Rapid 3D Model Generation with Intuitive 3D Input

Tianrun Chen[1,4*] Chaotao Ding[2*] Shangzhan Zhang[1*]
Chunan Yu[2] Ying Zang[2†] Zejian Li[3†] Sida Peng[3] Lingyun Sun[1]
[1]College of Computer Science and Technology, Zhejiang University
[2]School of Information Engineering, Huzhou University
[3]School of Software Technology, Zhejiang University
[4]KOKONI3D, Moxin (Huzhou) Technology Co., LTD.
tianrun.chen@kokoni3d.com {zhang3z, zejianlee, sunly}@zju.edu.cn
{2021388117, 2022388238}@stu.zjhu.edu.cn 02750@zjhu.edu.cn

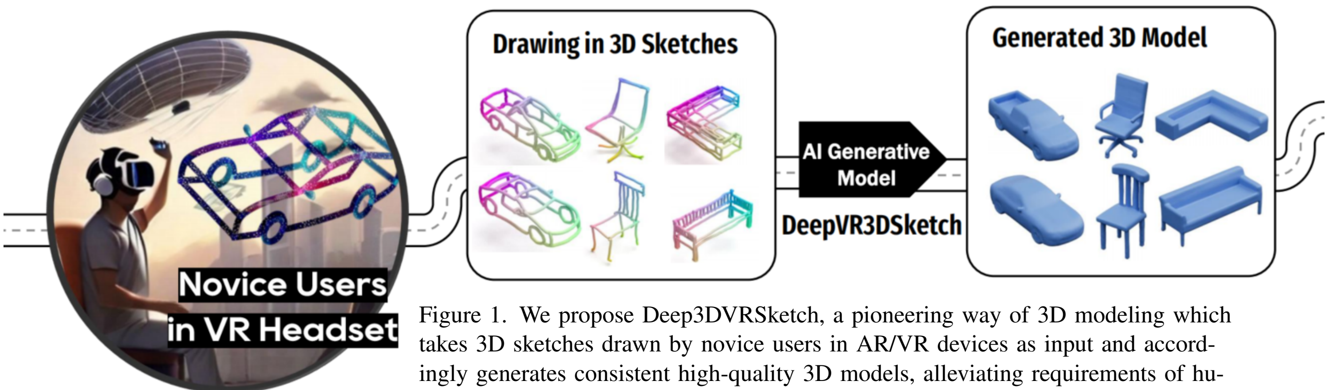

Figure 1. We propose Deep3DVRSketch, a pioneering way of 3D modeling which takes 3D sketches drawn by novice users in AR/VR devices as input and accordingly generates consistent high-quality 3D models, alleviating requirements of human skill and effort in traditional modeling.

Abstract

With the emergence of AR/VR, 3D models are in tremendous demand. However, conventional 3D modeling with Computer-Aided Design software requires much expertise and is difficult for novice users. We find that AR/VR devices, in addition to serving as effective display mediums, can offer a promising potential as an intuitive 3D model creation tool, especially with the assistance of AI generative models. Here, we propose Deep3DVRSketch, the first 3D model generation network that inputs 3D VR sketches from novice users and generates highly consistent 3D models in multiple categories within seconds, irrespective of the users' drawing abilities. We also contribute KO3D+, the largest 3D sketch-shape dataset. Our method pre-trains a conditional diffusion model on quality 3D data, then fine-tunes an encoder to map 3D sketches onto the generator's manifold using an adaptive curriculum strategy for limited ground truths. In our experiment, our approach achieves state-of-the-art performance in both model quality and fidelity with real-world input from novice users, and users can even draw and obtain very detailed geometric structures. In our user study, users were able to complete the 3D modeling tasks over 10 times faster using our approach compared to conventional CAD software tools. We believe that our Deep3DVRSketch and KO3D+ dataset can offer a promising solution for future 3D modeling in metaverse era. Check the project page at http://research.kokoni3d.com/Deep3DVRSketch.

1. Introduction

Today, a surge in demand for versatile and customizable 3D content has been catalyzed by the emerging trend of AR/VR [6, 55]. The virtual landscapes of the metaverse are calling out for creators with visions of immersive experiences waiting to be actualized. Yet the traditional tools of 3D modeling with computer-aided design (CAD) software pose barriers to crafting such imaginative domains. Existing widely-used CAD platforms demand substantial technical proficiency, including both command knowledge to understand numerous software functions and strategic knowledge to decompose designs into sequential modeling commands

*Equal contribution
†Corresponding Author

[3, 15]. This combination of required expertise and the labor-intensive manual process poses challenges for rapid iteration and scalable 3D content production in the metaverse era [18]. *Is there an alternative approach to obtain 3D models that can achieve rapid creation, customization based on user intent, and ease of use even for beginners?*

This paper provides an exploratory solution to this problem. We find that AR/VR devices, beyond acting as effective display platforms, hold significant potential for facilitating user-friendly and intuitive 3D model creation, especially with the assistance of recent advances in AI generative models. Past experiences have demonstrated that even beginners can effortlessly sketch 3D paths in the air using AR/VR 3D sketching tools [2, 16, 27, 31, 60]. We utilize these user-drawn, rudimentary 3D sketches in conjunction with our carefully designed AI network for 3D model generation to infer complete 3D geometry that matches the user's intention within seconds (Fig.1). By leveraging AI generative models after the user completes a drawing, the 3D model creation process can be greatly simplified, eliminating the need for skillful line drawings and laborious plane-filling. This enables even novices to produce high-quality 3D models from basic and rough sketches, while preserving the users' intended design.

Although promising, 3D model generation with 3D VR sketches is still a challenging task. The first challenge is the ambiguous connection between the 3D sketch and the corresponding 3D model, as novice users often struggle to produce accurate line drawings. Consequently, deterministic regression or supervision to minimize the sketch-model distance, as in prior works [40, 63], reflects the user's sketching skills (or the lack thereof) in the final 3D model, which is not friendly to novice users with a desire for high-quality models. The second challenge is the limited data availability. As of now, the only dataset available in 3D sketches and the corresponding shapes is from Luo et al. [39], which consists of merely one thousand samples and only one "chair" category. This volume of data is clearly insufficient for training a robust and generalized generative model.

Facing these challenges, we propose a novel framework, Deep3DVRSketch, for creating high-quality 3D shapes from input 3D sketches, regardless of the user's drawing skills. Instead of deterministic regression, our approach formulates this task as a conditional generation problem and designs a three-stage decoupled-generator training strategy to fully use limited training samples. Specifically, we first pre-train a conditional diffusion model on 3D object datasets with 3D shapes and rendered images, which have a large amount of data thanks to the fast development of 3D content creation. Then, the diffusion model that is capable of generating high-quality shapes is fixed, and an encoder is trained to map the 3D sketch onto feature vectors within a shared latent space with the pre-trained model, which interact with the intermediate feature maps of the diffusion model, guiding it to generate corresponding 3D shapes. Finally, joint fine-tuning is performed for the diffusion model and the encoder to improve alignment between 3D sketches and shape generation. During the training in 3D sketch mapping, we found the networks falter in generalizing across wide-ranging sketching styles and geometries with limited samples, leading to occasional failures in accurately reconstructing intricate local details. We thus propose an adaptive curriculum learning strategy to better use the limited data to learn diverse and complex shapes. To further address the issue of limited data, we also introduce a new dataset, KO3D+, which comprises thousands of sketch-model pairs drawn by humans across seven categories. This dataset provides a more expansive and diverse resource for the new field of 3D modeling from 3D VR Sketches for the academic community.

Our method's effectiveness is validated in comprehensive experiments. DeepVR3DSketch surpasses existing benchmarks in terms of model quality and fidelity, even with unseen inputs from novice users. Moreover, we demonstrate that even some very detailed structures can be generated based on users' detailed 3D drawings. In our user study, participants expressed higher satisfaction levels with the generated models. Users can use our approach to perform 3D modeling more than 10 times faster than conventional CAD based approaches. We believe that our Deep3DVRSketch can serve as a promising solution for future 3D modeling in the impending metaverse era.

2. Related Work

2.1. 3D Sketching in AR/VR

3D sketching tools have been developed with the emergence of AR/VR technologies. An early 3D sketching system in AR/VR was Holosketch [16], which enabled the creation of primitive shapes, freeform tubes, and 3D wireframes. Later work expanded the possibilities as hardware advanced [2, 27, 60]. Now, there are commercial tools like Tilt Brush, GravitySketch, and Quill. However, these systems rely heavily on manual operations that are labor-intensive and time-consuming. Recent solutions, such as the one proposed by Yu et al. [63], transform unstructured 3D sketches into smooth surfaces via optimization, but still rely on precise line drawings.

In light of this, to make possible 3D modeling for novice users with 3D sketches, in this paper, we propose to incorporate AI generative models to produce 3D shapes. The closest work to us is from Luo et al [40], which investigate shape prototyping and exploration with generative models based on normalizing flow and optimize the network based on the minimizing the distance between shapes and 3D sketches. However, as discussed earlier, matching sketches precisely leads to poor quality due to drawing imprecision by novice users. Their method is also limited to a single shape category per trained model. In contrast, our proposed generative approach can produce high-quality 3D shapes regardless of users' drawing skills. Moreover, our model supports multi-category shape generation, better suited for real-world applications.

2.2. 3D Model Generation with Generative Models

3D shape generation has seen substantial progress in recent years. A variety of generative models have been explored in the field of 3D generation, including Generative Adversarial Networks (GANs) [1, 12, 58, 59, 67], Variational Autoencoders (VAEs) [14, 45], autoregressive models [42, 57, 61], normalizing flows [49], and more recently diffusion models [13, 29, 43, 46, 71]. In this work, We opt to build upon our network based on diffusion models, which have achieved state-of-the-art results in 3D shape generation tasks and can produce high visual quality shapes with fine details.

We note there is also a significant advancement in image-conditioned and text-conditioned 3D model generation networks recently [11, 20, 33–36, 47, 49, 51–53]. Unlike previous works, here we choose a very new modality of input – the 3D VR sketch. For AI generative models, using 3D VR sketch as the input to get 3D models also has few advantages compared to other inputs. Image inputs don't allow for unrestricted, start-from-scratch 3D modeling; text is much less expressive or precise than a freehand sketch in conveying spatial or geometric information. While there are also some existing 3D model generation approaches focusing on 2D sketch input [4, 8–10, 21, 23, 32, 38, 44, 54, 64–66, 69], the 2D sketch is ambiguous and abstract, which arises from inherent limitations in 2D sketches, including missing information due to occlusion and limited viewpoints. Sketching in 3D, in contrast, provides the capacity to communicate more comprehensive information, such as the complex internal features of objects like car seats. Hence, it opens the potential for achieving highly detailed and superior quality 3D modeling as demonstrated in this paper.

3. Method

3.1. 3D Sketch Acquisition and the KO3D+ dataset

The scarcity of existing datasets has hindered the progression of sketch-to-3D research, so we build a new dataset, KO3D+, by recruiting participants to draw 3D sketches in VR. The protocol of data acquisition is akin to that used by Luo et al. [40], in which participants were asked to draw over existing 3D models. The 3D models utilized were man-made high-quality 3D shapes sourced from ShapeNet dataset. We selected seven categories from ShapeNet: car, sofa, airplane, bench, display, watercraft, and table. Each category contains 600 3D sketches along with their corresponding 3D shapes, making this the most extensive 3D sketch dataset currently available. For an in-depth description of the dataset, including sample illustrations, please refer to the supplementary materials.

3.2. The Deep3DVRSketch Network

The Deep3DVRSketch aims to translate the provided 3D sketches to high-quality 3D models. The main component of Deep3DVRSketch is a diffusion-based generation framework designed to produce high-quality 3D shapes. In Section 3.2.1, we introduce the diffusion network structure. In 3.2.2, we show how we train the network. Specifically, our diffusion-based generation framework is trained in three stages, namely Generative Pre-Training, 3D Sketch Mapping, and Joint Fine-Tuning. Finally, a curriculum learning strategy is introduced to better utilize limited 3D sketch data and produce higher quality results (3.2.3).

3.2.1 Preliminary: Conditional 3D Diffusion Model.

DeepVR3DSketch utilize a conditional diffusion model as the 3D shape generator, which has demonstrated success in obtaining high-quality and diverse 3D models compared to other methods such as using normalizing flows [17, 40].

Principle of Diffusion Models. A diffusion model is trained to sample from a target distribution by reversing a sequential noise diffusion process. Given a sample represented as z, we generate z_t for each t in the range from 1 to T by progressively introducing Gaussian noise, adhering to a predetermined variance schedule. Subsequently, a time-conditional 3D UNet, represented as ϵ_θ, is utilized for denoising. Finally, the UNet generate new 3D shapes by denoising from a new Gaussian noise.

3D Shape Representation. Here, we represent a 3D shape as a discrete 3D volume – a Signed Distance Function (SDF) volume. This SDF volume calculates the signed distance from the center of each grid cell to the nearest shape surface. The mesh, which is also the zero isosurface, can be derived from the the grids using the Marching Cube algorithm. The SDF volume can be converted into a discrete occupancy volume where each grid cell holds a binary occupancy based on whether the absolute value of its SDF is beneath a predefined threshold.

Coarse-to-Fine Diffusion. High-fidelity 3D shape representation requires modeling fine details using high-resolution discrete signed distance fields (SDFs). However, fully generating dense SDF grids incurs prohibitive computational and memory costs due to cubic complexity. To avoid a huge computational burden while still maintaining a high-quality model generation, we follow [68] and use a two-stage diffusion framework utilizing a self-conditioning continuous diffusion model. In specific, the first stage generates a low-resolution 3D occupancy volume $F \in \mathbb{R}^{n\times n\times n\times 1}$ to provide a preliminary approximation of the 3D shape. Subsequently, the second stage constructs a high-resolution sparse volume $S \in \mathbb{R}^{N\times N\times N\times 4}$.

In the two stages, we use a 5-level UNet in the first stage,

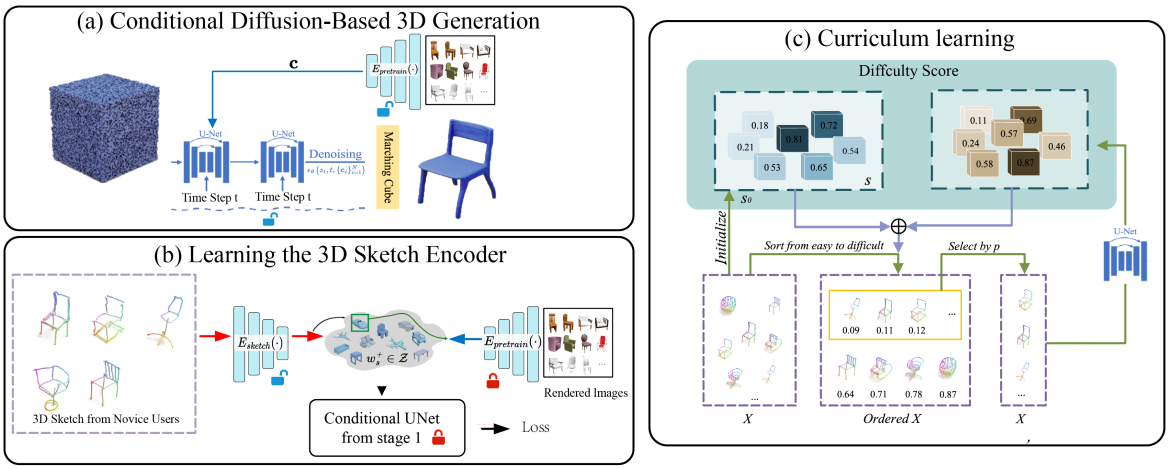

Figure 2. Key Designs of Deep3DVRSketch. (a) Conditional diffusion model pre-training with rendered images. (b) Sketch encoder fine-tuning to map sketches into diffusion manifold. (c) Curriculum learning exploits limited sketch-shape pairs.

and a 4-level UNet in the second stage, where an octree-based convolutional neural network handles the SDF data in sparse voxel format. For details of the network configuration, please refer to Supplementary Material. Both the UNets are trained with the denoising loss [26]:

$$L(\theta) = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \{\mathbf{c}_i\}_{i=1}^N\right)\right\|^2\right] \quad (1)$$

in which $\mathcal{N}(0, 1)$ denotes Gaussian distribution, $\{\mathbf{c}_i\}_{i=1}^N$ denote the condition applied to the generation process.

Learning the Conditional Distribution. Our 3D generation is a conditional generation setting in which the conditions $\{\mathbf{c}_i\}_{i=1}^N$ are injected into the diffusion-based generation process to accept user input. The conditioning signal is applied by using task-specific encoders to transform the conditioning signal $\mathbf{c}$ (e.g. images, sketches) into a 1024-dimensional latent code $\mathbf{l}$. Subsequently, multi-head cross-attention is used to infuse $\mathbf{l}$ into the UNet. The diffusion model may produce samples with limited diversity given the conditional input. To alleviate the issue, we adopt classifier-free guidance [25]. For more information, please refer to the Supplementary Material.

3.2.2 The Training of Deep3DVRSketch

In the following part, we show our multi-stage training strategy to make possible high-quality shape generation from 3D Sketches. We anticipate the diffusion model to create 3D shapes from a latent space, which will later be utilized for the downstream task of transforming 3D sketches into 3D models. In this context, the downstream fine-tuning is focused solely on comprehending the specifics of this task, while the intricate process of 3D shape generation leverages the pre-trained knowledge from the model. Specifically, in the Generative Pre-Training stage, we pre-train a conditional 3D diffusion generator that can produce high-quality shapes based on image conditioning. The images go through an image encoder $E_{pretrain}(\cdot)$ and derive a manifold $\mathcal{Z}$ used to condition the diffusion model to produce plausible 3D shapes.

Next, in 3D Sketch Mapping stage, we train a sketch encoder $E_{sketch}(\cdot)$ to map the input 3D sketch s into a latent code $w_s^+ = \mathcal{E}_s(s) \in \mathbb{R}^{1024}$ lying in the manifold $\mathcal{Z}$ while leaving the pretrained diffusion generator intact. The outputs of $E_{sketch}(\cdot)$ at this stage will be closer to $\mathcal{Z}$, but still cannot reach the perfect alignment. Therefore, we introduce a Joint Fine-Tuning stage, we fine-tune both the encoder $E(\cdot)$ and the diffusion generator altogether to obtain much-improved alignment in $E_{sketch}(\cdot)$ and $E_{pretrain}(\cdot)$ in the pre-training stage. Such a phased training approach is beneficial in maximizing the use of pre-trained knowledge and has been proven to be a key factor in significantly enhancing the final output quality.

Stage 1: Generative Pre-Training. The aim of this phase is to train a diffusion model to generate 3D shapes from a latent space that will subsequently be employed for the specific downstream task of converting 3D sketches into 3D models. To do so, we present the network with a vast collection of high-quality 3D shapes and condition the diffusion process with rendered images. These images are encoded by a pre-training encoder, denoted as $E_{pretrain}(\cdot)$.

Taking inspiration from previous works of visual-linguistic pretraining and their remarkable transferability [48, 70], we map the image condition to a CLIP latent space with a fixed pre-trained CLIP encoder. The diffusion model is conditioned by embedding from the CLIP latent space $\mathcal{Z}$ via cross-attention. By doing so, we obtain a diffusion model that can generate high-quality 3D shapes with global conditions from the CLIP latent space $\mathcal{Z}$.

Stage 2: 3D Sketch Mapping. Then, we map the 3D sketch input to the CLIP latent space $\mathcal{Z}$ and keep the pre-trained diffusion model in Stage 1 frozen. The 3D sketch is represented by point clouds, so we designed a transformer point encoder to encode the 3D sketch s into a latent code $w_s^+ = \mathcal{E}_s(s) \in \mathbb{R}^{1024}$, which is the same dimension as the CLIP feature in the previous stage.

We note that the CLIP latent space is a highly abstract domain. Mapping highly sparse and abstract 3D sketches represented by point clouds onto this space is a challenging task, especially when the quantity of 3D sketch data is limited. Addressing this challenge, we introduce additional prior knowledge about 3D space. Specifically, we utilize a pre-trained point encoder Uni3D [70], which is aligned with the CLIP latent space, to serve as the 3D sketch encoder. Structurally equivalent to the vanilla transformer of ViT, this encoder has already learned to represent a wealth of point cloud features from a substantial dataset consisting of millions of 3D shapes, corresponding images, and text entries under a multi-modal alignment learning objective. In the ablation study, we show this pre-training is critical for the network performance.

Stage 3: Joint Fine-Tuning. We observe that exclusively adjusting the 3D sketch $E_{sketch}(\cdot)$ is insufficient to ensure the optimal alignment of the shape generation process with the sketch input. Therefore, inspired by previous practice in image diffusion-based generation [56], we concurrently fine-tune both the 3D sketch $E_{sketch}(\cdot)$ and the diffusion model to ensure significantly enhanced spatial semantic alignment. This approach proves beneficial in maximizing the utilization of pre-trained knowledge and is also crucial for achieving improved quality.

3.2.3 Adaptive Curriculum Learning

As we mentioned in the introduction, the data scarcity and the complexity of 3D sketches are two main challenges in our task. In our framework, we find that with only limited training data of 3D sketches and 3D model pairs, networks struggle to generalize across large variations in sketching styles and geometries when mapping abstract 3D sketches to a latent space to condition a generative model. In our experiment, we can observe that the network occasionally fails to accurately reconstruct intricate local details. In these regions, the network struggles with precise parameterization due to less smooth areas in the implicit function. Even small errors can produce incorrect signs, leading to inaccurate surface reconstructions.

Drawing inspiration from curriculum learning, we intend to tackle these challenges by emulating the human learning process in sketching. Just as beginners in sketching start with simple, flexible shapes and gradually progress towards more complex and difficult ones in later training, we aim to incorporate this learning strategy into our framework.

Sample Difficulty Score. In curriculum learning, the careful selection and sequencing of samples, from simple to complex, is crucial for effective and incremental skill development. The selection is based on a difficulty score estimated on samples. Inspired by Curriculum DeepSDF [19], we consider points with incorrect estimates as hard samples, points with correct estimates as easy samples, and points between 0 and the ground truth as semi-hard samples. We use the following difficulty score.

$$s_{cur} = 1 + \lambda \text{sgn}(y)\text{sgn}(\hat{y} - y) \tag{2}$$

Here, y is the ground truth SDF value, $\hat{y}$ is the predicted SDF value, and $0 \leq \lambda \leq 1$ control the importance of hard and semi-hard samples. sgn(v)=1 if v>=0 and sgn(v)=-1 if v<0.

Adaptive Curriculum. Unlike traditional curriculum learning which designs the curriculum manually, we use an adaptive curriculum strategy. [30] This is, to the best of our knowledge, the first attempt at applying an adaptive curriculum in a conditional 3D generation task.

Specifically, we first leverage the pre-trained network to acquire the initial difficulty score and sort the initial dataset $\mathcal{X}$ in ascending order based on the current difficulty score s. We subsequently use the pacing function $p(\cdot)$ to form the sample pool $\mathcal{X}'$, from which we draw a mini-batch $\mathcal{B}' = [\mathcal{B}'_1, ..., \mathcal{B}'_M]$ to train the target network. The pacing function $p(\cdot)$ is a monotonically increasing function that determines the speed at which we learn from simpler to more complex samples. Finally, we update the difficulty score s at the end of the forward propagation and compute a new sample pool $\mathcal{X}'$.

The difficulty score is also adaptable to varying training duration. The difficulty score for the $(k+1)^{th}$ position can be expressed as:

$$s_{k+1} = (1 - \alpha)s_k + \alpha s_{cur} \tag{3}$$

where $k = \lfloor m/inv \rfloor$ where m denotes the m^{th} mini-batch. inv is controls the frequency of difficulty score updates, and α controls the speed of difficulty score updates.

Pacing Function. To manage the pace of sample learning, we require a monotonically increasing pacing function, $p(\cdot)$, to constrain the size of the sample pool $\mathcal{X}'$. The function can be expressed as:

$$p(i) = n \times \min(1, p_0 \times q^{\lfloor i/r_0 \rfloor}) \tag{4}$$

where n represents the number of samples. p_0 is the sample proportion at the initial step. q controls the speed of sample proportion growth. r_0 controls the frequency of sample proportion growth, and i is the current step. For more information, please refer to the Supplementary Material.

4. Experiment

4.1. Dataset

Our method first pre-trains a generative diffusion model on a large 3D shape dataset (ShapeNetCore-v2) for high-quality synthesis (Stage 1). The model is then fine-tuned on aligned and normalized sketch-shape pairs (Stages 2-3). The availability of 3D sketch and 3D model pair datasets is limited. Currently, there is only one dataset available by Luo et al [39]. This dataset consists of 1,005 sketch shape pairs in the chair class. We adopt their dataset and use their predefined split, 803 for training and 202 for testing. We use 5,721 samples of chairs from the ShapeNet dataset in the first stage pre-training. In addition to the Luo et al. dataset, we introduce our own dataset, KO3D+, which encompasses 7 categories from ShapeNet with the defined test-train split. We use 12,970 samples of chairs the corresponding categories in the ShapeNet dataset in the first stage pre-training.

4.2. Implementation Details

During the first pre-training stage, we trained the first UNet using the Adam optimizer [28] with a fixed learning rate of 2×10^{-4} for 800 epochs. For the subsequent UNet, we utilized the AdamW optimizer [37] with a fixed learning rate of 1e-4 for 500 epochs. In the second Sketch Mapping stage, we trained the sketch encoder for 300 epochs with a learning rate of 2e-4 and Adam optimizer. Finally, in the third Joint Fine-Tuning stage, we tuned the diffusion model and the sketch encoder jointly for 300 epochs with a learning rate of 2e-4 with Adam optimizer. The training process was conducted using 8 NVIDIA A100 graphics cards.

4.3. Evaluation Metrics

We quantitatively evaluate the generated 3D shapes in terms of the fidelity and quality. Following prior works in 3D generative models [22, 50], we choose the bidirectional Chamfer distance (CD) as the similarity metric between two 3D shapes to measure the fidelity of generated result and the Frèchet Inception Distance (FID) [24] to measure the visual quality of generated result. For more details about the evaluation metrics, please refer to the Supplementary Material.

4.4. Qualitative and Quantitative Evaluation

Our experiment is conducted in two settings: 1) Training a single-category 3D generation network using the dataset from Luo et al. [40] and the proposed KO3D+ dataset. 2) Training a multi-category 3D generation network using the proposed KO3D+ dataset. We find our method significantly outperforms existing 3D and 2D baselines in model fidelity and quality.

Single-Category 3D Generation. We first use the dataset from Luo et al. [40] to train our network and compare it with Luo et al. [40]. Our approach outperforms the existing methods in both model fidelity and quality as show in Tab. 1. This improved performance extends to the data in the KO3D+ dataset as well. Specifically, we selected the "car" category and trained our approach alongside Luo et al.'s method, and our approach consistently outperforms theirs. Importantly, our approach excels at reconstructing intricate details, such as the car's mirrors and interior seats (in Fig. 3). This not only demonstrate the strong capability of our algorithm, but also showcases the inherent advantage of utilizing 3D sketches, as these structures can be easily represented and conveyed through 3D sketches, whereas they pose significant challenges when using 2D sketches. The ability of our approach to accurately capture and reconstruct these fine-grained features further highlights the benefits of employing 3D sketches in the modeling process.

Category	Method	FID ↓	CD ↓
Chair [39]	Luo et al. [40]	11.5313	0.0305
	Ours	**8.7701**	**0.0220**
Car	Luo et al. [40]	20.7812	0.0303
	Ours	**7.9213**	**0.0047**

Table 1. Quantitative Result for Single-Category 3D Generation

Method	FID ↓	CD ↓
Luo et al. [40]	30.5313	0.0455
Ours	**15.5397**	**0.0304**

Table 2. Quantitative Result for Multi-Category 3D Generation

Method	FID ↓	CD ↓
Luo et al. [40]	11.5313	0.0305
Ours	**8.7701**	**0.0220**
LAS-Diffusion [68]	11.2195	0.0501
Deep3DSketch [9]	108.7884	0.1362
Deep3DSketch+ [8]	163.3932	0.1295

Table 3. Quantitative Comparison with 2D Sketch-Based Methods

Multi-Category 3D Generation. Next, we evaluate the cross-category performance of our approach by utilizing all the data from the KO3D+ dataset. As depicted in Fig. 4, our Deep3DVRSketch successfully generates high-quality 3D models across multiple categories. In contrast, the existing approaches fail to produce meaningful results (Tab. 2) and we have witnessed a mode collapse in existing methods. (details in Supplementary Material). This highlights the superior capability of our approach in handling diverse categories and consistently generating accurate and visually pleasing 3D models, making it a promising solution with broader applicability.

Comparison with 2D Sketch-Based Approaches. We also compare our 3D sketch-based generation approach with a solely 2D sketch-based method. The projection of a 3D sketch onto a 2D plane serves as a 2D sketch. We evaluate our approach against representative methods for generating 3D models from 2D sketches, namely LAS-Diffusion [68], Deep3DSketch [9], and subsequent work, Deep3DSketch+ [8]. Experimental results show Deep3DVRSketch with 3D sketch as the input produces more plausible shapes with better 3D awareness and local geometric control compared to 2D sketch-input approaches, as shown in Table 3 and Fig. 5 (see the arms in the first row and the legs and seat pad in the second row.) More examples in Supplementary Material.

4.5. User Study

3D Model Quality and Fidelity. To further validate the effectiveness of our sketch-to-model algorithm, we conducted

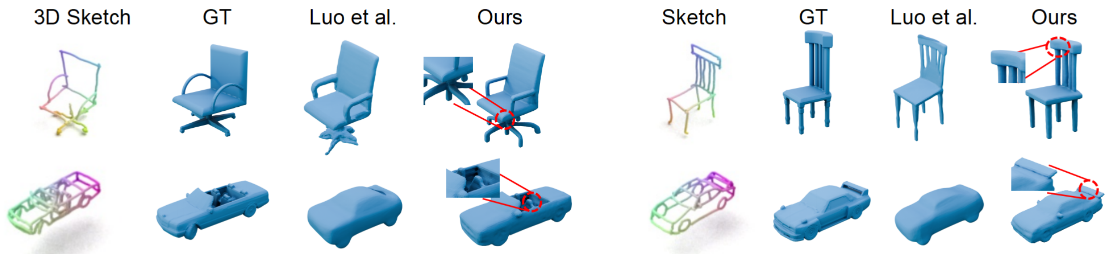

Figure 3. The Visualization of Performance Comparison. Our Deep3DVRSketch generates highly detailed, high-quality 3D shapes. Even small geometric features like car seats and spoilers are accurately reconstructed, as evidenced in the magnified regions.

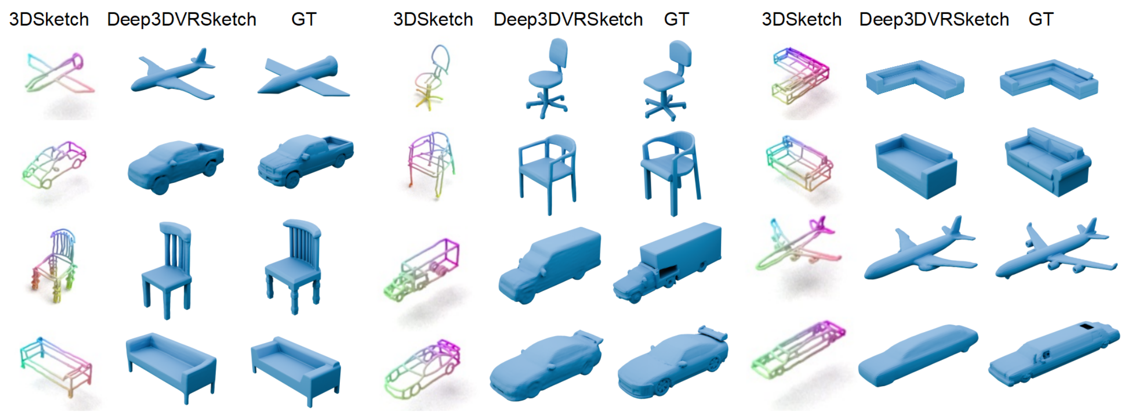

Figure 4. The Visualization of Multi-Category Generation Results from Our Deep3DVRSketch. In our Deep3DVRSketch, one network model is capable of generating high-fidelity and high-quality shapes at multiple categories from 3D sketch draw by novice users. It is not necessary to train multiple models for different categories.

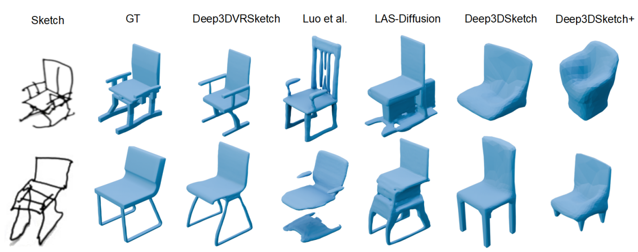

Figure 5. Comparison with 2D Sketch-Based Approaches. Our Deep3DVRSketch exhibits superior shape quality and fidelity compared to 2D sketch-based approaches.

a user study to evaluate the fidelity and quality of the generated 3D models. We adopted the widely used 5-point mean opinion score (MOS) metric, as in prior works [5, 7, 41, 62]. Specifically, users were asked to rate the following two factors on a scale of 1 to 5: Q1) How well does the output 3D model match the input sketch? (Fidelity); Q2) What is your opinion on the overall quality of the output 3D model? (Quality)

We recruited 12 designers from a 3D printing company who are familiar with 3D modeling and presented 48 results generated by our algorithm. Prior to the study, we explained the definitions of fidelity and quality. The average ratings are reported in Table 4. Compared to existing methods, our algorithm achieved higher user ratings, validating its effectiveness of our proposed approach.

Methods	(Q1): Fidelity	(Q2): Quality
Luo et al. [40]	2.02 ± 0.84	1.98 ± 0.86
Ours	**4.06 ± 0.76**	**4.03 ± 0.81**

Table 4. Mean Opinion Scores (1-5) from User Study

Deep3DVRSketch Make Rapid 3D Modeling Possible. To demonstrate the practical utility of our proposed method, we conducted a user study with 3 professional 3D designers from a 3D printing company. The experts were asked to model 9 reference shapes using both our VR sketch-based approach and their familiar CAD software tools (ZBrush). We recorded the average modeling time for each method in Tab. 5. With our approach, the designers could complete the modeling over 10 times faster, including sketching and network inference time (average 7.3 seconds on A100 GPU). The user study validates our approach as a practical tool for accelerating 3D design workflows. For more details, please refer to the Supplemantary Material.

4.6. Ablation Study

We perform extensive ablation study to validate the design choices in our Deep3DVRSketch network, as quantitative

Method	Time (s) ↓
Conventional CAD Software	869.2 ± 410.2
Ours	**89.4 ± 29.6**

Table 5. The Average Time Comparison of 3D Modeling

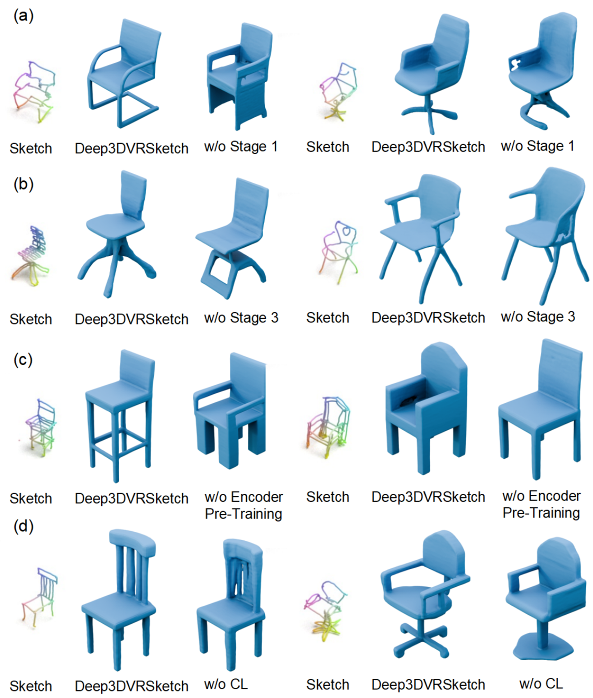

Figure 6. Qualitative Evaluation for Ablation Studies

results shown in Tab. 6. Experiments in this section is performed with the datasets from Luo et al. [39]

Benefits of Stage 1: Generative Pre-Training. Pre-training the diffusion model on large 3D datasets is crucial for enabling high-quality shape synthesis. Removing this stage and directly training the sketch-conditional diffusion model, with identical network architecture, causes significant performance drops. As evidenced in Fig. 6 (a) and in Tab. 6, the model fails to learn to generate plausible shapes without pretraining. This highlights the importance of leveraging abundant 3D data to first cultivate strong generative priors, before adapting the model to the sketch domain where labeled data is scarce.

Method	CD ↓	FID ↓
Deep3DVRSketch	0.0220	8.7701
w/o Stage 1: Generative Pre-Training	0.0305	25.0112
w/o Stage 3: Joint Fine-Tuning	0.0256	9.9729
w/o Encoder Pre-Training at Stage 2	0.0405	17.3267
w/o Curriculum Learning	0.0237	9.5470

Table 6. Quantitative Evaluation of Ablation Study

Benefits of Stage 3: Joint Fine-Tuning. The Joint Fine-Tuning stage is critical for aligning the sketch encodings with the generative model to produce shapes that match the input sketches. Without this fine-tuning, there is a performance drop as evident in Tab. 6. Interestingly, we find that without fine-tuning, the visual quality of generated shapes remains high, but they fail to correspond to the input sketches, as evidenced in Fig. 6 (b) for the legs of the chair and the arms of the chair.

Benefits of Encoder Pre-Training at Stage 2. To map 3D sketches into the abstract CLIP latent space with limited data, we leverage a pretrained point cloud encoder Uni3D [70] to provide inductive biases about 3D structure. In ablation studies, we use random initialization of this encoder, which can be found significantly harms performance. Fig. 6 (c) demonstrates that in the absence of a pre-trained 3D sketch encoder, the generated shapes exhibit high visual quality. However, they deviate from the intended correspondence with the input sketches, indicating a failure in successfully performing 3D sketch mapping.

Benefits of Curriculum Learning. The curriculum learning approach is employed to address the issue of limited data availability and helps mitigate the challenge of generalizing across large variations in sketching styles and geometries. By removing the curriculum learning, we observe that the network occasionally fails to accurately predict complex shape structures, as evidenced in Figure 6 (d), like the back of the chair and the wheel & arms of the chair.

5. Conclusion

This paper introduces Deep3DVRSketch, a 3D model generation network designed to generate high-fidelity and consistent 3D models in multiple categories from drawings in 3D VR space by novice users. Our approach combines VR 3D sketching with AI generative models to simplify the 3D model creation process. We formulate the task as a conditional generation problem and employ a three-stage training strategy, along with adaptive curriculum learning, to address the challenges posed by data scarcity and the complexity of sketches. To facilitate research in this field, we also introduce the KO3D+ dataset, which is currently the largest 3D sketching dataset containing ground truth 3D shapes. Extensive experiments demonstrate the effectiveness of our approach in terms of model quality, fidelity, and user satisfaction. We believe that Deep3DVRSketch, along with the proposed KO3D+ dataset, opens new avenues for research in 3D modeling and holds promise for the future of 3D modeling in the metaverse era.

Acknowledgments

This work was supported by National Key R&D Program of China (2022YFB3303301), NSFC (No. 62172364), Information Technology Center, and State Key Lab of CAD&CG, Zhejiang University. The research is also partially funded by Moxin (Huzhou) Technology Co., LTD.

References

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

[2] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. Symbiosissketch: Combining 2d & 3d sketching for designing detailed 3d objects in situ. In *CHI 2018*, pages 1–15, 2018. 2

[3] Suresh K Bhavnani, Bonnie E John, and Ulrich Flemming. The strategic use of cad: An empirically inspired, theory-based course. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 183–190, 1999. 2

[4] Alexandra Bonnici, Alican Akman, Gabriel Calleja, Kenneth P Camilleri, Patrick Fehling, Alfredo Ferreira, Florian Hermuth, Johann Habakuk Israel, Tom Landwehr, Juncheng Liu, et al. Sketch-based interaction and modeling: where do we stand? *AI EDAM*, 33(4):370–388, 2019. 3

[5] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021. 7

[6] Shu-Ching Chen. Multimedia research toward the metaverse. *IEEE MultiMedia*, 29(1):125–127, 2022. 1

[7] Tianrun Chen, Chaotao Ding, Lanyun Zhu, Ying Zang, Yiyi Liao, Zejian Li, and Lingyun Sun. Reality3dsketch: Rapid 3d modeling of objects from single freehand sketches. *arXiv preprint arXiv:2310.18148*, 2023. 7

[8] Tianrun Chen, Chenglong Fu, Ying Zang, Lanyun Zhu, Jia Zhang, Papa Mao, and Lingyun Sun. Deep3dsketch+: Rapid 3d modeling from single free-hand sketches. In *International Conference on Multimedia Modeling*, pages 16–28. Springer, 2023. 3, 6

[9] Tianrun Chen, Chenglong Fu, Lanyun Zhu, Papa Mao, Jia Zhang, Ying Zang, and Lingyun Sun. Deep3dsketch: 3d modeling from free-hand sketches with view-and structural-aware adversarial training. In *ICASSP*, pages 1–5. IEEE, 2023. 6

[10] Tianrun Chen, Runlong Cao, Zejian Li, Ying Zang, and Lingyun Sun. Deep3dsketch-im: rapid high-fidelity ai 3d model generation by single freehand sketches. *Frontiers of Information Technology & Electronic Engineering*, 25(1): 149–159, 2024. 3

[11] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 3

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

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

[14] Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. Cross-modal 3d shape generation and manipulation. In *European Conference on Computer Vision*, pages 303–321. Springer, 2022. 3

[15] Ivan Chester. Teaching for cad expertise. *International Journal of Technology and Design Education*, 17(1):23–35, 2007. 2

[16] Michael F Deering. Holosketch: a virtual reality sketching/animation tool. *TOCHI*, 2(3):220–238, 1995. 2

[17] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 3

[18] Trien V Do and Jong-Weon Lee. 3darmodeler: a 3d modeling system in augmented reality environment. *International Journal of Mathematical and Computational Sciences*, 4(3): 377–386, 2010. 2

[19] Yueqi Duan, Haidong Zhu, He Wang, Li Yi, Ram Nevatia, and Leonidas J Guibas. Curriculum deepsdf. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 51–67. Springer, 2020. 5

[20] Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. Shapecrafter: A recursive text-conditioned 3d shape generation model. *Advances in Neural Information Processing Systems*, 35:8882–8895, 2022. 3

[21] Chenjian Gao, Qian Yu, Lu Sheng, Yi-Zhe Song, and Dong Xu. Sketchsampler: Sketch-based 3d reconstruction via view-dependent depth sampling. In *European Conference on Computer Vision*, pages 464–479. Springer, 2022. 3

[22] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 6

[23] 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*, pages 13023–13032, 2021. 3

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[27] Daniel F Keefe, Daniel Acevedo Feliz, Tomer Moscovich, David H Laidlaw, and Joseph J LaViola Jr. Cavepainting: A fully immersive 3d artistic medium and interactive experience. In *Proceedings of the 2001 symposium on Interactive 3D graphics*, pages 85–93, 2001. 2

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

[29] Di Kong, Qiang Wang, and Yonggang Qi. A diffusion-refinement model for sketch-to-point modeling. In *Proceedings of the Asian Conference on Computer Vision*, pages 1522–1538, 2022. 3

[30] Yajing Kong, Liu Liu, Jun Wang, and Dacheng Tao. Adaptive curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5076, 2021. 5

[31] Kin Chung Kwan and Hongbo Fu. Mobi3dsketch: 3d sketching in mobile ar. In *CHI 2019*, pages 1–11, 2019. 2

[32] Changjian Li, Hao Pan, Yang Liu, Xin Tong, Alla Sheffer, and Wenping Wang. Robust flow-guided neural prediction for sketch-based freeform surface modeling. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 3

[33] Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*, 2023. 3

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

[35] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.

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

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[38] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. In *2017 International Conference on 3D Vision (3DV)*, pages 67–77. IEEE, 2017. 3

[39] Ling Luo, Yulia Gryaditskaya, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Fine-grained vr sketching: Dataset and insights. In *2021 International Conference on 3D Vision (3DV)*, pages 1003–1013. IEEE, 2021. 2, 6, 8

[40] Ling Luo, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song, and Yulia Gryaditskaya. 3d vr sketch guided 3d shape prototyping and exploration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9267–9276, 2023. 2, 3, 6, 7

[41] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 7

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

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

[44] 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

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

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

[47] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

[48] 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. 4

[49] 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. 3

[50] Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18339–18348, 2023. 6

[51] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[52] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023.

[53] Xi Tian, Yong-Liang Yang, and Qi Wu. Shapescaffolder: Structure-aware 3d shape generation from text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2715–2724, 2023. 3

[54] Jiayun Wang, Jierui Lin, Qian Yu, Runtao Liu, Yubei Chen, and Stella X Yu. 3d shape reconstruction from free-hand

sketches. In *European Conference on Computer Vision*, pages 184–202. Springer, 2022. 3

[55] Miao Wang, Xu-Quan Lyu, Yi-Jun Li, and Fang-Lue Zhang. Vr content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6(1):3–28, 2020. 1

[56] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 5

[57] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2021. 3

[58] 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. 3

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

[60] Pengfei Xu, Hongbo Fu, Youyi Zheng, Karan Singh, Hui Huang, and Chiew-Lan Tai. Model-guided 3d sketching. *TVCG*, 25(10):2927–2939, 2018. 2

[61] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 3

[62] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 7

[63] Emilie Yu, Rahul Arora, J Andreas Baerentzen, Karan Singh, and Adrien Bousseau. Piecewise-smooth surface fitting onto unstructured 3d sketches. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[64] Ying Zang, Chaotao Ding, Tianrun Chen, Papa Mao, and Wenjun Hu. Deep3dsketch+\+: High-fidelity 3d modeling from single free-hand sketches. *arXiv preprint arXiv:2310.18178*, 2023. 3

[65] Ying Zang, Chenglong Fu, Tianrun Chen, Yuanqi Hu, Qingshan Liu, and Wenjun Hu. Deep3dsketch+: Obtaining customized 3d model by single free-hand sketch through deep learning. *arXiv preprint arXiv:2310.18609*, 2023.

[66] 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. 3

[67] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, pages 52–63. Wiley Online Library, 2022. 3

[68] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *arXiv preprint arXiv:2305.04461*, 2023. 3, 6

[69] Yue Zhong, Yulia Gryaditskaya, Honggang Zhang, and Yi-Zhe Song. Deep sketch-based modeling: Tips and tricks. In *2020 International Conference on 3D Vision (3DV)*, pages 543–552. IEEE, 2020. 3

[70] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 4, 5, 8

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