A Diffusion-ReFinement Model for Sketch-to-Point Modeling

Di Kong, Qiang Wang, and Yonggang Qi(✉)

Beijing University of Posts and Telecommunications, Beijing, China
{dikong,wanqqiang,qiyg}@bupt.edu.cn

Abstract. Diffusion probabilistic model has been proven effective in generative tasks. However, its variants have not yet delivered on its effectiveness in practice of cross-dimensional multimodal generation task. Generating 3D models from single free-hand sketches is a typically tricky cross-domain problem that grows even more important and urgent due to the widespread emergence of VR/AR technologies and usage of portable touch screens. In this paper, we introduce a novel Sketch-to-Point Diffusion-ReFinement model to tackle this problem. By injecting a new conditional reconstruction network and a refinement network, we overcome the barrier of multimodal generation between the two dimensions. By explicitly conditioning the generation process on a given sketch image, our method can generate plausible point clouds restoring the sharp details and topology of 3D shapes, also matching the input sketches. Extensive experiments on various datasets show that our model achieves highly competitive performance in sketch-to-point generation task. The code is available at *https://github.com/Walterkd/diffusion-refine-sketch2point*.

1 Introduction

The challenge of being able to obtain a 3D model from a single sketch has been studied for decades. Our goal is to provide precise and intuitive 3D modeling for users with limited drawing experience. And we try to propose a method that is tailored for point cloud reconstruction from a single free-hand sketch, complementing to existing single-view reconstruction methods. Taking into account the abstraction and distortion common in sketches by novice users, a sketch contains far less information than an image due to its simplicity and imprecision. This leads to a lack of important information such as color, shading and texture. To tackle this cross-domain problem, a number of learning-based methods [49, 27] have been proposed that are trained by comparing rendered silhouettes or depth maps with ground truth ones, with no involvement of the ground truth in 3D shape. After obtaining depth and normal maps from the sketches, they use the maps to reconstruct the 3D shapes. The absence of the ground truth 3D shapes highly limits the capabilities of many sketch based 3D model generation networks to restore the topology and fine details in 3D shapes.

We find that denoising diffusion probabilistic models (DDPM) [16] and their variants [23, 28, 45] have achieved great success in the generation tasks. Especially in the 3D point cloud completion and generation task [29, 28]. Combined

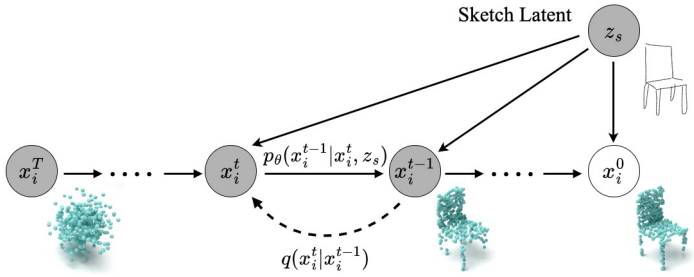

Fig. 1. The directed graphical model of the reconstruction process for point clouds.

with non-equilibrium thermodynamics, the reverse diffusion process can well simulate the generation process of 3D point cloud from disorder to order. While remarkable progress has been made in the point cloud generation tasks via GANs [1, 41, 37], the GANs based methods have some inherent limitations for modeling point clouds. Compared with GANs based generative work, the training procedure of diffusion probabilistic model is more stable and probabilistic generative model can achieve a better generation quality. However, the DDPM has rarely been studied in multimodal generation tasks. The conditional DDPM [4, 23] is also only applied in the homo-dimensional multimodal generation task and does not demonstrate its effectiveness in cross-dimensional problem. Sketch to 3D point cloud generation task can be treated as a conditional generation problem in the framework of DDPM [50, 28]. Indeed, we find the point clouds generated by a conditional DDPM often have a good overall distribution that uniformly covers the shape of the object. Nonetheless, due to the probabilistic nature of DDPM and the lack of a suitable network architecture to train the conditional DDPM for sketch to 3D point cloud generation in the previous works, we find DDPM reconstructed point clouds often lack flat surfaces and sharp details, which is also reflected by their high evaluation metric loss compared with state-of-the-art sketch to point cloud reconstruction methods in our experiments.

In this paper, we extend conditional DDPM to generate point cloud from a single free-hand sketch, adapting the score function in the reverse diffusion process to obtain the desired point distributions. Analogous to particles thermally diffusing in the 3D space, we use the diffusion process to simulate the transformation from the clean point distribution to a noise distribution. Likewise, we consider the reverse diffusion process to model the variation of point distribution in point cloud reconstruction, through which we recover the target point cloud from the noise. Our diffusion and sampling paradigm is shown in Fig. 2. Firstly, we use the Conditional Sketch-to-Point Reconstruction Network (CS2PRNet) to generate a point cloud by the DDPM conditioned on the input sketch image. It can iteratively move a set of Gaussian noise towards a desired and clean point cloud, corresponding to the input sketch. Following, the ReFinement Network (RFNet), a shape discriminator, further refines the desired point cloud generated from the CS2PRNet with the help of its capability to distinguish

the ground truth ones and the reconstructed ones. In addition, RFNet can be used to refine the relatively low quality point clouds generated by an accelerated DDPM [22], so that we could enjoy an acceleration up to 20 times, while minimizing the performance drop. In this way, the reconstruction results generated by our model demonstrate good overall density distribution, sharp local details and good match to input sketches.

To summarize, our contributions are as follows:

- For the first time, we extend conditional DDPM to be a good model with an effective and efficient loss function to generate point cloud from a single free-hand sketch, which provides a high-quality and easy-to-use 3D content creation solution.
- We address the importance of sketch latent representation in the sketch-based reconstruction task, and design a CS2PRNet to condition the generation process explicitly on sketch shape latent. And by using RFNet to refine the reconstructed point clouds, we can generate point clouds with both sharp 3D characteristic details and good matching to input sketches.
- Extensive experiments on various datasets demonstrate that our model achieves competitive performance in sketch based 3D modeling task.

2 Related Work

2.1 Single-View 3D Reconstruction

Restoration of 3D geometry from a single image is an ill-posed problem. Early approaches utilize shadings [17] and textures [46] to obtain clues about surface orientations. With the emergence of large-scale 3D model datasets [2], data-driven approaches are developed to infer category-specific shapes directly from image features [3, 5, 9, 10, 19, 30, 33, 44], in the formats of voxels [5, 10], point clouds [9], mesh patches [19, 33, 44, 8], and implicit representation [3, 30]. Recently, neural rendering techniques [21, 24, 31, 40, 48] are proposed to alleviate the necessity for ground truth 3D models in training, which is achieved by comparing the estimated shape silhouette with an input image, thus enabling supervision with 2D images only. Our method is based on the conditional generative encoder-decoder network architecture of [9]. We extend the original approach by disentangling image features into a latent shape space and utilizing the diffusion probabilistic model [38, 16] for reconstruction process.

2.2 Sketch-Based Modeling

Modeling based on sketches is a long established problem that has been investigated before deep learning methods become widespread. The earlier method [7], inspired by lofting technique, modeled shapes from a single image and user input strokes. Recent works [6, 13, 14, 18, 27, 36, 39, 42, 43, 49] utilizing deep learning methods to guide the 3D modeling from user inputs. Only a handful of them, however, focused on reconstructions from free-hand sketches [42, 43, 49]. Wang

et al. [43] presented a retrieval-based method to reconstruct 3D shapes. Wang *et al.* [42] adopted [9] for sketch-based 3D reconstruction by proposing an additional image translation network that aims at sketching style standardization to account for the variability of sketching styles. Sketch2model [49] solved the sketch ambiguities by introducing a view code. Different from their work, we regard the sketch latent representation, extracted from free-hand sketches, as a condition, and rely on it to guide the reverse diffusion process.

2.3 Diffusion Models

The diffusion models considered in our work can be interpreted as diffusion-based generative models, consisting of diffusion probabilistic models [38] and denoising diffusion probabilistic models [16]. Diffusion models are learned with two fixed Markov chains, controlling diffusion and reverse diffusion process. They can produce better samples than those of GANs. To tackle the generative learning trilemma, some denoising diffusion GANs have been proposed [47, 25], which reduce the sampling cost in diffusion models and perform better mode coverage and sample diversity compared to traditional GANs. In these works, a diffusion model whose reverse process is parameterized by condition GANs. Specifically to reduce the generation cost of diffusion-based generative models, Wang et al. [45] presented their latest Diffusion-GAN. The main distinction from denoising diffusion GANs is that Diffusion-GAN does not require a reverse diffusion chain during training and generation. Unlike their work, we introduced a shape discriminator in the final step of reverse diffusion process, which we used to help us better control the final 3D shape quality. Additionally, to address the slow sampling rate of the diffusion model, we apply [22], a more faster sampling strategy, through a defined bijection to construct the approximate diffusion process with less steps S. The length of the approximate reverse process S is relatively small.

Besides solving the problem of high sampling costs, denoising diffusion GANs can also be used in the study of multimodal generation task, such as text-to-speech (TTS) synthesis [25], multi-domain image translation [47]. They are capable to be applied in cross-domain tasks is because they model each denoising step using a multimodal conditional GAN. In parallel to the above GAN involved approaches, conditional DDPM [4, 23] has also been demonstrated to work for cross-modal generation tasks. However, they can currently only perform generative tasks in the same dimension, while our method can perform 2D-to-3D generation process by matching each shape latent variable with give 2D reference sketch image.

3 Methods

Given a single hand-drawn sketch in the form of line drawings, our method aims to reconstruct a 3D point cloud. We utilize the diffusion probabilistic model for the generation of 3D point cloud from a single free-hand sketch. Then we extend the diffusion model architecture by decomposing sketch image features

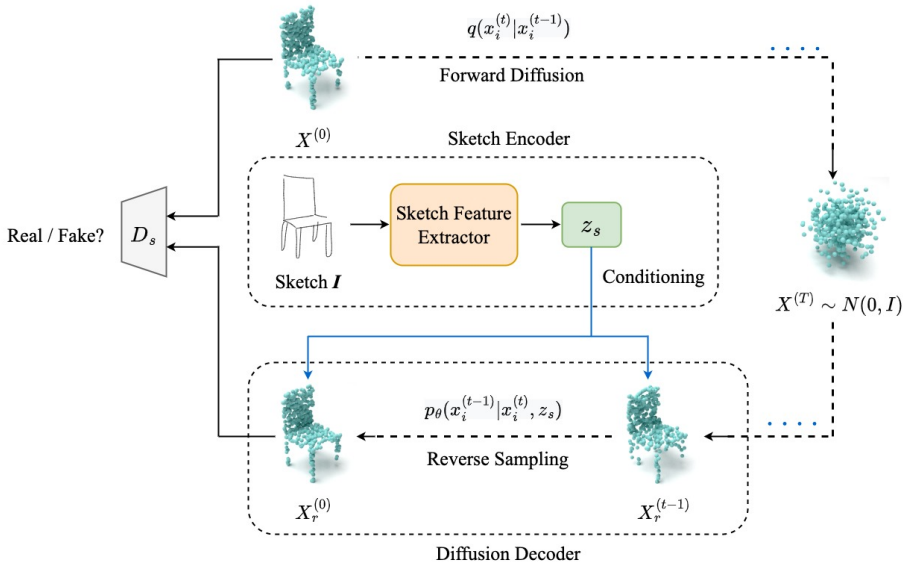

Fig. 2. The overview of our proposed model' s framework. It illustrates the training process and the reconstruction process. Our method utilizes the encoder-decoder architecture. The output of the decoder is refined by a shape discriminator D_s, obtaining a high quality point cloud by one step from a coarse point cloud.

into a latent shape space, and condition the generation process on the sketch shape latent. To better accommodate the cross-modal generation task, a new conditional reconstruction network is also provided. And a refinement network is applied to preserve the reconstructed shape quality.

3.1 Background on Diffusion Probabilistic Models

We assume $q_{data} \sim q(x_i^0)$ to be the distribution of the groundtruth point cloud x_i in the dataset. And $q_{ultimate} = \mathcal{N}(\mathbf{0}_{3N}, \boldsymbol{I}_{3N\times 3N})$ to be the ultimate latent distribution, where $\mathcal{N}$ is the standard Gaussian normal distribution and N is the amount of points per point cloud. Then, the conditional diffusion probabilistic model of T steps consists of a diffusion process and a reverse sampling process.

The Diffusion Process The diffusion process is implemented by a forward Markov chain. We use the superscript to denote the diffusion step t. For conciseness, we omit the subscription i in the following discussion. From clean data x^0, the diffusion process is to add Gaussian noise to get $x^{1:T}$:

$$q(x^{1:T}|x^0) = \prod_{t=1}^{T} q(x^t|x^{t-1}), \text{ where } q(x^t|x^{t-1}) = \mathcal{N}(x^t; \sqrt{1-\beta_t}x^{t-1}, \beta_t I) \ . \quad (1)$$

We define the Markov diffusion kernel as $q(x^t|x^{t-1})$. The role of the kernel is to add small Gaussian noise to disrupt the distribution of x^{t-1}. The whole process slowly injects Gaussian noise into the clean data distribution q_{data} until the output distribution is deformed to $q_{ultimate}$ according to a predefined variance schedule hyper-parameters $\beta_t, t = 1, ..., T$, which control the step sizes of the diffusion process.

The Reverse Sampling Process The points are sampled out of a noise distribution $p(x^T)$ which is an approximation to $q(x^T)$. Let $p(x^T) \sim p_{start}$ be the input noise variable. The reverse process, conditioned on sketch shape latent z_s, converts x^T to x_r^0 through a backward Markov chain:

$$p_\theta(x_r^{0:T-1}|x^T, z_s) = \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t, z_s) \ . \tag{2}$$

$$p_\theta(x^{t-1}|x^t, z_s) = \mathcal{N}(x^{t-1}; \mu_\theta(x^t, z_s, t), \sigma_t^2 \boldsymbol{I}) \ . \tag{3}$$

The mean $\mu_\theta(x^t, z_s, t)$ is a neural network parameterized by θ and the variance σ_t^2 is a time-step dependent constant closely connected to β_t. To generate a sample conditioned on z_s, we first sample from the starting distribution $p(x^T) \sim \mathcal{N}(\mathbf{0}_{3N}, \boldsymbol{I}_{3N\times 3N})$, then draw x^{t-1} via $p_\theta(x^{t-1}|x^t, z_s)$, where t decreases from T to 1. And x_r^0 is the sampled target shape.

Training To make likelihood $p_\theta(x^0)$ tractable to calculate, we use the variational inference to optimize the negative log-likelihood $-logp_\theta(x^0)$. [16] introduced a certain parameterization for μ_θ that can largely simplify the training objective, known as variational lower bound (ELBO). We use the notation $\alpha_t = 1 - \beta_t$, and $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The parameterization is $\sigma_t^2 = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, and $\mu_\theta(x^t, z_s, t) = \frac{1}{\sqrt{\alpha_t}}(x^t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x^t, z_s, t))$, where ϵ_θ is a neural network taking noisy point cloud $x^t \sim q(x^t|x^0) = \mathcal{N}(x^t; \sqrt{\overline{\alpha}_t}x^0, (1-\overline{\alpha}_t)I)$, diffusion step t, and conditioner z_s as inputs. The neural network ϵ_θ learns to predict the noise ϵ added to the clean point cloud x^0, which can be used to denoise the noisy point cloud $x^t = \sqrt{\overline{\alpha}_t}x^0 + \sqrt{1-\overline{\alpha}_t}\epsilon$. Then we minimizing the training objective L by adopting the variational bound:

$$L(\theta) = \mathbb{E}_q[\sum_{t=2}^{T}\sum_{i=1}^{N} D_{KL}(q(x_i^{t-1}|x_i^t, x_i^0)||p_\theta(x_i^{t-1}|x_i^t, z_s)) - \sum_{i=1}^{N} logp_\theta(x_i^0|x_i^1, z_s)] \ . \tag{4}$$

To be computable, we expand the ELBO into a sum of KL divergences, each of which compares two Gaussian distributions and therefore they can be computed in closed form. The detailed derivations, including the definition of $q(x_i^{t-1}|x_i^t, x_i^0)$, are provided in the supplementary material.

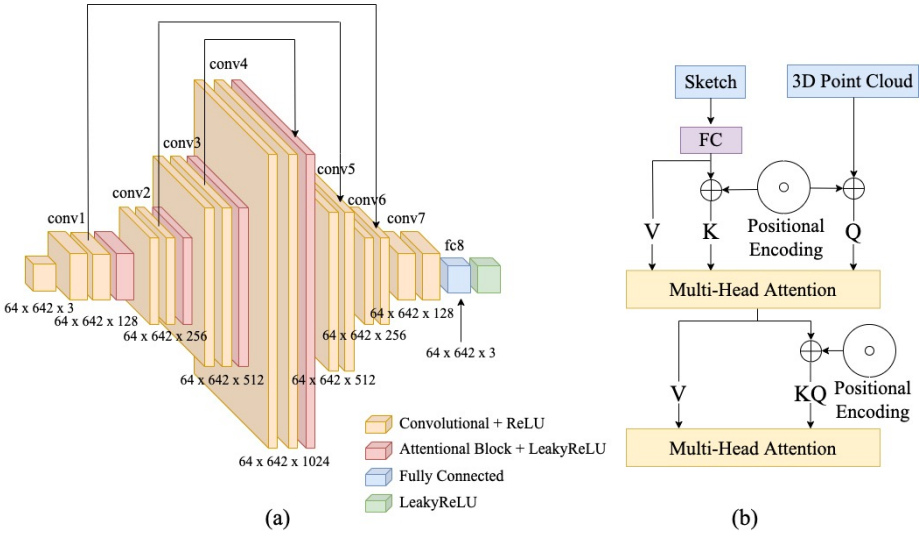

Fig. 3. The illustration of the Conditional Sketch-to-Point Reconstruction Network. Our reconstruction network uses long range convolution skip connections. We use two main types of blocks, convolutional block and attentional block.

3.2 Conditional Sketch-to-Point Reconstruction Network

In this section, we present the architecture of Conditional Sketch-to-Point Reconstruction Network ϵ_θ (see Fig. 3 for an illustration). With inputting the noisy point cloud x^t, the sketch shape latent z_s, the diffusion step t and the variance schedule β_t, the output of network ϵ_θ is per-point difference between x^t and x^{t-1}. In addition, ϵ_θ should also effectively incorporate information from z_s. The goal is to infer not only the overall shape but also the fine-grained details based on z_s. We design a neural network that achieves these features.

Before the introduction of the network, we need to resolve the role of the condition. Since the work in [28] has been shown to model the complex conditional distribution in 3D shape domain, we adopt them to approximate the true reverse sampling distribution $p_\theta(x^{t-1}|x^t, z_s)$, formulated similar to [16]. Unlike the [28], to achieve the goal of multimodal reconstruction, we use the latent shape representation extracted from sketch image. The main difference is that, in [28], x_0 is predicted as a deterministic mapping of x_t conditioned on a 3D object shape latent encoded in x_0 itself, while in our case x_0 is produced by the generator with latent variable z_s extracted from a sketch image corresponding to its 3D object. This is the key difference that allows our reverse sampling distribution $p_\theta(x^{t-1}|x^t, z_s)$ to become multimodal and complex in contrast to unimodal denoising diffusion probabilistic model in [28].

In [28], its MLP based generative network predicts the mean value of the next coordinate distribution of a point based on the latent shape representation z, with input of the coordinate x_i^{t+1} from the previous step. However, since

the information extracted from a sketch image is not as much as that encoded in a 3D object, we think that the MLP is not suitable in our task, because it may lose some information about the accurate positions of the points. Thus we need a generative network that can better effectively form an associative mapping between the information extracted from the 2D sketch image and the coordinates of the 3D object. We designed a conditional generative network to accomplish our goal satisfactorily. The overall architecture is shown in Fig. 3. The reverse diffusion kernel is parameterized by $\epsilon_\theta(x_i^t, t, z_s)$. We put the detail of reverse diffusion kernel to the supplementary material. We use a fixed linear variance schedule β_1, ..., β_T to represent the timetable of the reconstruction process. Time embedding vector is comprised of $[\beta_t, sin(\beta_t), cos(\beta_t)]$ and is used to ensure conditioning on t. We implement the reconstruction network using a variant of PVCNN [26], which consists of a series of ConcatSquashConv1d layers [12]. The dimension of the ConcatSquash-CNN used in our model is 3-128-256-512-1024-512-256-128-3, and we use the LeakyReLU nonlinearity between the layers. The input to the first layer is the 3D position of points x_i^t. And we use the quaternion $c = [\beta_t, sin(\beta_t), cos(\beta_t), z_s]$ as the context embedding vector. Then the quaternion c is inserted to every level of the reconstruction network. The features from z_s are transformed and then aggregated to the point x_i^t through attention mechanism. The attentional block are applied four times and features are eventually propagated to the original input point cloud. More details about convolutional block and attentional block are provided in supplementary material.

3.3 ReFinement Network: A Shape Discriminator

After training the diffusion process, the model is able to reconstruct the point cloud from a single free-hand sketch. However, due to the limited information extracted from sketch images, point clouds reconstructed from our model trained only by the diffusion process will show local distortions and deformations. Therefore, we introduce a shape discriminator D_s to alleviate such distortions. By introducing an adversarial loss, the shape discriminator is trained in an adversarial manner together with the encoder and decoder. It functions as a trade-off between denoising and shape quality. Under the influence of the discriminator, our generative network may not produce a shape in an exact match to the input sketch at a certain angle, but is more capable of taking into account some of the characteristics that a 3D object has, such as topological structure, to the generation results.

Given x^T, our conditional reconstruction network first generates x_r^0. The discriminator is trained to distinguish between the real x^0 and fake x_r^0. Then we train the conditional 3D point cloud generator $p_\theta(x^{t-1}|x^t, z_s)$ to approximate the true reverse sampling distribution $q(x^{t-1}|x^t)$ with an adversarial loss that minimizes a divergence D_{adv} in the last reverse sampling step:

$$\min_\theta \mathbb{E}_q[\sum_{i=1}^{N} D_{adv}(q(x_i^0|x_i^1)||p_\theta(x_i^0|x_i^1, z_s))] \ , \tag{5}$$

where fake samples from $p_\theta(x_r^0|x^1, z_s)$ are contrasted against real samples from $q(x^0|x^1)$. We denote the discriminator as $D_\phi(x_r^0, x^1)$, with parameters ϕ. It takes the 3-dimensional x_r^0 and x^1 as inputs, and decides whether x_r^0 is a plausible reverse sampled version of x^1. Given the discriminator, we train the generator by $max_\theta E_q E_{p_\theta}[log(D_\phi(x_r^0, x^1))]$, which updates the generator with the non-saturating GAN objective [11]. To summarize, the discriminator is designed to be diffusion-step-dependent and 3D topology-aware to aid the generator to achieve high-quality sketch-to-point reconstruction.

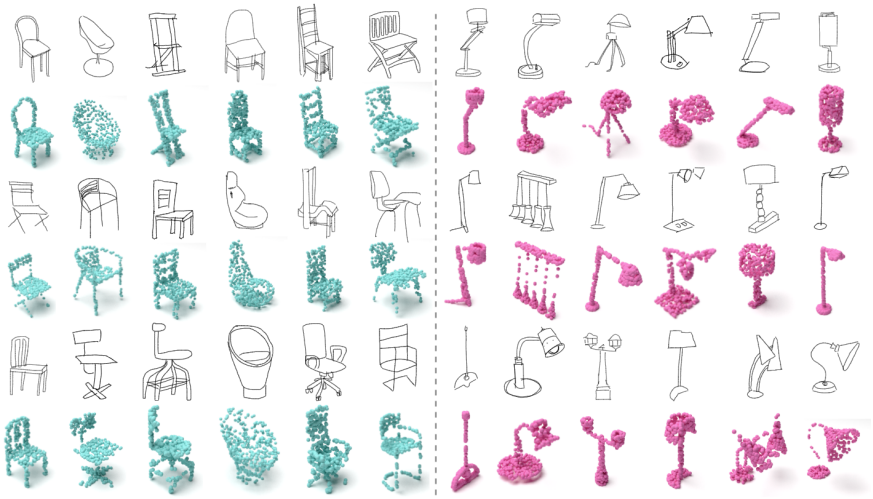

Fig. 4. Some representative examples of point clouds reconstructed by our model.

4 Experiments

In this section, we evaluate our proposed model's performance on sketch-to-point generation problem. We first perform case studies to show the effectiveness of our multimodal Diffusion-ReFinement model. Quantitative and qualitative evaluations on both synthetic and free hand-drawn sketches are presented in Section 4.3. We also provide some additional insight into our model by conducting an ablation study in Section 4.4 and proposing Feature Map module in Section 4.5.

4.1 Experimental Setup

Datasets For sketch to 3D point cloud generation experiments, we employ the 3D shapes from ShapeNetCorev2 [2] to match our corresponding sketch datasets. Every point cloud has 642 points. And every category in the dataset is randomly split into training, testing and validation sets by the ratio 80%, 15%,

5% respectively. For the corresponding sketch datasets, at the pretrain stage, we use ShapeNet-Synthetic sketches to train our model. We use the same approach as [49] to create the ShapeNet-Synthetic dataset using rendered images provided by [20]. We use this dataset to pretrain the model, under the same train/test/validation split as said before. At the fine-tuning stage, to quantitatively evaluate our method on free-hand sketches and benefit further research, we use ShapeNet-Amateur sketches [34] to fine-tune our pretrained model. It contains a chair dataset with 902 sketch-3D shape quadruplets and a lamp dataset with 496 sketch-3D shape quadruplets.

Evaluation Metrics Following prior works [1], we use two evaluation metrics to compare the quality of the reconstructed 3D point clouds to reference shapes: the Chamfer Distance (CD) and the Earth Mover's Distance (EMD) [35]. Chamfer distance measures the squared distance between each point in one point set to its nearest neighbor in the other set, while the EMD is the solution of the optimization problem that aims at transforming one set to the other.

4.2 Implementation Details

As the general training objective and algorithms in the previous section lay the foundation, we implement a model to reconstruct a point cloud from a single free-hand sketch based on the probabilistic model. We use a pretrained ResNet-18 [15] as our sketch image feature extractor, also can be called the sketch shape latent encoder and leverage the reverse sampling process presented in Section 3.1 for decoding. Expanding on Equation (4) and (5), we train the model by minimizing the following adapted objective:

$$L(\theta) = \mathbb{E}_q[\sum_{t=2}^{T}\sum_{i=1}^{N} D_{KL}(q(x_i^{t-1}|x_i^t, x_i^0)||p_\theta(x_i^{t-1}|x_i^t, z_s)) - \sum_{i=1}^{N} logp_\theta(x_i^0|x_i^1, z_s) + \sum_{i=1}^{N} D_{adv}(q(x_i^0|x_i^1)||p_\theta(x_i^0|x_i^1, z_s))] \ . \tag{6}$$

To decode a point cloud conditioned on the latent code z_s, we sample some points x_i^T from the noise distribution $p(x_i^T)$ and pass the points through the reverse Markov chain $p_\theta(x_i^{0:T}|z_s)$ defined in Equation (2) to acquire the reconstructed point cloud $X^0 = \{x_i^0\}_{i=1}^N$. For diffusion model, we set the variance schedules to be $\beta_1 = 0.0001$ and $\beta_T = 0.05$, and β_t' s $(1 < t < T)$ are linearly interpolated. For diffusion optimization, we use Adam optimizer with learning rate starting from e^{-3} and ended at e^{-4}. For discriminator, we also use Adam optimizer with learning rate e^{-4}. And we train a separate model for each category respectively.

4.3 Comparisons and Evaluations

We quantitatively and qualitatively compare our method with the following state-of-the-art single sketch image reconstruction models: TMNet [32], Pixel2Mesh

Table 1. Comparison of single sketch image to point cloud reconstruction performance. CD is multiplied by 10^3 and EMD is multiplied by 10^2.

Category	CD						EMD					
	TMNet	Pixel2Mesh	PSGN	3D-R2D2	OccNet	Ours	TMNet	Pixel2Mesh	PSGN	3D-R2D2	OccNet	Ours
AmaChair	3.716	5.084	**2.977**	4.145	3.450	3.250	12.59	14.34	10.23	13.69	**10.08**	10.19
AmaLamp	6.856	9.339	6.453	7.947	6.293	**6.152**	17.75	18.81	17.38	17.60	16.73	**16.33**
AmaMean	5.286	7.212	4.715	6.046	4.872	**4.701**	15.17	16.58	13.81	15.65	13.41	**13.26**
SynAirplane	1.788	2.656	1.577	2.147	1.834	**1.448**	7.03	8.25	7.55	6.74	6.61	**6.39**
SynBench	2.871	3.477	2.755	3.153	**2.425**	2.623	9.27	14.01	10.22	9.81	8.97	**8.84**
SynCabinet	5.106	5.859	4.936	5.533	4.824	**4.760**	11.03	12.50	12.09	11.95	11.19	**10.64**
SynCar	2.840	3.312	**2.116**	3.129	2.417	2.291	7.31	8.29	7.76	7.66	**7.08**	7.26
SynChair	3.190	4.340	**2.692**	3.079	2.913	2.865	9.92	12.57	10.15	9.75	9.87	**9.67**
SynMonitor	3.957	4.481	3.833	4.059	3.974	**3.713**	9.84	11.96	10.48	10.03	9.84	**9.77**
SynLamp	6.023	7.706	5.865	6.975	5.778	**5.564**	16.88	17.63	16.14	17.04	16.25	**15.92**
SynSpeaker	5.725	6.479	5.654	5.942	5.514	**5.323**	12.71	13.87	13.12	12.82	12.35	**12.29**
SynRifle	1.425	1.874	1.392	1.454	**1.238**	1.374	6.99	7.58	7.12	7.07	**6.81**	6.90
SynSofa	4.357	4.865	4.229	4.257	4.231	**4.152**	11.91	13.57	12.42	12.19	12.05	**11.62**
SynTable	4.581	5.827	4.428	5.077	**4.024**	4.164	**11.43**	12.90	12.54	12.05	11.89	11.70
SynTelephone	2.589	3.183	2.241	2.745	2.451	**2.172**	7.66	8.96	8.35	7.73	7.62	**7.51**
SynVessel	2.259	3.258	**2.041**	2.423	2.174	2.146	7.62	9.21	8.90	8.17	7.61	**7.23**
SynMean	3.593	4.409	3.366	3.844	3.369	**3.304**	9.97	11.64	10.53	10.23	9.86	**9.67**

Table 2. Model's performance on ShapeNet-Synthetic dataset and ablation study for Discriminator's effectiveness.

Training Strategy III	Airplane		Bench		Cabinet		Car		Chair		Monitor		Lamp	
	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD
Trained on Synthetic sketches	1.889	6.78	2.955	9.27	4.936	11.50	2.851	8.16	3.121	9.91	4.112	10.96	6.019	16.18
After Discriminator	**1.448**	**6.39**	**2.623**	**8.84**	**4.760**	**10.64**	**2.291**	**7.26**	**2.865**	**9.67**	**3.713**	**9.77**	**5.564**	**15.92**
Training Strategy III	Speaker		Rifle		Sofa		Table		Telephone		Vessel		Mean	
	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD
Trained on Synthetic sketches	5.542	12.78	1.893	7.38	4.722	12.07	4.477	11.90	2.524	7.86	2.532	8.03	3.659	10.21
After Discriminator	**5.323**	**12.29**	**1.374**	**6.90**	**4.152**	**11.62**	**4.164**	**11.70**	**2.172**	**7.51**	**2.146**	**7.23**	**3.304**	**9.67**

[44], PSGN [9], 3D-R2N2 [5], OccNet [30], using point clouds from thirteen categories in ShapeNet. The comparison is performed on ShapeNet-Synthetic and ShapeNet-Amateur datasets. For volume-based methods TMNet and Pixel2Mesh, both metrics are computed between the ground truth point cloud and 642 points uniformly sampled from the generated mesh. Since the outputs of Pixel2Mesh are non-canonical, we align their predictions to the canonical ground truth by using the pose metadata available in the dataset. Results are shown in Table 1. On chair of both datasets, our method outperforms other methods except PSGN when measured by CD. The EMD score of our method pushes closer towards the OccNet performance when tested on Amateur dataset and reaches the best performance when tested on Synthetic dataset. While on category of lamp, our approach outperforms other methods in both evaluation metrics for two datasets. Including these two categories, our approach outperforms other five baselines on most of the categories. Notably, from Tables 1 and 2, when both training and testing are conducted on the ShapeNet-Synthetic dataset, our model can have a better performance in both CD and EMD. Also, the visualization of reconstructed point clouds in Fig. 5 validates the effectiveness of our model compared with other baselines. While OccNet can reconstruct the rough shapes, it fails to capture the fine details of the geometry and is not able to model the topology of

Table 3. Comparison of 2D silhouette IoU score on ShapeNet-Synthetic and ShapeNet-Amateur datasets. Generated shapes are projected to the ground truth view and we calculate the IoU score between the projected silhouettes and the ground truth ones.

Category	Airplane	Bench	Cabinet	Car	Chair	Monitor	Lamp	Speaker	Rifle	Sofa	Table	Telephone	Vessel	Mean
TMNet	0.593	0.625	0.810	0.821	0.709(0.683)	0.784	0.606(0.585)	0.790	**0.672**	0.796	0.703	0.813	0.706	0.713
Pixel2Mesh	0.532	0.564	0.734	0.772	0.675(0.652)	0.729	0.548(0.530)	0.713	0.596	0.738	0.625	0.750	0.635	0.653
PSGN	0.652	0.633	0.832	0.866	**0.744**(0.710)	0.803	0.619(0.599)	0.808	0.633	0.813	0.712	0.848	0.724	0.733
3D-R2D2	0.565	0.573	0.786	0.796	0.718(0.707)	0.765	0.579(0.554)	0.766	0.618	0.805	0.678	0.797	0.688	0.693
OccNet	0.641	**0.684**	**0.873**	0.839	0.736(0.712)	0.788	0.643(0.622)	0.812	0.655	0.802	**0.745**	0.823	0.739	0.741
Ours	**0.679**	0.667	0.858	**0.868**	0.732(**0.715**)	**0.812**	**0.662**(**0.627**)	**0.827**	0.647	**0.820**	0.731	**0.854**	**0.756**	**0.750**

surface. PSGN performs generally better than OccNet in terms of the capability of modeling the fine structures. However, due to the limitations of the vanilla architecture, it struggles to reconstruct shapes with complex topology. In comparison, we believe that in the vast majority of cases, our approach has surpassed other approaches in terms of visual quality and is better at restoring detailed information of the corresponding sketches. We are able to generate point clouds with complex topology while maintaining high reconstruction accuracy.

We also compare projected silhouettes of generated 3D models with ground truth silhouettes, and show the results in Table 3. It shows our model is more powerful at matching input sketches.

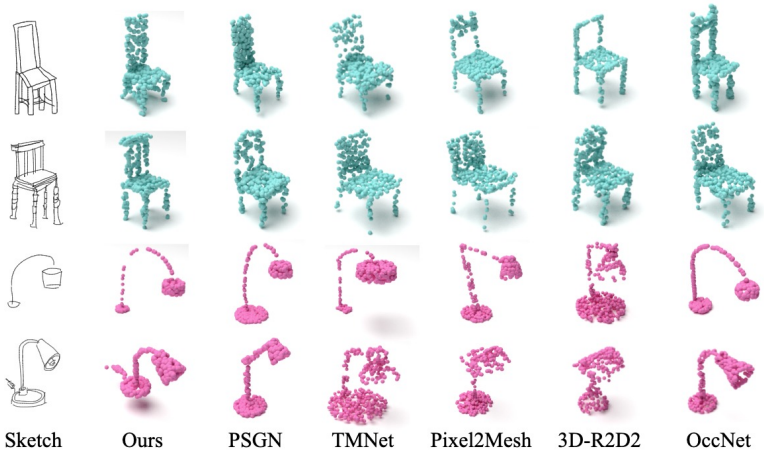

Fig. 5. Qualitative comparisons with other five baseline methods on free-hand sketches from the ShapeNet-Amateur Sketch dataset.

4.4 Ablation Study

In this section, we conduct controlled experiments to validate the importance of different components. Table 4 presents the comparison of the generator network

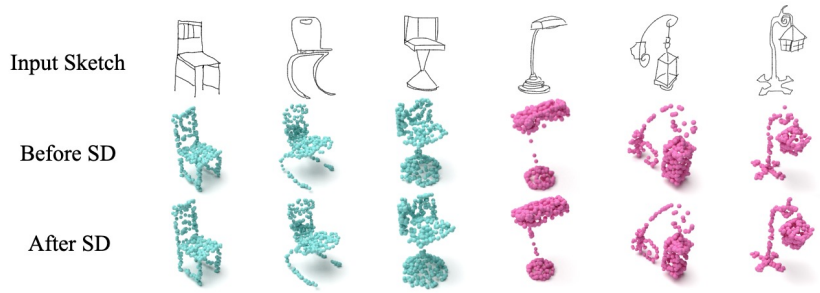

Fig. 6. Ablation study for shape discriminator. Without shape discriminator, the output shape may not resemble real objects(middle row).

using CNN and MLP respectively. Our approach has significantly outperformed the MLP based method across all datasets. In Table 5 and Table 2, we investigate the impact of the discriminator and validate the effectiveness of discriminator design. As is illustrated in Section 3.2 and Section 3.3, our generation result relies on both input sketch image and a discriminator refinement. With the adjustment of the discriminator, the network is expected to generate finer details to match the input sketch. Tables 5 and 2 show that with the discriminator being put into use, both metrics improved in all categories. Fig. 6 shows the importance of the proposed shape discriminator to the generation results. Without shape discriminator, the output point cloud may get distorted and not resemble a real object. While with the inclusion of a discriminator, we can get the desired point cloud with promising shape quality and better restore the topology of a 3D object. In Table 5, we also demonstrate the effectiveness of our Fine-Tune training strategy. Since the size of ShapeNet-Amateur dataset is relatively small and the diffusion model normally requires a large dataset, our model cannot perform promisingly if trained directly and only on the ShapeNet-Amateur dataset, so we decided to train on the ShapeNet-Synthetic dataset with a larger amount of data, and then fine-tune the model using the ShapeNet-Amateur dataset after it has basically converged. From the final results, subsequent to such a training strategy, applying the discriminator last when the fine-tuning has almost converged, our model is able to have a high performance on the ShapeNet-Amateur dataset.

Table 4. Comparison of the Conditional Reconstruction Network using CNN and MLP respectively.

Tested Dataset	Metric	MLP Based	CNN Based
AmaChair	CD	4.739	**3.25**
	EMD	10.84	**10.19**
SynChair	CD	3.983	**2.865**
	EMD	10.55	**9.67**
AmaLamp	CD	7.467	**6.152**
	EMD	17.42	**16.33**
SynLamp	CD	6.571	**5.564**
	EMD	16.75	**15.92**

Table 5. Model's performance on different training stages and ablation studies for effectiveness of Discriminator and Fine-Tune training strategy.

Training Strategy I	AmaChair		AmaLamp	
	CD	EMD	CD	EMD
Trained on Synthetic sketches	9.620	18.33	18.790	30.83
Fine tune on Amateur sketches	3.766	10.35	6.473	17.18
After Discriminator	**3.250**	**10.19**	**6.152**	**16.33**
Training Strategy II	**AmaChair**		**AmaLamp**	
	CD	EMD	CD	EMD
Trained on Amateur sketches	6.376	12.78	8.078	19.45
After Discriminator	5.575	12.08	7.867	16.65

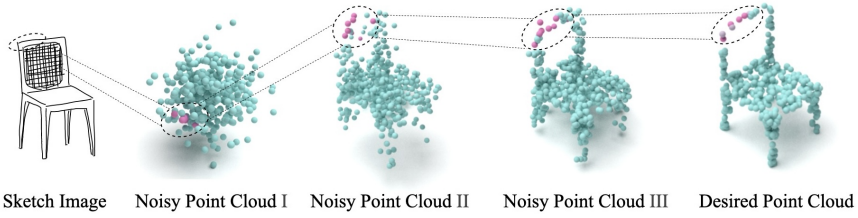

Fig. 7. The Feature Map module maps features from the sketch image to the noisy point cloud and demonstrates how features guide the movement of points from the noisy point cloud to the desired point cloud.

4.5 Sketch-Point Feature Map (FM)

Further, to study how the sketch image influence the point cloud reconstruction process, we implement a Feature Map (FM) module. The FM module transmits information from the sketch image to the denoising process. The FM module maps the features from sketch image to noisy points in the reconstruction network, which are $\{x_i|1 \leq i \leq N\}$. Fig. 7 illustrates this process. In this way, the reconstruction network can utilize local features of the sketch image to manipulate the input noisy point cloud to form a clean and desired point cloud. The key step in this process is to map features from sketch image to $\{x_i|1 \leq i \leq N\}$. We adopt a similar strategy from [29], except that we replace incomplete point clouds with sketches. Features of sketch image are transformed through a shared MLP, and then aggregated to the points x_i through the attention mechanism, which is a weighted sum of the features from part of the sketch. We set a large distance to define sketch parts in FM module. This makes FM module have large receptive fields, so that we can query a large part of the sketch image. And we leverage the spatial correspondence between the different parts of the sketch image and the point cloud through the proposed Feature Map module to infer high level 2D-to-3D structural relations.

5 Conclusions

In this paper, we propose the Sketch-to-Point Diffusion-ReFinement model for cross-domain point cloud reconstruction. A novel conditional reconstruction network is presented, to condition the generation process explicitly on sketch shape latent, which emphasizing the importance of sketch latent and brings controllability to the output point cloud. From observation and evaluation, a refinement network provides users to restore the point clouds with sharp 3D characteristic details and topology. Experimental results demonstrate that the proposed model achieves the state-of-the-art performance in sketch to point cloud generation task. We hope our method can inspire further researches in cross-domain 3D reconstruction and sketch-related areas.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International conference on machine learning. pp. 40–49. PMLR (2018)
2. Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. CoRR **abs/1512.03012** (2015), http://arxiv.org/abs/1512.03012
3. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
4. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
5. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
6. Delanoy, J., Aubry, M., Isola, P., Efros, A.A., Bousseau, A.: 3d sketching using multi-view deep volumetric prediction. Proceedings of the ACM on Computer Graphics and Interactive Techniques **1**(1), 1–22 (2018)
7. Deng, C., Huang, J., Yang, Y.L.: Interactive modeling of lofted shapes from a single image. Computational Visual Media **6**(3), 279–289 (2020)
8. Deng, Z., Liu, Y., Pan, H., Jabi, W., Zhang, J., Deng, B.: Sketch2pq: freeform planar quadrilateral mesh design via a single sketch. IEEE Transactions on Visualization and Computer Graphics (2022)
9. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
10. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: European Conference on Computer Vision. pp. 484–499. Springer (2016)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
12. Grathwohl, W., Chen, R.T., Bettencourt, J., Sutskever, I., Duvenaud, D.: Ffjord: Free-form continuous dynamics for scalable reversible generative models. arXiv preprint arXiv:1810.01367 (2018)
13. Han, X., Gao, C., Yu, Y.: Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling. ACM Transactions on graphics (TOG) **36**(4), 1–12 (2017)
14. Han, Z., Ma, B., Liu, Y.S., Zwicker, M.: Reconstructing 3d shapes from multiple sketches using direct shape optimization. IEEE Transactions on Image Processing **29**, 8721–8734 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
17. Horn, B.K.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view (1970)

18. Huang, H., Kalogerakis, E., Yumer, E., Mech, R.: Shape synthesis from sketches via procedural models and convolutional networks. IEEE transactions on visualization and computer graphics **23**(8), 2003–2013 (2016)
19. Huang, Q., Wang, H., Koltun, V.: Single-view reconstruction via joint analysis of image and shape collections. ACM Transactions on Graphics (TOG) **34**(4), 1–10 (2015)
20. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. Advances in neural information processing systems **30** (2017)
21. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
22. Kong, Z., Ping, W.: On fast sampling of diffusion probabilistic models. arXiv preprint arXiv:2106.00132 (2021)
23. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
24. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7708–7717 (2019)
25. Liu, S., Su, D., Yu, D.: Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. arXiv preprint arXiv:2201.11972 (2022)
26. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems **32** (2019)
27. Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., Wang, R.: 3d shape reconstruction from sketches via multi-view convolutional networks. In: 2017 International Conference on 3D Vision (3DV). pp. 67–77. IEEE (2017)
28. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
29. Lyu, Z., Kong, Z., Xu, X., Pan, L., Lin, D.: A conditional point diffusion-refinement paradigm for 3d point cloud completion. arXiv preprint arXiv:2112.03530 (2021)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
31. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
32. Pan, J., Han, X., Chen, W., Tang, J., Jia, K.: Deep mesh reconstruction from single rgb images via topology modification networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9964–9973 (2019)
33. Pontes, J.K., Kong, C., Sridharan, S., Lucey, S., Eriksson, A., Fookes, C.: Image2mesh: A learning framework for single image 3d reconstruction. In: Asian Conference on Computer Vision. pp. 365–381. Springer (2018)
34. Qi, A., Gryaditskaya, Y., Song, J., Yang, Y., Qi, Y., Hospedales, T.M., Xiang, T., Song, Y.Z.: Toward fine-grained sketch-based 3d shape retrieval. IEEE transactions on image processing **30**, 8595–8606 (2021)
35. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)

36. Shen, Y., Zhang, C., Fu, H., Zhou, K., Zheng, Y.: Deepsketchhair: Deep sketch-based 3d hair modeling. IEEE Transactions on Visualization and Computer Graphics **27**(7), 3250–3263 (2020)
37. Shu, D.W., Park, S.W., Kwon, J.: 3d point cloud generative adversarial network based on tree structured graph convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3859–3868 (2019)
38. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
39. Su, W., Du, D., Yang, X., Zhou, S., Fu, H.: Interactive sketch-based normal map generation with deep neural networks. Proceedings of the ACM on Computer Graphics and Interactive Techniques **1**(1), 1–17 (2018)
40. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2626–2634 (2017)
41. Valsesia, D., Fracastoro, G., Magli, E.: Learning localized generative models for 3d point clouds via graph convolution. In: International conference on learning representations (2018)
42. Wang, J., Lin, J., Yu, Q., Liu, R., Chen, Y., Yu, S.X.: 3d shape reconstruction from free-hand sketches. arXiv preprint arXiv:2006.09694 (2020)
43. Wang, L., Qian, C., Wang, J., Fang, Y.: Unsupervised learning of 3d model reconstruction from hand-drawn sketches. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1820–1828 (2018)
44. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–67 (2018)
45. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-gan: Training gans with diffusion. arXiv preprint arXiv:2206.02262 (2022)
46. Witkin, A.P.: Recovering surface shape and orientation from texture. Artificial intelligence **17**(1-3), 17–45 (1981)
47. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. In: International Conference on Learning Representations (ICLR) (2022)
48. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems **29** (2016)
49. Zhang, S.H., Guo, Y.C., Gu, Q.W.: Sketch2model: View-aware 3d modeling from single free-hand sketches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6012–6021 (2021)
50. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)