Sketch and Text Guided Diffusion Model for Colored Point Cloud Generation

Zijie Wu[1*] Yaonan Wang[1] Mingtao Feng[2†] He Xie[1] Ajmal Mian[3]
[1]Hunan University [2]Xidian University [3]The University of Western Australia

Abstract

Diffusion probabilistic models have achieved remarkable success in text guided image generation. However, generating 3D shapes is still challenging due to the lack of sufficient data containing 3D models along with their descriptions. Moreover, text based descriptions of 3D shapes are inherently ambiguous and lack details. In this paper, we propose a sketch and text guided probabilistic diffusion model for colored point cloud generation that conditions the denoising process jointly with a hand drawn sketch of the object and its textual description. We incrementally diffuse the point coordinates and color values in a joint diffusion process to reach a Gaussian distribution. Colored point cloud generation thus amounts to learning the reverse diffusion process, conditioned by the sketch and text, to iteratively recover the desired shape and color. Specifically, to learn effective sketch-text embedding, our model adaptively aggregates the joint embedding of text prompt and the sketch based on a capsule attention network. Our model uses staged diffusion to generate the shape and then assign colors to different parts conditioned on the appearance prompt while preserving precise shapes from the first stage. This gives our model the flexibility to extend to multiple tasks, such as appearance re-editing and part segmentation. Experimental results demonstrate that our model outperforms recent state-of-the-art in point cloud generation.

1. Introduction

Denoising Diffusion Probabilistic Models (DDPM) [19, 45] can generate novel images and videos, from their text descriptions. Recent diffusion models [38, 33, 39, 42] contain billions of parameters and are trained on millions of text-image [13] pairs sourced from the Internet, including existing datasets like MS-COCO [25]. The availability of such large corpora of text-image pair datasets is, arguably, the major driving force behind the success of research in image and video generation. However, such datasets of text-shape pairs are almost nonexistent causing a major bottleneck for advancing research in this direction.

3D shape generation has a wide range of applications including data augmentation, virtual/augmented reality [12], manufacturing, and reverse engineering. Pioneering works in diffusion based 3D shape generation include Luo *et al.* [26] and Point Voxel Diffusion (PVD) [66]. Luo *et al.* [26] extend the DDPMs to 3D shape generation conditioned on shape latent that follows a prior distribution parameterized via normalized flows [9]. Their probabilistic generative model treats points as particles that, under heat, are diffused to a noise distribution. It then exploits the reverse diffusion process to learn the point distribution to be able to generate plausible point clouds. PVD [66] combines the denoising diffusion model with the point-voxel 3D shape representation. It performs the denoising steps by optimizing a variational lower bound to the conditional likelihood function to produce 3D shapes or multiple 3D completions [14] from a partial observation.

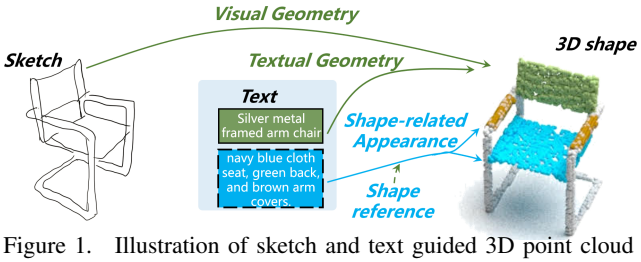

Figure 1. Illustration of sketch and text guided 3D point cloud generation. The text and sketch complement each other.

The problem of text guided 3D model or point cloud generation is slightly different from that of images in the sense that there is a higher emphasis, at least implicitly, on geometry rather than appearance which puts yet another constraint on the training datasets. Current methods [26, 66] have mostly used the ShapeNet [3] dataset for text-shape learning combined with the ModelNet [55] datasets for 3D representation learning and validation. However, these datasets are relatively small and lack detailed textual descriptions. To work around the data paucity, CLIP-forge [42] extends the CLIP model [37] for 3D shape generation by aligning 3D shapes with text in the latent space. Although these text guided methods maintain diversity in the generated shapes, their common limitation is that the generated shapes are often not the desired ones, which is understandable given the shape ambiguity in the text.

In this paper, to reduce the shape ambiguity, we propose a sketch and text guided probabilistic diffusion (STPD) model for colored point cloud generation that conditions the

*Work performed during visit at the University of Western Australia
†Correponding author.

denoising process with sketch inputs, in addition to textual descriptions. Sketches can be hand drawn and give much more geometric details compared to text. We argue that, combined with text, sketches are a viable option for conditioning the denoising of 3D shapes given their direct relevance to the problem at hand and their history of successful use for 3D object reconstruction [65]. However, cross-modality inputs (sketch and text) [11] always suffer from large cross-modality discrepancies and intra-modality variations bringing in additional challenges of cross-modality fusion [28] and data sparsity. Especially, sketches are inherently sparse with only a very small proportion of informative pixels, and contain much less information than natural images.

The proposed STPD model can produce the geometry and appearance simultaneously and unambiguously from a sketch-text input, combining ease of access and completeness of description. It takes the sketch as the main shape description and shape information present in the text to complement the sketch input. Hence, shape information is extracted from the sketch as well as the text when there is shape information in the text. Appearance is extracted from the text input to give color to the generated point clouds as shown in Fig. 1. To summarize, our contributions are:

1. We propose a staged probabilistic diffusion model for better control over the geometry and appearance to generate colored point clouds. Our model conditions the denoising process jointly with sketch and text inputs to generate 3D shapes. It can also be used for part-segmentation and appearance re-editing.
2. We propose an attention capsule based feature extraction module for encoding sketches, which are inherently sparse. Our method robustly gives more attention to the entities of the useful pixels and ignores the large number of meaningless capsules corresponding to the blank pixels in the sketch.
3. We present a sketch-text fusion network that efficiently abstracts the shape and color information from both sketch and language descriptions to guide the reverse 3D shape diffusion process.

Extensive experiments on the ShapeNet dataset [3] and comparison to existing diffusion probabilistic model based generation as well as some classical shape reconstruction methods show that our model achieves state-of-the-art performance for colored point cloud generation. We also show the representation learning ability of our model by conducting 3D object classification experiments on the ModelNet40 [55] dataset and the application of our model to part segmentation on the ShapeNet-Parts dataset [3].

2. Related Work

We first survey diffusion models for generating images. Next, we discuss 3D shape reconstruction methods followed by classical 3D shape generation methods. Finally, we discuss diffusion models for 3D shape generation.

2.1. Text-based 2D Image Generation

Generative Adversarial Network (GAN): Early methods for text based image generation were based on GANs. Xu *et al.* [59] used attention mechanism for fine-grained text to image generation. TediGAN [56] uses GAN inversion based on the pretrained StyleGAN [1] for text to image generation. Zhang et al.[63][64] used multiple GAN architectures that stacked generators and discriminators to progressively generate multi-scale images. GANs are known to be difficult to train and have recently been surpassed by diffusion models [8].

Probabilistic Diffusion Models: Sohl-Dickstein et al. [45] proposed diffusion probabilistic models that are both tractable and flexible. Inspired by non-equilibrium thermodynamics, the diffusion process uses Markov chain [6] to convert a simple Gaussian distribution to a target data distribution. Keeping the steps of the chain small enough, each step can be analytically evaluated [16], and hence, the full chain can also be evaluated. At each step of the forward pass, a small perturbation is added to the data until it reaches the Gaussian distribution. Since the perturbation at each step is small, a model can be learned to estimate and remove the perturbation. It was not until 2020 that Ho et al. [19] showed the real power of diffusion models with high quality image generation by training on a weighted variational bound and denoising score matching with Langevin dynamics. Dhariwal and Nichol [8] showed the superiority of diffusion models over GANs. However, directly optimizing in pixel space comes with a high computational cost. To reduce the computational requirements, [39] performs diffusion in the latent feature representation space. Large models such as Imagen [41] and DALL-E2 [38], which uses the CLIP model [37], can generate diverse photo-realistic images. The success of diffusion models is largely attributed to fundamental research as well as the availability of huge datasets in the image domain.

2.2. 3D Reconstruction from Images

Reconstructing 3D shapes from images has been studied for decades. Single-view reconstruction techniques are popular, however, recovering 3D shapes from a single image is an ill-posed problem, as it requires a strong prior shape knowledge. Fan *et al.* [10] proposed to directly regress point clouds from a single image. Wang *et al.* [51] deformed the ellipsoid mesh to form the 3D shape to learn the shape prior. Sketch2Mesh [17] employed the silhouettes matching match between a given single sketch and the learned model to refine the 3D shape. TMNet [32] uses triangular meshes and images to supervise the reconstruction of 3D shapes.

Multiview images contain more geometric information useful for 3D reconstruction. Pix2Vox++ [58] fuse the mul-

tiple coarse 3D volumes, which are generated from each individual view, to form the well-refined 3D shapes. Wang *et al.* [49] include transformers to recover 3D information from multiviews. Reconstruction based methods usually demand exact image-to-shape match and strong prior shape knowledge, whereas generative methods offer a balance between diversity and alignment with user specifications.

2.3. Sketch, Text and 3D Shape Generation

Sketch to 3D Shape: Sketches have been used as a sparse representation of natural images and 3D shapes. Sketches are quite illustrative, despite their abstract nature and simplicity. Sketch based 3D generation [31] can be coarsely categorized into evocative [50] and constructive [29] methods, which depend on retrieval from model collections or free-form modeling. Deep models learn the pattern of the sketch for 3D model generation. For example, Zhang *et al.* [65] proposed a view-aware technique to gain priors from existing 3D models using a single freehand sketch. DeepSketchHair [44] proposed a learning based system for intuitive 3D hairstyling from 2D sketches. These methods have less dependence on a shape prior to the generation phase, but the tremendous requirement of shape-related data during training for pattern learning.

Text to 3D Shape: Text to shape generation is a relatively new direction. Examples include Text2shape [4] and CLIP-Forge [42]. Text2shape [4] generates colored 3D shapes using joint embedding of text and 3D shapes. CLIP-Forge [42] generates 3D shapes from the pre-trained text-to-image CLIP [37], addressing the lack of text-to-shape priors data. Although previous studies have considered sketch and text as flexible methods for 3D shape generation, they only refine the model in an implicit pattern that takes the shape priors library into optimization while ignoring the intrinsic information mining at the sketch and text level.

2.4. Diffusion Models for 3D Shape Generation

Recent methods [26, 66, 62, 33] generate 3D shapes with denoising diffusion probabilistic models. DPM [26] introduced the flow-based model for point cloud embedding that conditions the diffusion model on geometry representation. PVD [66] trains a diffusion model on the point-voxel representation for 3D shapes. It performs the variational lower bound optimizing on unconditional shape generation and conditioned shape completion to produce complete 3D shapes. These methods show great promise in extending the diffusion theory to 3D shape generation and achieving good performance. However, their generations are relatively uncontrolled and lack certainty for our generation problem which needs to follow alignment with user instructions more closely. LION [62] proposed a text-based 3D generation model by combining the CLIP-forge to align the text-to-3D embedding. DreamFusion [33] is a generative method for text based 3D rendering, which replaces CLIP with a loss derived from the distillation of 2D diffusion models. Point-E [30] starts from a language prompt to generate the single view image with fine-tuned GLIDE model and then uses the generated image for image-to-shape diffusion. Nonetheless, these text-based 3D diffusion models rely on feature pre-alignment from text to image and usually fail when there are huge domain gaps. Our method focuses on information mining for 3D diffusion by exploring the 3D structure from sketch and text to facilitate efficient controlled generation of high quality 3D shapes.

3. Proposed Approach

Given a sketch $\mathbb{S}$ and text description $\mathbb{T}$ of an object, the goal is to generate its 3D geometry and appearance. To achieve this, the proposed STPD model consists of three modules: sparse feature extraction network for sketch embedding (Section 3.1), sketch-text feature fusion (Section 3.2), and staged diffusion for shape and color generation (Section 3.3). Figure 2 shows an overview of our model.

3.1. Sparse Feature Extraction of a Sketch

Feature extraction from a sketch faces three major challenges. Firstly, sketches are sparse and consist of very few informative pixels. Secondly, visual models are generally trained on large amounts of natural images and their attention mechanisms exploit the information in dense pixels to focus on the more relevant parts. This becomes challenging when the image (i.e. sketch) contains copious amounts of futile background pixels and the training dataset is also small. Finally, the relationship between different object parts is ambiguous due to self-occlusions in a single sketch. Below, as shown in Fig. 3, we present our capsule network using dynamic routing [40] with attention to effectively learning the sparse features from a sketch image.

Primary Capsule Layer: Our sketch extraction starts from a pixel-wise CNN embedding $S_{N,D_{input},W,H} = \boldsymbol{f}_{cnn}(\mathbb{S})$, where N is the core number of caps, D is the dimension of each caps, and W, H are the width, height of the sketch. Based on these embeddings, we perform a 1×1 $Conv$ of filter size D to share an affine transformation on each capsule channel and then turn it to a primary capsule $u^{i,0}_{N,D,w,h}$

$$u^{i,0}_{(N,D,w,h)} = \text{Conv}(S_{(N,D,w,h)}). \tag{1}$$

Here, i denotes the index of capsules in the current layer, w, h is the size of current caps layers and Conv is a simple network comprising a 1×1 convolution layer with stride 1, followed by batch normalization, ReLU, and max pooling.

Dynamic Routing with Attention: Dynamic routing is applied to strengthen the structural relationship inference between different sketch parts. This facilitates robust feature learning from sketches while keeping the network depth in control and the number of parameters low. We add an attention module to force the capsules to focus on the useful sketch parts and ignore the background. Each s^j can be acquired by the sum of previous layer capsules,

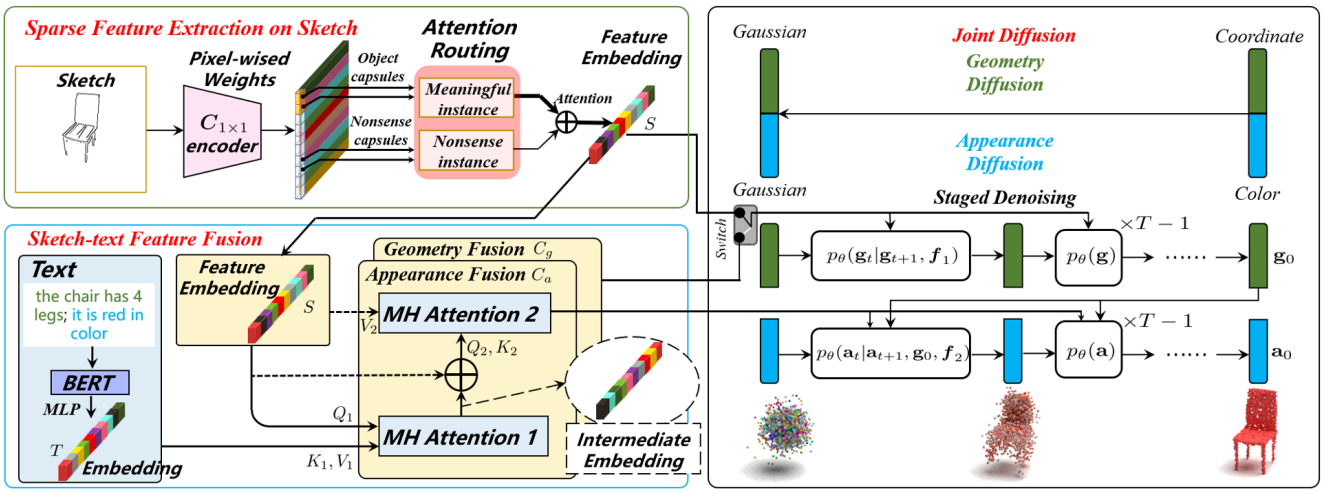

Figure 2. Overview of STPD framework. The sketch feature is first extracted with two level attention modules for sparse embedding and then flows with multi-head attention fusion to text. With the embedded features, we proposed a staged denoising probabilistic diffusion model for both the geometry and appearance of 3D shapes.

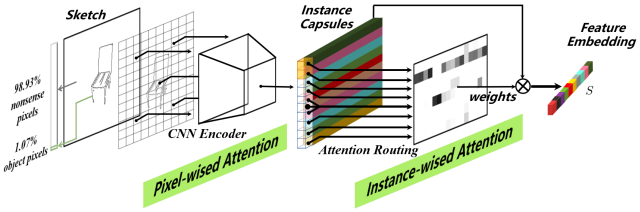

Figure 3. Proposed sketch feature extractor. Pixel-wise attention and instance-wise attention are implemented by a CNN and attention routing, respectively. The two level attentions are designed to extract the sparse information from hand drawn sketches.

$$s^j_{(N,D,w,h)} = \sum_{i=1,2,...} c^{ij}_{w,h} a^{ij}_{(N,D,1,1)} u^{i,j-1}_{(N,D,w,h)}, \quad (2)$$

where c^{ij} are the dynamic routing parameters computed by softmax layers and a^{ij} is the attention weight obtained by 1×1 Conv on lower level primary capsules and hence associated with each instance of the original sketch.

Similar to the l times dynamic routing, the capsules are squashed [40], to cluster the informative parts, and then reshaped to a one dimension vector $S_{(1,D_f)}$ (where $D_f = N \times D$) from the last s^j. This S forms the sketch embedding used to condition the reverse diffusion process.

3.2. Sketch-Text Feature Fusion Embedding

We encode the input language expression as a sequence of textual embeddings with BERT [7]. Applying a linear projection to the BERT embedding, we get embedding T of the same dimension as the sketch embedding S. Both embeddings are fused with two multi-head attention modules to guide the reverse diffusion process.

The attention-ed embedding is formulated as a set of cascaded multi-attention modules. The first attention takes the sketch feature S as query to extract the relevant *intermediate* semantic information I from the text embedding. The second attention performs text-guided self-attention for the sketch feature to produce the conditional embedding that contains geometry/color information based on I. Architectures for geometry and color diffusion are the same but have different parameters. The first multi-head attention (MH Attention 1 in Fig. 2) is defined as:

$$I_{(1,D_f)} = Atten_1(Q_1 = S, K_1 = V_1 = T), \quad (3)$$

where Q_1, K_1, V_1 are the Query, Key, Value and I is an intermediate attention-ed feature passed to the second multi-head attention module (MH Attention 2) defined as:

$$C_{p(1,D_f)} = Atten_2(Q_2 = K_2 = I + S, V_2 = S), \quad (4)$$

where Q_2, K_2, V_2 are the Query, Key, Value and $p = g$ for geometric diffusion and $p = a$ for appearance diffusion. Specifically, we take the sum of intermediate features I with the sketch feature S as the input of the query and key in the second multi-head attention. Then, the text prompt is introduced as the vector input to support the appearance denoising using the color information from the attention module. Note that, C_g and C_a are from the same cascaded attentions but focus on different descriptions (geometry or appearance) in each diffusion process. Appearance condition also requires information from sketch (shape) because different object parts can have different colors.

3.3. Staged Colored Point Cloud Generation

Markov's chain: Given a sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, noise is systematically added to $\mathbf{x}_0$ in the forward process [19] step-by-step following the Markov's chain assumption until the Gaussian distribution is reached:

$$\begin{aligned} q(\mathbf{x}_{0:T}) &= q(\mathbf{x}_0) \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \\ q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) &= \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \end{aligned} \quad (5)$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes a Gaussian distribution. Here, $\mathbf{x}_0$ can represent either coordinates or colors from 3D shapes. We train geometry and appearance diffusion separately to generate colored point clouds. The reverse process, $p_\theta\left(\mathbf{x}_t \mid \mathbf{x}_{t+1}\right)$, starts from a standard Gaussian prior and ends with the desired $\mathbf{x}_0$:

$$p_\theta\left(\mathbf{x}_{0:T}\right) = p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right), \\ p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mu_\theta\left(\mathbf{x}_t, t\right), \sigma_t^2 \mathbf{I}\right). \tag{6}$$

This cross-entropy minimization is designed such that it forces the reverse process to follow the pattern of noise addition of the forward process, and eventually, recover $\mathbf{x}_0$ from the standard Gaussian noise distribution. Hence, the loss function can be derived as:

$$L_t = \mathbb{E}_{\mathbf{x}_0, \epsilon}[C \left\|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_t(\mathbf{x}_t, t)\right\|^2], \tag{7}$$

where C is a constant.

Geometry Diffusion: Denoting the colored point cloud as $\mathbf{x}_0 \in \mathbb{R}^{N\times 6} = \{\mathbf{g}_0 \in \mathbb{R}^{N\times 3}, \mathbf{a}_0 \in \mathbb{R}^{N\times 3}\}$, consisting N points with geometry $\mathbf{g}_0$ and appearance $\mathbf{a}_0$ coordinates. We first perform geometry diffusion to generate the 3D shape. STPD is trained by minimizing the cross-entropy loss between two diffusion chains with respect to the geometric conditional feature C_g, and learning of model parameters θ:

$$\min_\theta \mathbb{E}_{g_0\sim q(g_0), g_{1:T}\sim q(g_{1:T})}[\sum_{t=1}^{T} \log p_\theta(\mathbf{g}_{t-1}|\mathbf{g}_t, C_g)], \tag{8}$$

where C_g is derived from sketch-text embedding when the text also contains shape description, or only from sketch when there is no text. We add a switch to adjust the feature map selection for different scenarios. The joint posterior $q(g_{1:T})$ can be factorized by the assumption of Markov's chain $\prod_{t=1}^{T} q(\mathbf{g}_{t-1}|\mathbf{g}_t, \mathbf{g}_0)$, where each is analytically tractable by a re-parameterized Gaussian distribution:

$$\tilde{\boldsymbol{\mu}}_t = \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0, \tag{9}$$

$$q\left(\mathbf{g}_{t-1} \mid \mathbf{g}_t, \mathbf{g}_0\right) = \mathcal{N}\left(\mathbf{g}_{t-1}; \tilde{\boldsymbol{\mu}}_t, \tilde{\beta}_t \mathbf{I}\right), \tag{10}$$

where $\tilde{\beta}_t$ is the known variance defined in the forward process, $\alpha_t = 1 - \beta_t$, and $\tilde{\alpha} = \prod_{s=1}^{t} \alpha_s$. From Eq. 7 and this property, our diffusion model learns to match each $q\left(\mathbf{g}_{t-1} \mid \mathbf{g}_t, \mathbf{g}_0\right)$ and $p_\theta\left(\mathbf{g}_{t-1} \mid \mathbf{g}_t\right)$ by estimating the output noise $\epsilon_\theta(\mathbf{g}_t, C_g, t)$ to match noise ϵ:

$$\left\|\epsilon - \epsilon_\theta(\mathbf{g}_t, C_g, t)\right\|^2, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{11}$$

The noise prediction network ϵ_θ progressively generates the 3D shape using the reverse chain

$$g_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{g}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(g_t, C_g, t)\right) + \sqrt{\beta_t}\epsilon.$$

Appearance Diffusion: Appearance diffusion is performed while freezing the geometry parameters. The optimization function, in this case, can be simply modified from Eq. 8 to learn a conditional generative model for appearance:

$$\min_\theta \mathbb{E}_{a_0\sim q(a_0), a_{1:T}\sim q(a_{1:T})}[\sum_{t=1}^{T} \log p_\theta(\mathbf{a}_{t-1}|\mathbf{a}_t, p_\theta(\mathbf{g}_0), C_a)]$$

The training loss for appearance diffusion is minimized and conditioned on appearance features C_a and $p_\theta(\mathbf{g}_0)$:

$$\left\|\epsilon - \epsilon_\theta(\mathbf{a}_t, p_\theta(\mathbf{g}_0), C_a, t)\right\|^2, \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{12}$$

Colored Point Cloud Generation: We can now formally define our joint generative model for colored point cloud generation as

$$p_{\theta_2,\theta_1,\phi}(\mathbf{a}_0, \mathbf{g}_0, C_g, C_a) = p_{\theta_2}(\mathbf{a}_0|\mathbf{g}_0, C_a) p_{\theta_1}(\mathbf{g}_0|C_g) p_\phi(C_g, C_a),$$

where $\{\mathbf{a}_0, \mathbf{g}_0\}$ denote the colored point cloud $\mathbf{x}_0$. $p_\phi(C_g, C_a)$ refers to the sketch and textual embedding, and $p_{\theta_2}(\mathbf{a}_0)$, $p_\phi(\mathbf{g}_0)$ are the appearance and geometry denoising processes respectively. Note that the appearance diffusion is conditioned on the geometry diffusion as well. We also tried extended diffusion to denoise from a 6D Gaussian rather than treating the shape and appearance separately. However, integrating the unrelated features causes undesirable interference in the diffusion process and consequently, STPD does not learn to produce point clouds with fine details.

4. Experiments

4.1. Experimental Setup

Datasets: We use the *chair*, *table*, *aeroplane*, *car* categories of ShapeNet dataset [3], randomly split it into training-test sets at 80-20% ratio and render a single-image from each 3D object. We apply Canny edge detection [24] on the rendered ShapeNet images to acquire their single-channel edge images. We manually remove the background so that they resemble clean hand drawn sketches. For colored point clouds, we use the *chair*, *table* categories as only they have text descriptions [4]. To generate ground-truth point clouds, each 3D shape is randomly sampled at 2048 points and then translated to zero mean and normalized in the range $[-1, 1]^3$ so that the Gaussian distribution assumption is satisfied during the forward pass.

Evaluation Metrics: Previous works [26] used Chamfer Distance (CD) and Earth Mover's Distance (EMD) to evaluate the reconstruction quality of point clouds. Based on these, existing works report Coverage (COV) and Minimum matching distance (MMD). Although these metrics have their limitations, they still serve as a good reference to measure and compare the performance of generative methods. Specifically, COV evaluates the certainty of conditioned generation whereas MMD calculates the average distance of the point clouds in the reference set and their closest neighbors in the generated set. Another metric used in numerous works to evaluate the generation quality and diversity is 1-nearest neighbor accuracy (1-NNA) which directly quantifies distribution similarity between two sets. However, since

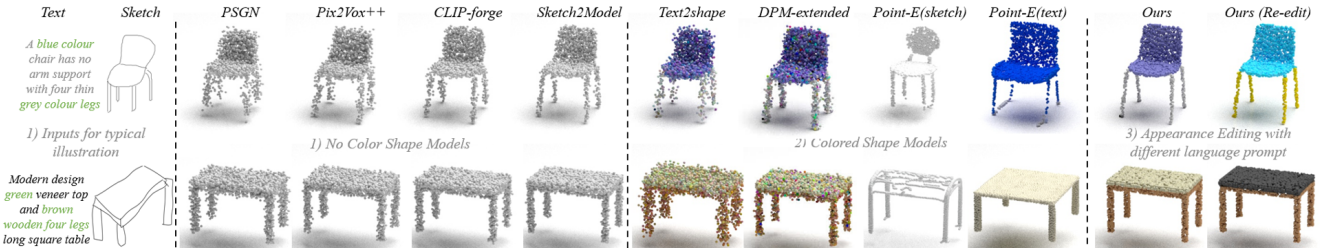

Figure 4. Visual comparison of the generated objects. 1) Inputs include text and free-hand sketch. 2) Shape methods with no color. 3) Colored shape generation methods. 4) Our method generates both geometry and appearance using sketch and text. The appearance can be re-edited by changing the language prompt while preserving geometry.

Categories	Input	Methods	MMD (↓) CD	MMD (↓) EMD	COV (%, ↑) CD	COV (%, ↑) EMD
Reconstruction Shape (N×3)	Sketch	PSGN [10]	8.174	7.253	64.13	69.76
		Pix2Mesh [51]	9.771	7.347	65.61	67.86
		TMNet [32]	9.18	8.11	63.55	65.24
		Sketch2Mesh [17]	3.018	2.786	80.24	78.23
	Sketches	Pix2Vox++ [57]	5.976	5.760	71.29	69.55
		EvolT [49]	5.718	5.307	73.09	74.84
Generative Shape (N×3)	Text	Text2Shape [5]	9.873	6.348	33.76	31.28
		CLIP-forge [43]	4.367	3.818	35.26	33.17
	Sketch	MixNF [34]	7.689	6.224	69.89	67.24
		DPF-Net [22]	7.531	6.176	72.34	67.68
		Sketch2Model [65]	3.176	2.967	78.69	77.82
	×	DPM [26]	3.764	1.986	46.87	44.59
	Sketch,text	STPD (ours)	**1.635**	**1.043**	**93.53**	**91.37**
Generative Shape&Color (N×6)	Text	Text2Shape	10.34	7.476	34.59	33.26
		Point-E (text) [30]	2.523	2.265	51.96	49.23
	×	DPM-extended	4.765	3.256	45.76	39.68
	sketch	Point-E (sketch)	3.092	3.117	61.12	62.09
	Sketch,text	STPD (ours)	**1.346**	**1.017**	**95.53**	**94.97**

Table 1. Comparison of 3D shape generation for the *chair*, *table* categories of ShapeNet [3]. CD is multiplied by 10^3 and EMD by 10^1. Our method generates the best shape, with or without color.

our aim is to generate 3D shapes faithful to the provided sketch, we stick to the first four metrics. Where other methods do not provide color, we compare shapes only.

Baselines: We compare with classical as well as diffusion based methods. In the former category, we select the major sketch [65], image [34] [22], and text [43] [5] based 3D object generation methods. We also compare to reconstruction based methods including single-view [10, 51, 32, 17] and multiviews [58, 49]. In the latter category, we select 3D diffusion (DPM) [26], Point-E [30], and Point Voxel Diffusion (PVD) [66]. Point-E (sketch) results are obtained after fine-tuning the Point-E image-to-shape diffusion model on our sketch-shape data for a fair comparison. We take ten generations for all methods and use the average error from ground truth for evaluation.

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

Table 1 compares the point cloud generation quality of our method with existing state-of-the-art. For a fair comparison, we re-train all methods with our sketch-shape data of the *chair*, *table* categories, or their text descriptions when the method takes text only as input. Multiview methods are given sketches from 3 viewpoints. Only our method performs (single view) sketch and text guided 3D object generation. Since DPM [26] is unconditional, we generate many shapes and pick the best matches. For shape only (upper half of Table 1), the proposed STPD outperforms all methods, achieving 45.8% lower CD-MMD error from the nearest reconstruction method Sketch2Mesh [17] and 48.5% lower CD-MMD error from the nearest generative method Sketch2Model [65]. The multiview methods do not perform well since they are unable to establish correspondences between the sketches. Text guided generative methods have more diversity but struggle with alignment to user specifications given the inherent shape ambiguity in the text. Sketch guided methods perform better in general.

Table 1 also compares colored point cloud (shape and color) generation. For this, we extend DPM to a 6D Gaussian and also compare with Text2Shape [4] and Point-E [30]. The proposed STPD again outperforms all methods, achieving 46.7% CD-MMD error from the nearest competitor Point-E [30]. In Fig. 4, we compare our results visually to the existing methods using hand-drawn sketches from SketchX [23]. Our method also offers to re-edit the color (appearance) once the geometry is finalized (last column). Table 2 reports comparative results on the *chair*, *table*, *aeroplane*, *car* categories of ShapeNet. We choose these categories because PVD [66] provides comparisons on the *chair*, *aeroplane*, *car* categories and the *table* category is used in our previous experiment. From Table 2, we can see that STPD achieves the best performance on both evaluation metrics on all four object categories. SoftFlow [21], PointFlow [60], DPF-Net [22], DPM [26], Point-E [30], PVD [66] achieve an average Chamfer Distance of 2.464, 2.429, 2.168, 2.509, 2.182, 2.539 and an average EMD of 3.105, 3.191, 2.851, 2.553, 2.490, 2.661 respectively. Our STPD achieves 1.172, 1.261 average CD, EMD errors which are 46.3%, 49.3% lower than the nearest competitor Point-E [30].

Human Evaluations: We also conducted human evaluations on 100 objects generated by our method and other methods. Table 3 shows the results where 30 people rate

Methods	chair		table		aeroplane		car	
	CD	EMD	CD	EMD	CD	EMD	CD	EMD
SoftFlow [21]	2.786	3.295	4.815	5.139	0.404	1.198	1.850	2.789
PointFlow [60]	2.707	3.649	4.804	5.082	0.403	1.180	1.803	2.851
DPF-Net [22]	2.763	3.320	3.986	4.661	0.528	1.105	1.396	2.318
DPM [26]	3.201	2.718	4.601	4.298	0.432	1.027	1.804	2.167
Point-E [30]	2.901	2.667	3.809	4.290	0.412	1.005	1.609	1.997
PVD [66]	3.211	2.939	4.731	4.527	0.442	1.030	1.774	2.146
STPD (ours)	**1.209**	**1.117**	**2.117**	**1.897**	**0.311**	**0.897**	**1.049**	**1.136**

Table 2. Comparison with baselines on 4 categories of the ShapeNet dataset [3]. CD is multiplied by 10^3 and EMD by 10^1.

Human rate (↑)	SoftFlow	PointFlow	DPF-Net	DPM	Point-E	PVD	STPD
chair	1.8	1.7	1.9	2.1	3.9	3.5	**4.8**
table	2.1	1.9	3.5	3.0	3.4	2.5	**4.6**
aeroplane	3.4	3.3	2.2	3.1	3.7	3.0	**4.0**
car	2.5	2.4	3.4	3.5	3.2	2.9	**4.3**

Table 3. Human evaluations: 100 generated objects rated on a scale of 1 to 5 by 30 people.

Input Modality/Model for Conditioning		CD	EMD
Sketch	Text	($\times 10^3$)	($\times 10^1$)
×	BERT	28.325	20.457
$C_{1\times1}$	×	10.965	8.658
$C_{1\times1}$	BERT	9.651	8.459
ResNet50	BERT	6.409	5.480
ResNet50 + Atten.Routing	BERT	1.935	1.587
$C_{1\times1}$ + Atten.Routing	×	1.689	1.348
$C_{1\times1}$ + Atten.Routing (concatenate)	BERT	3.748	4.127
$C_{1\times1}$ + Atten.Routing (our fusion)	BERT	**1.390**	**1.101**

Table 4. Ablation study on visual & textual modules. $C_{1\times1}$ is a simple visual encoder comprising 1×1 convolution followed by batch normalization and ReLU.

the generations on a scale of 1 to 5 (higher is better). STPD outperforms the baseline methods on all individual object categories and achieves an average rating of 4.4, compared to average rating of 3.3 by the nearest competitor Point-E.

4.3. Ablation Study

We perform ablation study on the sketch and text modules to verify their contributions. Table 4 shows our results using CD and EMD metrics. Text conditioning, using BERT embeddings, alone achieves the worst performance. $C_{1\times1}$ is a simple visual encoder comprising 1×1 convolution followed by batch normalization and ReLU. Sketch conditioning, using simple $C_{1\times1}$ embeddings, reduces the errors significantly since a sketch can describe shaping much better than text. Using sketch and text conditioning with $C_{1\times1}$ with BERT embeddings, the errors do not reduce much. However, when a more sophisticated sketch encoding is performed with ResNet50 [18] and combined with BERT, the CD and EMD reduce notably. To account for the sparsity of sketches, when we add capsule network with attention routing on top of ResNet50, we see a dramatic improvement in both metrics. Our simple visual encoder $C_{1\times1}$ with attention routing further improves the performance even without using text conditioning. As seen in the last row of the table, the best performance is achieved when we use sketch conditioning with $C_{1\times1}$ + attention routing embeddings combined with text embeddings from BERT [7] using the proposed fusion method. The second last row of the table shows that a simple concatenation of the two embeddings does not work as well as our proposed fusion mechanism.

Model	ModelNet10	ModelNet40	ScanObjectNN
PointFlow [60]	93.7	86.8	78.2
PDGN [20]	94.2	87.3	78.1
3D-GAN [54]	91.0	83.3	76.7
AtlasNet [15]	91.9	86.6	76.9
ShapeGF [2]	90.2	84.6	75.8
DPM [26]	94.2	87.6	79.7
STPD	**94.8**	**88.1**	**84.3**

Table 5. Object classification accuracy (%).

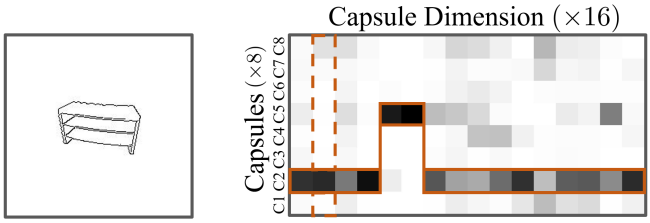

Figure 5. Visualization of attention score maps for a sketch with respect to capsules. The 8× 16 size denotes 8 capsules of 16-dimension each. Capsule attention maps are learned from attention-routing. Dark parts denote high attention.

4.4. Representation Learning Evaluation

We evaluate the representation learned by our sketch embedding used to condition the diffusion model. We follow the work of MVCNN [46], ScanObjectNN [48] and convert rendered images from different viewpoints of the ModelNet dataset [55] to sketches and conduct object classification experiments. We train our STPD on the generated sketches and then take the conditioned embedded features to learn a linear SVM classifier for classification based on a single sketch embedding. The essence of the conditional diffusion model is that the conditioning should guide the diffusion to retrieve the 3D shape from Gaussian noise. This means that our conditional network should be able to represent specific object categories to aid classification. The results in Table 5 demonstrate that our architecture can extract discriminative features, that are also useful for accurate classification.

4.5. Effectiveness of Capsule Network

Fig. 5 visualizes the generated attention scores from a typical sketch. More examples are given in the supplementary material along with quantitative results using Instance Scores for Sketch (ISS i.e. without attention routing) and Instance Score on Capsule (ISC i.e. with attention routing) metrics. ISS is the percentage of useful pixels divided by the total pixels of the sketch. For ISC, we compute the aver-

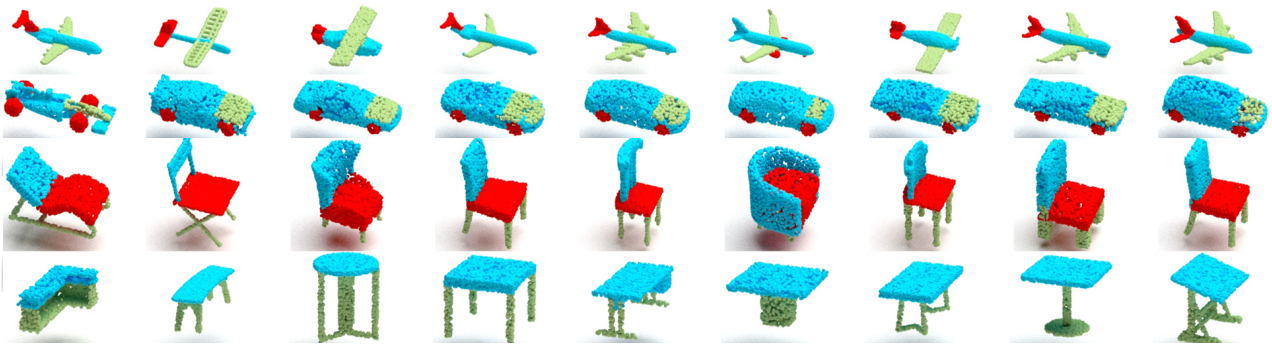

Figure 6. Visualization of part segmentation results for 4 categories of ShapeNet-Parts.

age of the highest attention score for each of the 16 capsule dimensions (column wise) and convert it to a percentage.

Before the attention capsule, the sketches have ISS of 1.07% on average. After our capsule-attention module, ISC becomes 58.68% on average for better diffusion. This demonstrates the effectiveness of the Capsule Attention Network which attends to sparse data in sketches during diffusion training. Each capsule (1×16) denotes a learned feature, e.g. the table, from a sketch. It can be observed that given the distribution of the useful pixel in the sketch, only a few attention score values are dark i.e, attentioned. Caps number 2 and a few dimensions of Caps number 5 are given high weight to contribute to the final visual feature, while the other Caps that represent the white background of the sketch are ignored. The attention-based capsule network enables our method to effectively locate the referred objects in the sparse sketch input.

5. Extending STPD to Part Segmentation Task

To further show the feature learning ability of our STPD model, we use it for the task of part segmentation in point clouds. For this, we use the ShapeNet-Parts [3] dataset, focusing on the four object categories of Table 2.

The part segmentation task can be extended from our colored point cloud generation model by assigning different colors to each part. We consider only two parts for the $table$ and three parts for remaining three categories e.g. for $aeroplane$, the ($body$, $wing$, $tail$) are assigned ($blue$, $green$, red) colors respectively.

In our STPD , each coordinate is generated by our sketch-text guided geometric diffusion model. However, in the part segmentation task, the shape $\mathbf{g}_0$ is given as input to the model. Here, we directly apply the text feature as the condition C_a of appearance diffusion. For instance, we provide a generic text sentence "*An blue aeroplane with green wings and red tail*" for all aeroplanes.

With the given input geometry and appearance description, we can formally define our joint generative model for point cloud segmentation as

$$p_{\theta,\phi}(\mathbf{a}_0, \mathbf{g}_0, C_a) = p_\theta(\mathbf{a}_0|\mathbf{g}_0, C_a)p_{input}(\mathbf{g}_0)p_\phi(C_a), \quad (13)$$

Model	mIoU$_c$	mIoU$_I$	chair	table	aeroplane	car
PointNet [35]	82.1	83.5	89.6	80.6	83.4	74.9
PointNet++ [36]	83.3	84.6	90.8	82.6	82.4	77.3
DGCNN [52]	83.8	84.9	90.6	82.6	84.0	77.8
Point-BERT [61]	84.7	85.6	91.0	81.5	84.3	79.8
PointMLP [27]	84.6	84.8	90.3	84.3	83.5	80.5
KPConv [47]	85.1	85.9	91.1	83.6	84.6	81.1
P2P [53]	85.0	85.9	91.6	83.7	84.3	80.4
STPD	**85.9**	**86.7**	**92.4**	**84.1**	**85.6**	**81.3**

Table 6. Comparative results of part segmentation on four categories of ShapeNet-Parts dataset.

where the combination $\{\mathbf{a}_0, \mathbf{g}_0\}$ denotes the desired segmentation of the 3D point cloud. Note that, $\mathbf{a}_0$ is generated as colors which are then used as segmentation labels. We apply K-means clustering for final refinement of the colors.

Results for 3D Shape Segmentation: Fig. 6 shows some visualizations of our results and Table 6 compares our method to some existing baselines. The proposed STPD achieves the highest mIoU compared to existing methods.

6. Conclusion

We proposed a novel sketch and text guided probabilistic diffusion model for 3D object generation. Our model takes a hand drawn sketch and text as inputs to condition the reverse diffusion process. With the strong generative ability and effective guidance by the sketch-text fused features, our model is able to achieve high alignment with user specifications. Our visual encoder is designed to focus on the informative sparse parts of sketches. The sketch-text fusion part gathers informative object context from visual and textual descriptions to improve the shape and appearance of the generated 3D objects. Experimental results demonstrate that our model achieves state-of-art performance in 3D object generation and shows promising applications for 3D object classification and part segmentation.

7. Acknowledgement

Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

References

[1] A.H. Bermano, R. Gal, Y. Alaluf, R. Mokady, Y. Nitzan, O. Tov, O. Patashnik, and D. Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. *Computer Graphics Forum*, 41(2):591–611, 2022. 2

[2] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 5, 6, 7, 8

[4] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pages 100–116. Springer, 2018. 3, 5, 6

[5] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pages 100–116. Springer, 2018. 6

[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, January 2023. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1

[10] 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 conference on computer vision and pattern recognition*, pages 605–613, 2017. 2, 6

[11] Mingtao Feng, Syed Zulqarnain Gilani, Yaonan Wang, Liang Zhang, and Ajmal Mian. Relation graph network for 3d object detection in point clouds. *IEEE Transactions on Image Processing*, 30:92–107, 2020. 2

[12] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023. 1

[13] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3722–3731, 2021. 1

[14] Zhiheng Fu, Siyu Hong, Mengyi Liu, Hamid Laga, Mohammed Bennamoun, Farid Boussaid, and Yulan Guo. Multi-stage information diffusion for joint depth and surface normal estimation. *Pattern Recognition*, 141:109660, 2023. 1

[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 7

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, June 2022. 2

[17] Benoit Guillard, Edoardo Remelli, Pierre Yvernay, and Pascal Fua. Sketch2mesh: Reconstructing and editing 3d shapes from sketches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13023–13032, October 2021. 2, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 4

[20] Le Hui, Rui Xu, Jin Xie, Jianjun Qian, and Jian Yang. Progressive point cloud deconvolution generation network. In *European Conference on Computer Vision*, pages 397–413. Springer, 2020. 7

[21] Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33:16388–16397, 2020. 6, 7

[22] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete point flow networks for efficient point cloud generation. In *European Conference on Computer Vision*, pages 694–710. Springer, 2020. 6, 7

[23] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. 2014. 6

[24] Yibo Li and Bailun Liu. Improved edge detection algorithm for canny operator. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 10, pages 1–5. IEEE, 2022. 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[26] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1, 3, 5, 6, 7

[27] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 8

[28] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. *arXiv preprint arXiv:2307.13537*, 2023. 2

[29] Yongwei Miao, Feixia Hu, Xudong Zhang, Jiazhou Chen, and Renato Pajarola. Symmsketch: Creating symmetric 3d free-form shapes from 2d sketches. *Computational Visual Media*, 1(1):3–16, 2015. 3

[30] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3, 6, 7

[31] Luke Olsen, Faramarz F Samavati, Mario Costa Sousa, and Joaquim A Jorge. Sketch-based modeling: A survey. *Computers & Graphics*, 33(1):85–103, 2009. 3

[32] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 2, 6

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3

[34] Janis Postels, Mengya Liu, Riccardo Spezialetti, Luc Van Gool, and Federico Tombari. Go with the flows: Mixtures of normalizing flows for point cloud generation and reconstruction. In *2021 International Conference on 3D Vision (3DV)*, pages 1249–1258. IEEE, 2021. 6

[35] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

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

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 1, 2

[39] 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, June 2022. 1, 2

[40] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017. 3, 4

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022. 2

[42] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 1, 3

[43] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 6

[44] Yuefan Shen, Changgeng Zhang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Deepsketchhair: Deep sketch-based 3d hair modeling. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3250–3263, 2021. 3

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2

[46] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 7

[47] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 8

[48] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019. 7

[49] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Com-*

puter Vision (ICCV), pages 5722–5731, October 2021. 3, 6

[50] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1883, 2015. 3

[51] 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)*, September 2018. 2, 6

[52] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 8

[53] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *arXiv preprint arXiv:2208.02812*, 2022. 8

[54] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 7

[55] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2, 7

[56] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, June 2021. 2

[57] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 6

[58] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020. 2, 6

[59] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[60] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 6, 7

[61] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. 8

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

[63] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[64] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

[65] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2model: View-aware 3d modeling from single free-hand sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6012–6021, 2021. 2, 3, 6

[66] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 1, 3, 6, 7