with ν' at each time step.

Compared with a naive denoising diffusion model as in Eq. 5, the self-condition indicates the estimated unsigned distance of each point to the GT shape surface so that richer information about the target shape can be provided to assist the noise prediction.

4. Experiments

Datasets We use CO3D-v2 [46] as the main dataset following MCC [61] and NU-MCC [28]. It consists of around 37k videos of 51 object categories, of which 10 are held out for evaluation and the remaining 41 for training. The 10 held-out categories are the same as the ones selected by MCC and NU-MCC, as listed in the supplementary materials. In CO3D-v2, the object shape annotations are obtained via COLMAP [50, 51] and thus inevitably contain noise and voids. We test the zero-shot generalization ability of the proposed method on the dataset of MVImgNet [65], which is a real-world dataset with 220k object videos in 238 categories, and their 3D annotations are also obtained via COLMAP. Note that videos of MVImgNet are captured in
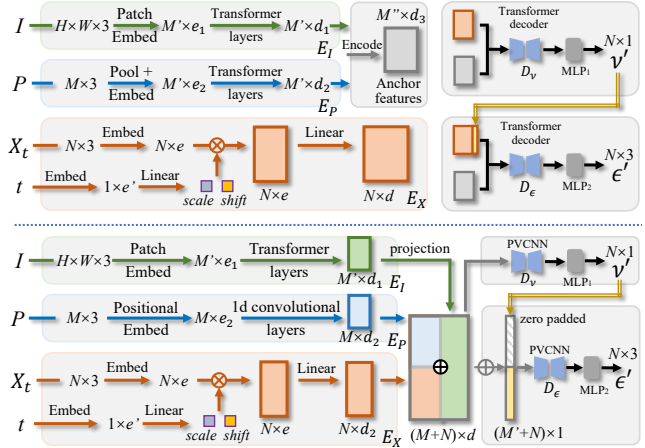

Figure 3. Illustration of the Transformer-based (upper part) and the PVCNN-based (lower part) implementations. $\otimes$ denotes the affine operation. The yellow arrow with double lines indicate the proposed self-conditioning mechanism.

180° and thus can only provide incomplete 3D shape annotations. We also contribute a dataset with 100k cleaned point clouds from MVImgNet. We hire annotators to manually filter the 3D annotations with low quality and remove the background noise caused by COLMAP estimation for the rest of the point clouds.

Evaluation metrics Following previous works [28, 39, 56, 61], we use 6 metrics to evaluate the reconstruction quality. The metrics can be divided into two groups for measuring (i) the absolute distance: the Chamfer distance (CD) and its two components that measure the distance in two different directions (Acc and Comp); and (ii) the relative degree of recovery: the precision ratio (Prec) that indicates the percentage of predicted points within a small distance threshold ρ to any GT point, the recall ratio (Recall) that indicates the percentage of GT points within ρ to any predicted point, and their F-score (F1). Among them, F1 holds the dominant since both accuracy and completeness are considered. The mathematical formulation of all metrics are detailed in the supplementary materials.

Baselines We compare the proposed method with four baselines. The first one is PC^2 [35] based on the diffusion model, implemented with a PVCNN backbone. It is designed for single-image reconstruction and conducts category-specific learning. We propose a modified PC^2 (PC^2-depth) by integrating the depth condition into the diffusion model, where the single-view depth is unprojected into a partial point cloud as the diffusion condition and concatenated with X_t before being fed into the PVCNN for X_t denoising. Another baseline is MCC [61] which is implemented with a Transformer-based encoder-decoder, which conducts implicit learning based on an occupancy field. The last one is NU-MCC [28] that improves MCC by proposing a Repulsive UDF to replace the occupancy field and applying anchor representations for higher efficiency.

Table 1. Results on CO3D-v2, averaged on all samples from 10 held-out categories. The best results are highlighted in bold font.

Method	Backbone	Acc↓	Comp↓	CD↓	Prec↑	Recall↑	F1↑
PC^2 [35]	PVCNN	0.342	0.214	0.556	24.2	56.2	33.0
PC^2-depth	PVCNN	0.209	0.103	0.312	61.7	87.6	70.7
MCC [61]	Transformer	0.172	0.144	0.316	68.9	72.7	69.8
NU-MCC [28]	Transformer	0.121	0.146	0.266	79.2	84.0	80.9
Ours1	PVCNN	0.163	0.089	0.252	69.0	89.7	76.2
Ours2	Transformer	**0.104**	**0.087**	**0.190**	**85.1**	**90.1**	**87.2**

Implementation details The size of input image I is 224×224, P and I are divided into patches of size 16×16 for ViT encoding. The encoded feature dimensions of I and P are 768 ($d_1 = d_2 = 768$), and the decoded feature dimension is also 768 ($d' = 768$). The UDF value supervision is clamped with a max value of 0.5. Besides, T is set as 1,000 in our diffusion model, $\lambda = 1.0$ in training, and $N = 50k$, the distance threshold $\rho = 0.1$ for evaluation. Our model is trained with a batch size of 64 for 100 epochs (taking around 48 hours on NVIDIA V100 GPUs), and an Adam [24] optimizer with a base learning rate of 10^{-4} is used following MCC and NU-MCC. We implement the proposed approach with two versions that use PVCNN (Ours1) and Transformer (Ours2) as the decoder backbone, respectively. The version of "Ours1" is constructed by adding the implicit learning branch and the self-conditioning mechanism into the diffusion-based PC^2 [35], and for "Ours2", we integrate the query points denoising branch into NU-MCC [28] that constructed based on implicit field learning and use the predicted UDF values as the self-condition to guide the denoising. See supplementary materials for more model architecture details of PVCNN-based and Transformer-based implementations.

4.1. Results on CO3D-v2

We show the evaluation results of the proposed method on the CO3D-v2 dataset in Tab. 1. No matter when implemented with PVCNN or Transformer as the backbone, our method can get better results than other existing methods. With PVCNN, our method improves the performance of the baseline PC^2-depth by 19.2% on Chamfer distance and 7.8% on F-score. Based on Transformer, our method achieves SOTA performance, which surpasses the previously best algorithm NU-MCC overall metrics, specifically by 28.6% on Chamfer distance (0.266→0.190) and by 7.8% on F-score (80.9%→87.2%).

We visualize the reconstruction results of each method in Fig. 6. As shown, the proposed method can produce better generations than the baseline method (Ours1 v.s. PC^2-depth and Ours2 v.s. NU-MCC) on both higher precision in local details and better completeness in coarse shapes. Besides, our method implemented based on Transformer achieve the best qualitative results over all baseline methods, especially for objects with more complex geometry structures. More
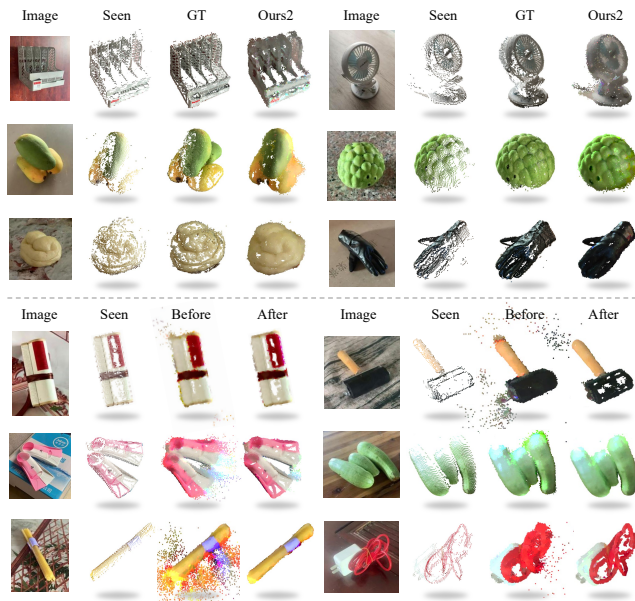

Figure 4. Visualization on MVImgNet data. Upper: generalization results by Ours2; Lower: the comparison of generalization results before and after fine-tuning on cleaned MVImgNet data.

qualitative results of our method on the CO3D-v2 dataset are included in the supplementary materials. We also visualize reconstructions of samples from seen categories in the supplementary materials.

4.2. Results on MVImgNet

We evaluate the generalizability of our method implemented based on Transformer (Ours2) on the MVImgNet dataset. Considering the incompleteness of GT shapes in MVImgNet, we do not consider the quantitative results. As presented in Fig. 4, our method can well generalize to more various categories of objects than in the CO3D-v2 dataset. Meanwhile, there are also some cases of over-predicting as in the lower part of Fig. 4, where Ours2 predicts extra noisy points in results. We further use the cleaned MVImgNet point clouds to fine-tune the network and found that the generations are endowed with higher accuracy, which indicates that the cleaned data can further improve the model's generalizability. Note that the categories in the cleaned data for fine-tuning have no overlap with the ones for evaluation. This experiment can also validate the generalizability of
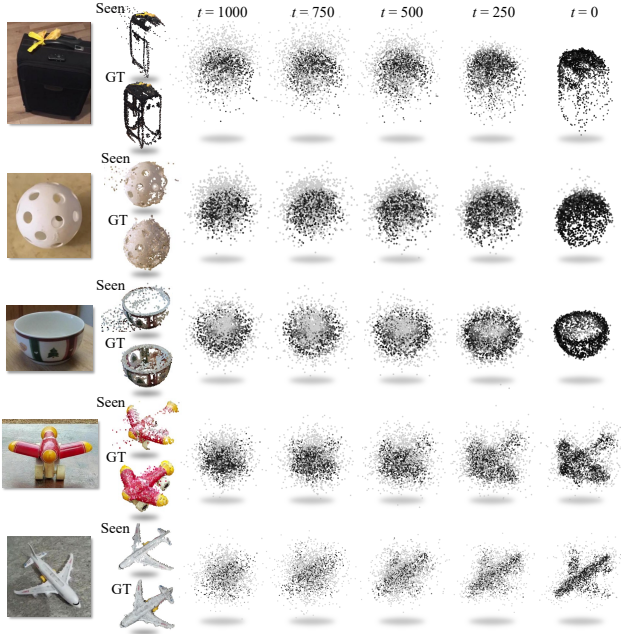
Figure 5. Visualization of the denoising process ($t=\{1000, 750, 500, 250, 0\}$) of our method in inferring. Note we only sample 2k points in each noisy sample for better visualization. The darkness of each point indicates the magnitude of the predicted UDF value (the smaller, the darker) at this position.

point diffusion models when learning from large-scale data. More generalization results of our method are included in the supplementary materials.

4.3. Ablation Study

Our approach mainly consists of three components: (i) the point diffusion learning provides adaptive query points for more efficient implicit learning, (ii) we supervise the UDF value of each point in a noisy point cloud for implicit field learning, and (iii) a self-conditioning mechanism is designed to facilitate its benefit to the denoising learning. Ablative analyses are conducted on them.

Individual impact To analyze the impact of the three components above, we evaluate the precision, recall, and F-score of each variant. Adding diffusion learning only into NU-MCC can bring an obvious improvement by absolute 4.9% on F-score (80.9%→85.8%), and further adding self-conditioning also makes a positive effect (85.8%→87.2%), as shown in Tab. 2(c, d). When adding UDF learning into PC^2-depth, the F-score is raised absolutely by 2.6% (70.7%→73.3%), and adding self-conditioning can further bring an absolute 2.9% gain (73.3%→76.2%), which proves the effectiveness of self-conditioning.

Point diffusion We visualize the denoising process in Fig. 5 to better validate the effectiveness of the point diffusion learning. Starting from purely noisy samples, the point clouds gradually get close to the target shapes according to the inputs. Notably, when addressing objects with

Table 2. Ablative results on three components: (i) Diff.: using diffusion learning; (ii) UDF: using UDF supervision; (iii) Self.: using the proposed self-conditioning mechanism. The upper section of the table includes results when implemented with PVCNN, and the lower includes Transformer-based results.

Method	Diff.	UDF	Self.	Prec↑	Recall↑	F1↑
PC^2-depth	✓			61.7	87.6	70.7
(a)	✓	✓		65.8	87.4	73.3
(b)	✓	✓	✓	**69.0**	**89.7**	**76.2**
NU-MCC		✓		79.2	84.0	80.9
(c)	✓	✓		84.2	85.9	85.8
(d)	✓	✓	✓	**85.1**	**90.1**	**87.2**

Table 3. Results of using different variants of self-condition. "None" denotes not using any self-conditioning.

Self-condition	Prec↑	Recall↑	F1↑
(a) UDF value	**69.0**	89.7	**76.2**
(b) UDF + RGB value	68.9	89.8	**76.2**
(c) Occupancy value	67.8	87.8	74.5
(d) $\widetilde{X_0}$ (classical)	63.2	**90.3**	72.7
(e) None	61.7	87.6	70.7

Table 4. Effects of self-conditioning at different denoising stages. The whole denoising process with 1k steps in inferring is evenly divided into four stages.

1k-750	750-500	500-250	250-0	Prec↑	Recall↑	F1↑
	✓	✓	✓	67.5	88.8	75.1
✓		✓	✓	56.0	86.7	65.5
✓	✓		✓	63.7	88.0	72.3
✓	✓	✓		59.9	85.1	68.1
✓	✓	✓	✓	**69.0**	**89.7**	**76.2**

relatively simple geometry, the point cloud at $t = 0$ itself can well capture the shape. However, for more complex geometries (*e.g.* toy plane in the last two rows of Fig. 5), the noise in point clouds can not be perfectly diminished. In this situation, implicit values can still well indicate the accurate shapes that complement the denoised point clouds.

Self-conditioning We conduct another group of analysis on the choice of self-condition: (a) using predicted UDF values (ours), (b) using predicted UDF and RGB values, (c) using predicted Occupancy values, (d) using $\widetilde{X_0}$ as in a classical manner, and (e) using none of them above. As shown in Tab. 3, using RGB values additionally can not lead to an extra performance gain, and replacing UDF with the Occupancy field can cause a performance drop. A potential reason is that UDF values can provide more fine-grained information about the difference between the noisy sample and the target shape. Using $\widetilde{X_0}$ results in a worse F-score than using occupancy values, but better than not using. This result validates the claim in Sec. 1 that the proposed self-conditioning mechanism can provide more accurate and useful information about the target shape to improve the noise prediction in point diffusion learning.

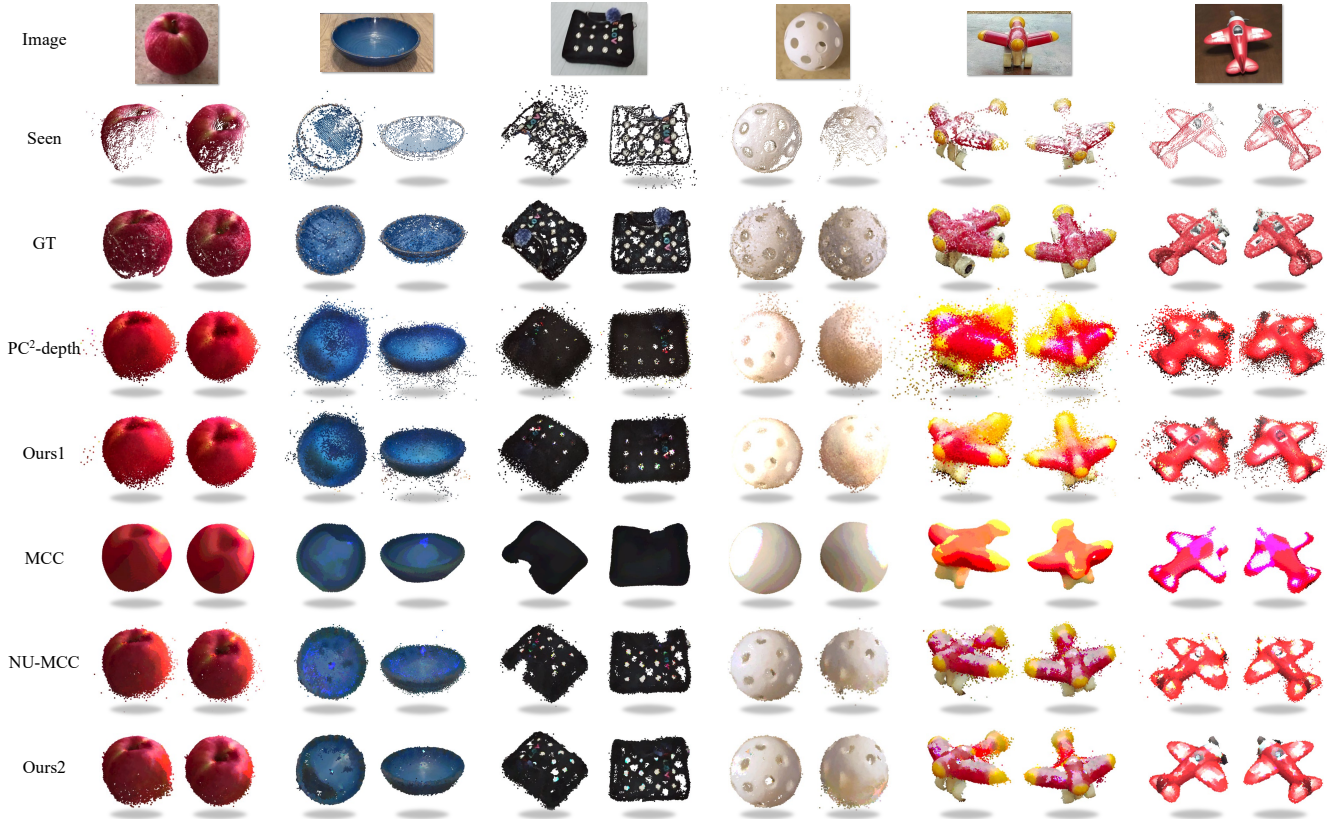Besides, we also delve into an in-depth analysis of the

Figure 6. Visualization of reconstructions by different methods on CO3D-v2 unseen categories. We choose two views for each sample.

impact of self-conditioning at different denoising stages. Dividing the whole denoising process into four stages of even length (per 250 time steps), we close the self-conditioning at each single stage in inferring, which leads to the results in Tab. 4. As shown, the first stage has the smallest effect, while the second and the last stages have significant impacts on the reconstruction quality. A possible reason is the second stage takes a role in capturing the coarse shape from a pure noise and the last stage is important for points' getting close enough to the target shape to ease the final UDF estimating.

5. Conclusion

In this paper, we have introduced IPoD, a powerful approach based on implicit field learning and point diffusion, to address the problem of generalizable 3D object reconstruction from single RGB-D images. We propose to conduct implicit field learning with adaptive queries through point denoising that helps the model better capture both the global coarse shape and local fine details. We also develop a self-conditioning mechanism to leverage implicit predictions to reversely assist the noise estimation in diffusion learning, which eventually forges a cooperative system. We implement our method based on two kinds of backbones, PVCNN and Transformer, and evaluate them on the CO3D-

v2 dataset. Experiments show that our method can achieve impressive reconstruction results, which quantitatively outperforms previous SOTA methods on both reconstruction completeness and precision. Results on MVImgNet also show that our method can well generalize to more various categories from another dataset.

Limitations We have not validated the effectiveness of our method on 3D human and scene reconstruction. Human shapes often contain more fine-grained details that may bring new challenges, and 3D scenes are also hard to reconstruct considering the serve occlusion causing poor quality in single-view inputs. These will be our future work.

References

[1] 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. 2

[2] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. 2

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[4] Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2023. 2, 5

[5] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 2

[6] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4054–4062, 2023. 2

[7] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4456–4465, 2023. 2

[8] 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 (CVPR)*, pages 6970–6981, 2020. 3

[9] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33:21638–21652, 2020. 3

[10] 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 (ICCV)*, pages 2262–2272, 2023. 2

[11] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. Diffcomplete: Diffusion-based generative 3d shape completion. *arXiv preprint arXiv:2306.16329*, 2023. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 1, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2020. 4

[14] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. 2

[15] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–499, 2016. 2

[16] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9785–9795, 2019. 2

[17] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2020. 2

[18] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 1, 2

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[21] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[22] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18423–18433, 2023. 2

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 1

[24] Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2015. 6

[25] Nilesh Kulkarni, Justin Johnson, and David F Fouhey. What's behind the couch? directed ray distance functions (drdf) for 3d scene reconstruction. *arXiv preprint arXiv:2112.04481*, 2021. 2

[26] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single

image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[27] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023. 1

[28] Stefan Lionar, Xiangyu Xu, Min Lin, and Gim Hee Lee. Numcc: Multiview compressive coding with neighborhood decoder and repulsive udf. *arXiv preprint arXiv:2307.09112*, 2023. 1, 2, 3, 4, 5, 6

[29] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 1

[30] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *arXiv preprint arXiv:1901.05567*, 2019. 2

[31] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Pointvoxel cnn for efficient 3d deep learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 4

[32] 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 (CVPR)*, pages 2837–2845, 2021. 2, 3

[33] Zhaoyang Lyu, Zhifeng Kong, XU Xudong, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[34] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021. 3

[35] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12923–12932, 2023. 2, 3, 4, 5, 6

[36] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2021. 2

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. 2

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF:

Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 3

[39] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 414–431, 2020. 5

[40] 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. 2

[41] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the ACM International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021. 2

[42] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5379–5389, 2019. 3

[43] 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 (CVPR)*, pages 165–174, 2019. 3

[44] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 523–540, 2020. 3

[45] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2023. 1, 2

[46] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 1, 2, 3, 5

[47] 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 (CVPR)*, pages 10684–10695, 2022. 2

[48] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 45(4): 4713–4726, 2022. 1, 2

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 3

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 4104–4113, 2016. 5

[51] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2016. 5

[52] Jaehyeok Shim, Changwoo Kang, and Kyungdon Joo. Diffusion-based signed distance fields for 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20887–20897, 2023. 2

[53] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023. 2

[54] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34:12533–12548, 2021. 2

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the The International Conference on Learning Representations (ICLR)*, 2020. 1, 2

[56] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15598–15607, 2021. 5

[57] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 2

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 4

[59] 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 (ICCV)*, pages 3818–3827, 2019. 2

[60] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2

[61] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9065–9075, 2023. 1, 2, 3, 4, 5, 6

[62] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape recon-

struction via 2.5 d sketches. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2

[63] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 2

[64] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. 2

[65] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9150–9161, 2023. 1, 2, 5

[66] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[67] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *arXiv preprint arXiv:2305.04461*, 2023. 2, 3

[68] 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 (ICCV)*, pages 5826–5835, 2021. 2, 3

[69] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12588–12597, 2023. 2