Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Deep3DSketch-im: rapid high-fidelity AI 3D model generation by single freehand sketches*

Tianrun CHEN[1], Runlong CAO[3], Zejian LI[2], Ying ZANG‡[3], Lingyun SUN‡[1]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*School of Software Technology, Zhejiang University, Hangzhou 310027, China*

[3]*School of Information Engineering, Huzhou University, Huzhou 313000, China*

E-mail: tianrun.chen@zju.edu.cn; crl1567@163.com; zejianlee@zju.edu.cn; 02750@zjhu.edu.cn; sunly@zju.edu.cn

Received Apr. 30, 2023; Revision accepted Nov. 26, 2023; Crosschecked Jan. 15, 2024

Abstract: The rise of artificial intelligence generated content (AIGC) has been remarkable in the language and image fields, but artificial intelligence (AI) generated three-dimensional (3D) models are still under-explored due to their complex nature and lack of training data. The conventional approach of creating 3D content through computer-aided design (CAD) is labor-intensive and requires expertise, making it challenging for novice users. To address this issue, we propose a sketch-based 3D modeling approach, Deep3DSketch-im, which uses a single freehand sketch for modeling. This is a challenging task due to the sparsity and ambiguity. Deep3DSketch-im uses a novel data representation called the signed distance field (SDF) to improve the sketch-to-3D model process by incorporating an implicit continuous field instead of voxel or points, and a specially designed neural network that can capture point and local features. Extensive experiments are conducted to demonstrate the effectiveness of the approach, achieving state-of-the-art (SOTA) performance on both synthetic and real datasets. Additionally, users show more satisfaction with results generated by Deep3DSketch-im, as reported in a user study. We believe that Deep3DSketch-im has the potential to revolutionize the process of 3D modeling by providing an intuitive and easy-to-use solution for novice users.

Key words: Content creation; Sketch; Three-dimensional (3D) modeling; 3D reconstruction; Shape from X; Artificial intelligence (AI)

https://doi.org/10.1631/FITEE.2300314 **CLC number:** TP31

1 Introduction

The abilities of artificial intelligence (AI) content generation have made significant strides in recent years, with notable progress made in generating images and text (Zhou et al., 2023; Huang and Wang, 2024; Lei and Li, 2024). However, despite the increasing demand for three-dimensional (3D) models in various applications, most existing artificial intelligence generated content (AIGC) methods have been focused on language and two-dimensional (2D) image generation, resulting in a significant lack of progress in generating 3D content. Generating 3D content poses several significant challenges, such as the need for 3D spatial awareness and the difficulty in representing complex 3D shapes. Furthermore, the lack of readily available 3D content data for training models is a significant challenge in the development of AI-generated 3D content—for years, researchers have to use computer-aided design (CAD) based approaches to create 3D models. However, the

‡ Corresponding authors

* Project supported by the National Key R&D Program of China (No. 2022YFB3303301), the National Natural Science Foundation of China (Nos. 62006208, 62107035, and 62207024), and the Public Welfare Research Program of Huzhou Science and Technology Bureau, China (No. 2022GZ01)

ORCID: Tianrun CHEN, https://orcid.org/0000-0003-0177-0157; Ying ZANG, https://orcid.org/0000-0002-1361-1500; Lingyun SUN, https://orcid.org/0000-0002-5561-0493

complexity and steep learning curve of CAD software pose a significant challenge for novice users, as observed in Chester (2007). The time-consuming and labor-intensive nature of the CAD-based approach also limits its scalability and hinders its potential to democratize 3D modeling, as noted by Reddy and Rangadu (2018). Recently, software like Tinkercad offers better usability, but the strategic knowledge (decomposing steps of constructing models) is still a challenge for users to learn (Mahapatra et al., 2019). As a result, there is a growing need for more intuitive and user-friendly 3D modeling methods that can better serve the needs of a broader range of users.

To address the challenges of 3D content generation and the limitations of CAD-based approaches, this study explores sketch-based 3D modeling as a promising alternative. By leveraging the intuitive and natural form of computer–human interaction, we aim to allow users to create 3D models using freehand sketches as input, thereby greatly simplifying the learning progress of 3D modeling, thus resulting in an increase in the amount of quality 3D content produced.

Existing sketch-based 3D modeling approaches are far from perfect. Many existing approaches require precise line drawings from multiple perspectives or follow a step-by-step workflow that assumes a strategic understanding of 3D modeling (Cohen et al., 1999; Deng et al., 2020). These methods, although effective, can be challenging for novice users and time-consuming. Additionally, other approaches that employ template primitives or retrieval-based techniques (Chen DY et al., 2003; Wang F et al., 2015) lack the customizability which is necessary to allow users to fully express their creative ideas. Thus, a more balanced approach is needed to provide both ease of use and flexibility for novice users, thereby enabling them to create custom 3D models with the minimum effort and maximum creativity.

To accomplish the task of efficient and user-friendly 3D modeling, our approach aims to generate a detailed 3D model from a single freehand sketch input. This is a challenging task as the input is limited to a single sketch with the minimum information. Previous studies have used deep neural networks to achieve the sketch-to-model translation, using an encoder–decoder architecture that compresses the input sketch into a coarse representation (latent code) capturing information such as the semantic category and conceptual shape, followed by recovery of the 3D shapes using a decoder that calculates the offsets of a given number of points or vertices (Guillard et al., 2021; Zhang SH et al., 2021; Chen TR et al., 2023a, 2023b, 2023c; Zang et al., 2023). However, these approaches have difficulties in capturing intricate details due to the significant domain gap between a sketch and a 3D shape domain, and the limited resolution in points or vertex representations.

Here, as shown in Fig. 1, we propose Deep3DSketch-im, which can elevate the resolution of sketch-to-3D modeling to a new level. This resolution enhancement is enabled by different data representations—those other than the aforementioned points or voxels. Securing an integration of the implicit 3D surface representation, namely signed distance functions, into the sketch-to-model process would be the first step toward achieving this enhancement. The signed distance function encodes the distance of each point sample in 3D from the boundary of the shape, with a sign indicating whether the point is inside or outside the shape (Fig. 2). For our sketch-to-model task, a convolutional neural network (CNN) first encodes the input sketch into a feature vector. Then, we use this feature vector to predict the signed distance field (SDF) value of a 3D point. By sampling different points with infinite possible locations, Deep3DSketch-im generates an implicit field of the underlying surface with infinite resolution.

We have performed extensive experiments using synthetic and real hand-drawn datasets. Experimental results show that our approach achieves state-of-the-art (SOTA) performance. Our approach can reconstruct more details with higher-fidelity results. Our user study also shows that users are more satisfied with the models obtained by our approach. We believe that our approach allows for a wide range of applications, such as 3D printing, virtual reality, and video game development. Additionally, the ability to create custom 3D models quickly and easily could have significant impacts on industries such as architecture, product design, and entertainment. Deep3DSketch-im could become a powerful tool for democratizing 3D modeling, making it accessible to a wider range of users and potentially revolutionizing the way by which we design and create in 3D.

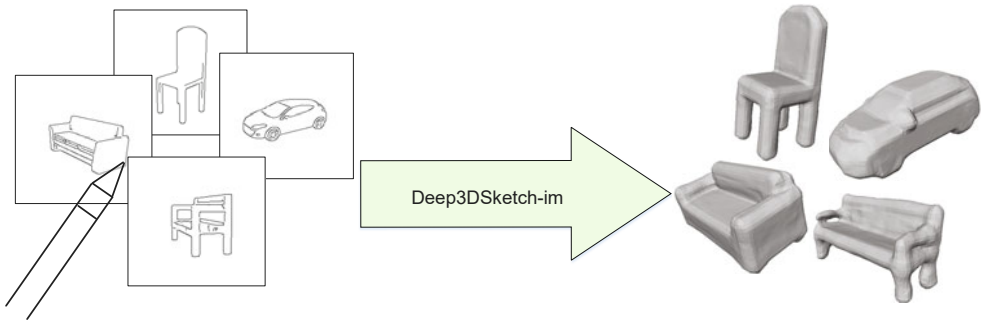

Fig. 1 Pipeline of our sketch-based 3D modeling approach

Our approach takes a single-view freehand sketch and feeds it into an end-to-end neural network; a high-fidelity full 3D model is obtained using the given sketch

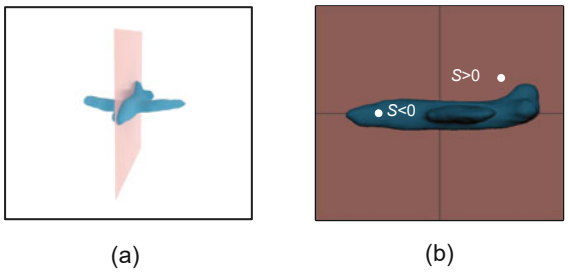

Fig. 2 Illustration of the signed distance field (SDF): (a) rendered 3D surface with $S=0$; (b) cross-section of the SDF

A point is outside the surface if $S>0$, inside the surface if $S<0$, and on the surface if $S=0$ (S is the distance from the point to the iso-surface)

2 Related works

2.1 Sketch-based 3D modeling

Sketch-based 3D modeling has been an active research area for many years, with numerous approaches proposed by researchers. One category of sketch-based 3D modeling approaches is interactive approaches, which requires breaking down the task into sequential steps or using specific drawing gestures or annotations. These methods have been shown to require significant strategic knowledge, making them challenging for novice users. For instance, Li CJ et al. (2020) used a two-stage approach for coarse-to-fine reconstruction, while Cohen et al. (1999) used annotation-based feedback to refine the 3D model.

In contrast, end-to-end approaches such as template primitives or retrieval-based methods tend to be more straightforward but lack customizability. These approaches involve generating the 3D model directly from the sketch without any intermediate steps. For example, Chen DY et al. (2003) used 3D geometric primitives for sketch-based modeling, while Wang F et al. (2015) introduced a retrieval-based approach that uses a database of 3D models to find the closest match to the input sketch. Recently, deep learning based approaches have been proposed for single-view 3D reconstruction, including sketch-based 3D modeling. For example, Zhang SH et al. (2021) and Chen TR et al. (2023a, 2023b, 2023c) proposed the use of an encoder–decoder backbone to output the offset of a round shape template, but the adoption of such an approach is able to result in only the representation of coarse shapes characterized by a lack of structural details. Gao et al. (2022) used density maps and point clouds as the representation; though the representation is capable of generating high-fidelity results, further processing needs to be done for most applications requiring mesh representation. Guillard et al. (2021) proposed methods that reconstruct the 3D model with a two-stage refinement scheme, but it cannot provide high-fidelity generation in real time. Moreover, these approaches face substantial challenges due to the sparse and abstract nature of sketches lacking fine boundary information and texture information for depth estimation, making it difficult to produce high-quality 3D shapes. In Zhong et al. (2020), these challenges were illustrated and analyzed in detail.

In this study, we introduce a novel approach called Deep3DSketch-im, which leverages SDF to represent 3D surfaces, rather than using point clouds or voxels, to generate higher-fidelity 3D models. By incorporating SDFs, our approach can capture more structural details in the input sketches and produce

models with a theoretically infinite resolution. This sets our approach apart from existing methods, as it does not require a fixed topology assumption and can produce accurate ground truths without approximating metrics (Xu et al., 2022, 2023; Lin GY et al., 2023; Yang et al., 2023). This makes it suitable for both novice and experienced users who want to create high-quality 3D models from sketches.

2.2 Single-view 3D reconstruction

The task of reconstructing 3D geometry from a single 2D image has been a challenging problem in the fields of computer vision and computer graphics for many years. In recent years, data-driven approaches have gained popularity with the advent of large-scale datasets like ShapeNet (Chang et al., 2015). Some works (Chen ZQ and Zhang, 2019; Park et al., 2019) use category-level information to infer 3D representations, while others (Kato et al., 2018; Liu et al., 2019a, 2019b) directly generate 3D models from 2D images using differentiable rendering techniques. More recently, unsupervised methods for implicit function representations using differentiable rendering have been proposed (Lin CH et al., 2020; Yu et al., 2021).

However, most of these methods focus on learning 3D geometry from 2D colored images. In contrast, our approach aims to generate 3D meshes from 2D sketches, which are a more abstract and sparse form of image representation. Sketches lack important information like texture, lighting, and shading, making it challenging to infer 3D geometry accurately (Chen TR et al., 2023a, 2023b; Zang et al., 2023; Zhang SZ et al., 2023). Additionally, sketches are often incomplete, and the same set of strokes can have different interpretations in 3D, adding ambiguity to the problem. Therefore, it is critical to develop a method that can accurately interpret and reconstruct 3D shapes from sparse and ambiguous sketches.

In this study, we propose a novel approach that addresses these challenges and provides an efficient and accurate solution for sketch-based 3D modeling. Our approach uses a deep learning based approach that learns to interpret the abstract representation of sketches and reconstructs a high-quality 3D mesh. We show that our approach outperforms SOTA methods on benchmark datasets for sketch-based 3D modeling.

3 Method

3.1 Signed distance field

Our objective is to generate a highly detailed and accurate 3D model of an object from a given image. To achieve this, we adopt a novel approach that represents the 3D shape as an SDF (Tong X, 2022). By representing the shape as an SDF, we can model the object's surface as a level set of the function, allowing us to generate a high-resolution mesh of the object's surface. As illustrated in Fig. 2, the signed distance function is a continuous function that maps a given spatial point $p = (x, y, z) \in \mathbb{R}^3$ to a real value: $s = S_{\mathrm{DF}}(p)$, where S_{DF} refers to the signed distance function. In contrast to commonly used 3D representations such as depth, the absolute value of $S_{\mathrm{DF}}(p)$ indicates the distance of a point p to the surface, and the sign of $S_{\mathrm{DF}}(p)$ indicates whether p is inside or outside the surface. The iso-surface $\mathcal{S}' = \{p|S_{\mathrm{DF}}(p) = 0\}$ implicitly represents the 3D shape. In implementation, we first define a dense 3D grid and predict signed distance function values for each point in the grid. With these values calculated, we can then use the Marching Cubes algorithm to obtain the 3D mesh that corresponds to the iso-surface $\mathcal{S}'$.

3.2 Network architecture

As illustrated in Fig. 3a, the conventional sketch-to-3D modeling approach is composed of an encoder–decoder network structure, where the encoder E transforms the sparse and ambiguous input sketch into a latent shape code z_{s} that summarizes the sketch at a coarse level. The decoder D is then used to transfer the latent shape code z_{s} to the mesh $M_{\Theta} = D(z_{\mathrm{s}})$. Nevertheless, this design has the limitation that only coarse shapes are generated when, owing to the fact of the number of vertices or points being limited, the resolution is low. Moreover, the significant domain gaps between sketches and 3D shapes make it challenging to generate high-fidelity 3D shapes.

On the contrary, our approach involves projecting each 3D query point onto the image plane and gathering multiscale CNN features for the corresponding image patch. The collected features are then used by the neural network to decode the given spatial point into an SDF value, using the multiscale

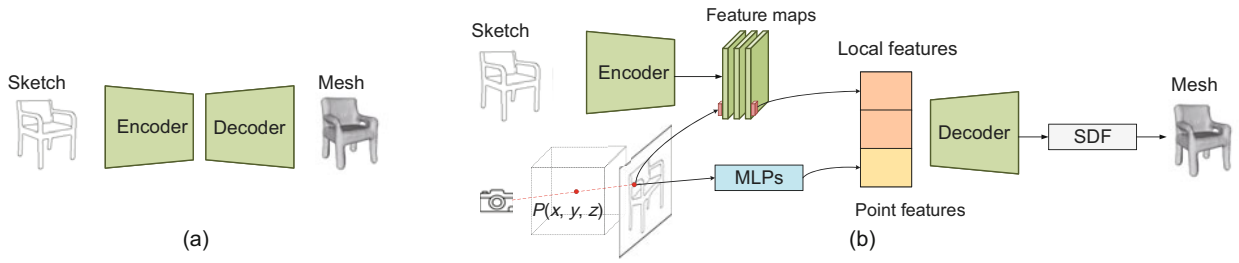

Fig. 3 Network architecture: (a) existing encoder–decoder structure of neural networks; (b) pipeline of our Deep3DSketch-im, which samples points at mesh and projects the points onto the image plane

The network calculates the signed distance field (SDF) value of the point. The mesh is obtained by applying the Marching Cubes algorithm to the SDF. MLP: multilayer perceptron

local image features.

The details of our network structure are illustrated in Fig. 3b. The input sketch can be seen as a binary image $I \in \{0,1\}^{W\times H}$ (W and H are the width and height of the sketch, respectively), in which $I[i,j]=0$ if marked by the stroke, and $I[i,j]=1$ otherwise. The image of the sketch is fed into an image encoder for multiscale image feature extraction, obtaining feature maps at different scales. Meanwhile, we project a 3D point $p \in \mathbb{R}^3$ onto the image plane, obtaining a 2D location $q \in \mathbb{R}^2$. We obtain the local image features by retrieving the features on each feature map corresponding to location q and concatenating them. As the feature maps in the later layers have smaller dimensions than the original image, we resize them to the original size with bilinear interpolation and extract the resized features at location q.

Additionally, as proposed in Wang WY et al. (2019), we extract point features for each point q by applying 1×1 convolution and rectified linear unit (ReLU) activation functions after the convolution, resulting in feature vectors of increased length. We concatenate the point features and local features into a new feature vector, and then feed the new feature vector into a decoder that involves multiple 1×1 convolutions and ReLU activation functions to obtain the SDF prediction value. The local feature is used to capture detailed information in the sketch input, which is demonstrated effectively in our later experiment.

3.3 Sketch view prediction

Note that the abovementioned network needs the sketch pose as the input. However, for a more user-friendly experience, it is better to allow users to use the network without additional input. We find that we can predict pose based on synthesized data. Specifically, following the approach adopted in Chen TR et al. (2023a, 2023c), we design a separate pose-estimation network and train it in a fully supervised manner. We use an encoder E to produce latent code z_l from the sketches and input it to the viewpoint prediction module, which consists of two fully connected layers D_v aimed at producing the viewpoint estimate ξ_pred, represented by an Euler angle. The viewpoint prediction module is optimized in a fully supervised manner with the input of the ground-truth (GT) viewpoint ξ_gt, supervised by a viewpoint prediction loss $\mathcal{L}_\mathrm{v}$, which adopts the mean-squared error (MSE) loss for the predicted and GT poses, defined as

$$\mathcal{L}_\mathrm{v} = \|\xi_\mathrm{gt} - \xi_\mathrm{pred}\|_2 = \|\xi_\mathrm{gt} - D_\mathrm{v}(z_\mathrm{l})\|_2 . \quad (1)$$

The viewpoint prediction model is trained along with a 3D model generation process as in Chen TR et al. (2023b, 2023c). In the experiment, we find that the predicted pose can effectively guide the network with very little performance drop.

3.4 Loss function

We adopt a continuous signed distance function regression approach that allows us to extract surfaces corresponding to different iso-values as in Wang WY et al. (2019). To focus the network on recovering details near and inside the iso-surface, we use a weighted loss function. The loss function is defined as

$$\mathcal{L} = \sum_p m \left| f(I,p) - S^I_\mathrm{DF}(p) \right|, \quad (2)$$

where $|\cdot|$ is the L_1-norm, $S^I_\mathrm{DF}(\cdot)$ is our GT signed distance function, f is our neural network, and m is a

hyperparameter satisfying $m{=}m_1$ if $S^I_{\mathrm{DF}}(\cdot)$ is smaller than a certain threshold δ and $m{=}m_2$ in other cases.

4 Experiments

4.1 Dataset

There is a limited availability of datasets that include both sketches and their corresponding 3D models for research purposes. Zhang SH et al. (2021), recently used synthetic data from the ShapeNet-Synthetic dataset for their training data. The ShapeNet-Synthetic dataset includes 13 categories of 3D objects, and synthetic data are generated using a canny edge detector on rendered images from Kar et al. (2017). The trained model is then evaluated on real-world data from the ShapeNet-Sketch dataset, which includes 1300 sketches and their corresponding 3D shapes. These sketches are drawn by human volunteers with varying levels of skills, based on images of 3D objects from Kar et al. (2017).

4.2 Implementation details

The image encoder of the sketch is VGG-16. The decoder consists of three layers, with 1×1 convolution followed by an ReLU function on the first two layers, and no activation function on the last layer, which is a 1×1 value prediction. The decoder takes the concatenation of point features and local multi-scale features. During training, we focus on points near the iso-surface $\mathcal{S}'$, which is achieved by employing Monte Carlo sampling. Specifically, we randomly select 2048 grid points from a Gaussian distribution $N(0, 0.1)$. Additionally, we set the parameters $m_1 = 4$, $m_2 = 1$, and $\delta = 0.01$ in Eq. (2) to ensure effective recovery of details near and inside the iso-surface. Our network is implemented using PyTorch and optimized using the Adam optimizer with a learning rate of $1{\times}10^{-4}$ and a batch size of 16, and employs an NVIDIA GeForce RTX 3090 Graphics Card.

4.3 Experimental results and performance comparison

We assess the performance of our method by comparing it with that of the SOTA model, following the same protocol as in Zhang SH et al. (2021). The model is trained for each category. We use the official training–evaluation–testing split and evaluate both the ShapeNet-Synthetic (edge-detected sketch) and ShapeNet-Sketch (hand-drawn sketches) datasets. In Table 1, the ShapeNet-Synthetic dataset is used for experiments, which provides accurate GT 3D models for training and evaluation. To evaluate the fidelity of the generated meshes, we employ the Chamfer distance metric, which is a widely used measure for 3D reconstruction. Some latest baseline approaches, namely Sketch2Model (Zhang SH et al., 2021), Deep3DSketch (Chen TR et al., 2023a), Sketch2Mesh (Guillard et al., 2021), and the deep implicit surface network (DISN) (Wang WY et al., 2019) are used for comparison. For a fair comparison, we use only the first feed-forward stage of Sketch2Mesh and do not perform the post-processing optimization step (shown as Sketch2Mesh). In the experiment, our approach demonstrates high effectiveness and achieves SOTA performance. The visualization presented in Fig. 4 also emphasizes the significant performance elevation of our proposed Deep3DSketch-im.

In Table 2, we further evaluate the performance in relation to real-world human drawings through the ShapeNet-Sketch dataset. Due to the limited number of samples, we train the model on the ShapeNet-Synthetic dataset and use the ShapeNet-Sketch dataset for evaluation. The results also show that our approach can consistently produce higher-fidelity results when it comes to real hand-drawn datasets, as illustrated in Fig. 5. Particularly, many detailed structures are accurately captured by our proposed Deep3DSketch-im, for example, the top and the side mirror of the car. The results show that Deep3DSketch-im is a robust network that can generalize well in real-world data, while it is trained only on synthetic data from edge detectors. The visualization results illustrate the obvious effectiveness of Deep3DSketch-im in producing models with higher quality and fidelity in structure. Further research can be conducted in terms of domain generalization and domain adaptation to better mitigate the domain gap between the synthetic and real-world data (Tong YZ et al., 2023; Zhu et al., 2023a, 2023b, 2023c) or collect more real data for training.

4.4 Using Deep3DSketch-im w/o view input

As mentioned above, despite inputting GT pose during the evaluation in Tables 1 and 2, we argue that the view information is not needed as the user

Table 1 Quantitative evaluation of the ShapeNet-Synthetic dataset

Method	Chamfer distance (1×10^{-3})						
	Car	Sofa	Airplane	Bench	Display	Chair	Table
Sketch2Model (Zhang SH et al., 2021)	15.2595	42.3509	22.9386	23.3147	24.0722	61.9558	21.8722
Deep3DSketch (Chen TR et al., 2023a)	12.5674	43.9631	23.4149	23.5360	23.3090	61.2654	20.8446
Sketch2Mesh* (Guillard et al., 2021)	11.4795	25.7326	**9.2896**	**9.0149**	**15.6744**	16.8339	**17.8063**
DISN (Wang WY et al., 2019)	7.8952	**17.6470**	11.5923	12.9683	16.6323	**15.1664**	25.8620
Ours	**7.4214**	17.7643	11.7236	13.6058	16.9012	15.3529	25.7168

Method	Chamfer distance (1×10^{-3})						
	Telephone	Cabinet	Loudspeaker	Watercraft	Lamp	Rifle	Mean
Sketch2Model (Zhang SH et al., 2021)	18.8203	18.6660	20.7303	15.7249	60.3407	19.0046	28.0808
Deep3DSketch (Chen TR et al., 2023a)	16.1120	18.3639	22.2328	15.2500	56.4119	19.3038	27.4290
Sketch2Mesh* (Guillard et al., 2021)	17.6221	20.4426	**12.0591**	**8.9920**	33.2874	8.8732	15.9314
DISN (Wang WY et al., 2019)	8.7911	16.0838	18.6744	16.2405	40.3034	8.0286	16.5710
Ours	**8.6561**	**15.8494**	19.0436	16.1779	**30.3812**	**7.9500**	**15.8880**

The best results are in bold. * For a fair comparison, we use only the first feed-forward stage of Sketch2Mesh and do not perform the post-processing optimization step

Table 2 Quantitative evaluation of the ShapeNet-Sketch dataset

Method	Chamfer distance (1×10^{-3})						
	Car	Sofa	Airplane	Bench	Display	Chair	Table
Sketch2Model (Zhang SH et al., 2021)	16.5391	35.1184	21.0765	24.6853	22.9207	66.5329	22.2305
Deep3DSketch (Chen TR et al., 2023a)	13.5601	37.4461	21.4698	26.3255	20.8404	67.5253	21.7184
Sketch2Mesh* (Guillard et al., 2021)	10.9910	**12.1899**	**9.5388**	**9.2354**	**14.9054**	16.6400	**18.6540**
DISN (Wang WY et al., 2019)	7.8952	17.1633	10.5889	14.4459	15.0854	17.6491	18.7602
Ours	**7.8444**	16.8443	11.2969	14.4989	14.9134	**16.2878**	28.4224

Method	Chamfer distance (1×10^{-3})						
	Telephone	Cabinet	Loudspeaker	Watercraft	Lamp	Rifle	Mean
Sketch2Model (Zhang SH et al., 2021)	18.9218	22.2107	23.0350	17.4914	51.9116	19.8603	27.8872
Deep3DSketch (Chen TR et al., 2023a)	17.0649	22.8680	24.9334	16.8122	52.8507	19.6919	27.9313
Sketch2Mesh* (Guillard et al., 2021)	17.6146	20.8006	27.5137	**10.3312**	35.6106	8.9314	16.4671
DISN (Wang WY et al., 2019)	11.1023	20.4927	**21.6061**	15.7442	32.1064	7.8680	**16.1071**
Ours	**11.0535**	**19.7184**	22.1876	16.1833	**31.0735**	**7.7296**	16.7734

The best results are in bold. * For a fair comparison, we use only the first feed-forward stage of Sketch2Mesh and do not perform the post-processing optimization step

input because the view prediction network can predict the view and input the same to the model generation process, which is also beneficial for user experience in real-world applications. We test the performance with no GT pose but only predicted poses. As shown in Fig. 6, Deep3DSketch-im without viewpoint as the input works as well as the network with GT poses. The quantitative evaluation results are shown in Table 3.

4.5 Runtime & time complexity evaluation

Once the network is adequately trained, we measure its performance on a computer equipped with a consumer graphics card (NVIDIA GeForce RTX 3090). Our approach demonstrates a generation speed of 0.97 s, which is sufficient for fluent human–computer interaction.

Note that in contrast with the approaches adopted in some NeRF studies (Fu et al., 2022;

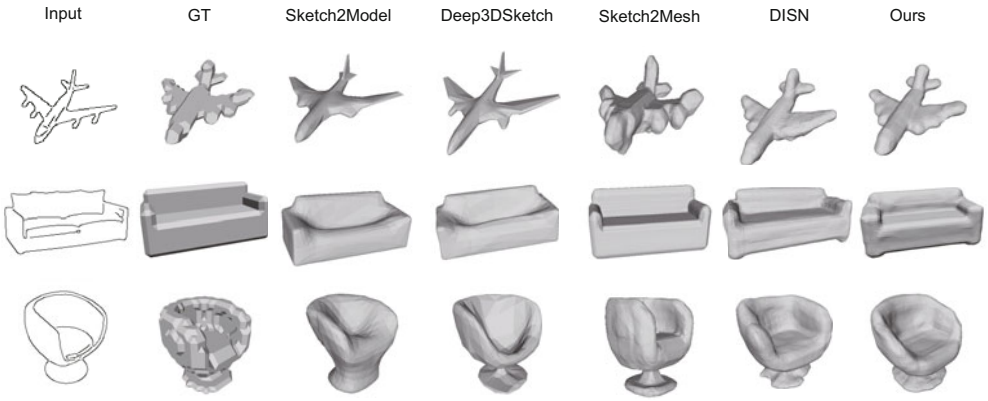

Fig. 4 Qualitative evaluation with state-of-the-art (SOTA) method in a synthesized dataset

The visualization of 3D models generated from the ShapeNet-Synthetic dataset demonstrates that our method is capable of synthesizing higher-fidelity of 3D structures. The detailed structures such as the two engines in the first row are reconstructed by our approach but do not appear in the results obtained by other approaches. For Sketch2Mesh, for a fair comparison, only the model generated by a feed-forward neural network is used without further post-processing. GT: ground-truth; DISN: deep implicit surface network

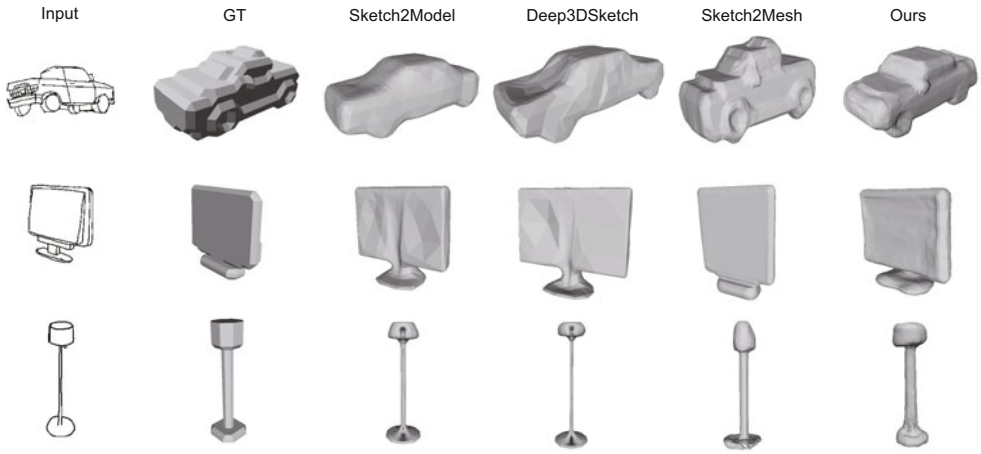

Fig. 5 Qualitative evaluation with state-of-the-art (SOTA) method for a real-world dataset

The visualization of 3D models generated from the ShapeNet-Sketch dataset demonstrates that our method is capable of synthesizing higher-fidelity 3D structures. The detailed structures such as the side-view mirrors of the car in the first row are reconstructed by our approach but do not appear in the results obtained by other approaches. For Sketch2Mesh, for a fair comparison, only the model generated by a feed-forward neural network is used without further post-processing. GT: ground-truth

Metzer et al., 2022; Jo et al., 2023), our approach is a generalized approach that does not require time-consuming per-object optimization. The time complexity is related to the number of points. If we assume that the voxelized grid used to represent the target model has a size of $N \times N \times N$, where N resprents the resolution or density of the grid in each dimension, our algorithm's complexity is $O(N^3)$ as we calculate the SDF value for each grid cell.

4.6 User study

To further assess the effectiveness of our sketch-to-model algorithm, we carry out a user study using the mean option score (MOS) metric, which ranges from 1 to 5 (Seufert, 2019). The experiments follow the settings adopted in various studies (Cai et al., 2021; Michel et al., 2022; Yao et al., 2022; Chen TR et al., 2023c). We present 36 3D models generated by our algorithm to 12 designers who are well-versed in 3D content. The designers are asked to rate the

Table 3 Quantitative evaluation of ShapeNet-Synthetic dataset w/o view input

Method	Chamfer distance (1×10^{-3})						
	Car	Sofa	Airplane	Bench	Display	Chair	Table
Ours w/o GT pose	7.9350	**17.0330**	**11.2280**	14.3931	**14.8556**	16.4007	28.2873
Ours w GT pose	**7.4214**	17.7643	11.7236	**13.6058**	16.9012	**15.3529**	**25.7168**

Method	Chamfer distance (1×10^{-3})						
	Telephone	Cabinet	Loudspeaker	Watercraft	Lamp	Rifle	Mean
Ours w/o GT pose	11.2339	19.7097	22.0595	16.2492	30.6835	74.9817	21.9267
Ours w GT pose	**8.6561**	**15.8494**	**19.0436**	**16.1779**	**30.3812**	**7.9500**	**15.8880**

The better results are in bold. w/o: without; w: with

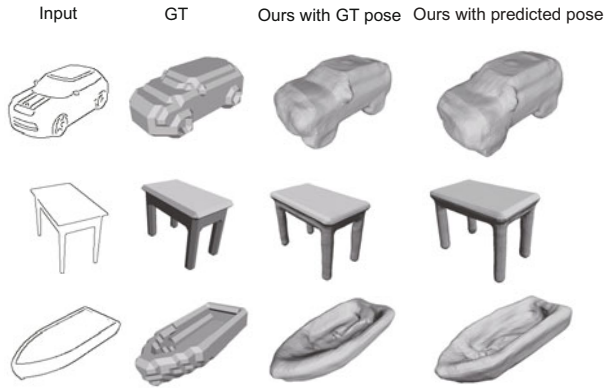

Fig. 6 Qualitative evaluation with predicted pose

We can use the predicted pose inputted to the Deep3DSketch-im network. GT: ground-truth

models based on two factors:

Q1: How well does the output 3D model match the input sketch (fidelity)?

Q2: How do you rate the quality of the output 3D model (quality)?

Before the experiment, each participant is given a brief one-on-one introduction to the concepts of fidelity and quality. We average the scores and report the rating results in Table 4. The results indicate that our method outperforms the SOTA method in terms of users' subjective ratings.

Table 4 Mean opinion scores (1–5) for Q1 (fidelity) and Q2 (quality)

Method	Score	
	Q1: fidelity	Q2: quality
Deep3DSketch	3.39 ± 0.26	3.01 ± 0.33
Ours	**3.97 ± 0.31**	**4.04 ± 0.41**

The better results are in bold

5 Real-world implication of the proposed approach

Sketching is the most natural and intuitive way to express new ideas, and a single-view sketch is definitely the easiest approach for users to implement their thoughts. Although the downstream application is not the particular focus of the study, we note one particular application of sketch-based 3D modeling that is rapid home interior design, which is the follow-up work to Chen TR et al. (2023c). With the involvement of context information, users can design the 3D model using sketches and place models within a real environment for a more immersive experience. In other words, the sketch can precisely define the six-dimensional (6D) pose and position of the generated object. We have demonstrated that the sketch-guided approach is more efficient and easier to use compared to a "touch-based" approach concerning users' utility in terms of manipulating the generated 3D objects within the scene. Note that future research can focus on expanding the application of this sketch-based modeling tool, with appropriate shape representations (Xu et al., 2022; Yang et al., 2023; Zang et al., 2023), to eventually CAD/computer-aided manufacturing (CAM) and many other fields.

6 Conclusions

We have presented a novel deep learning network, Deep3DSketch-im, for generating 3D models from a single 2D sketch. Considering the lack of fine details characterizing the existing sketch-to-model approaches, we first introduce SDF to represent the 3D shape for infinite resolution. We design a

network to capture the local features for fine-grained structure 3D modeling. Through experiments on the ShapeNet-Synthetic dataset, we have shown that our approach outperforms SOTA methods in terms of both quantitative metrics and user study ratings. Our method also expands the existing language- or vision-centered AIGC tools. We believe that our study opens up exciting possibility for creating 3D content from simple 2D sketches, which can have significant applications in industries such as gaming, animation, and architecture.

Contributors

Tianrun CHEN designed the research. Tianrun CHEN and Runlong CAO processed the data and performed the experiments. Tianrun CHEN drafted the paper. Zejian LI, Ying ZANG, and Lingyun SUN revised and finalized the paper.

Compliance with ethics guidelines

Lingyun SUN is an editor-in-chief assistant of this special issue, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

Our project data can be found at https://tianrunchen.github.io/Deep3DSketch-im. Other data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Cai YJ, Wang YW, Zhu YH, et al., 2021. A unified 3D human motion synthesis model via conditional variational auto-encoder. IEEE/CVF Int Conf on Computer Vision, p.11625-11635. https://doi.org/10.1109/ICCV48922.2021.01144

Chang AX, Funkhouser T, Guibas L, et al., 2015. ShapeNet: an information-rich 3D model repository. https://arxiv.org/abs/1512.03012

Chen DY, Tian XP, Shen YT, et al., 2003. On visual similarity based 3D model retrieval. *Comput Graph Forum*, 22(3):223-232. https://doi.org/10.1111/1467-8659.00669

Chen TR, Fu CL, Zhu LY, et al., 2023a. Deep3DSketch: 3D modeling from free-hand sketches with view- and structural-aware adversarial training. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096348

Chen TR, Fu CL, Zang Y, et al., 2023b. Deep3DSketch+: rapid 3D modeling from single free-hand sketches. Proc 29th Int Conf on Multimedia Modeling, p.16-28. https://doi.org/10.1007/978-3-031-27818-1_2

Chen TR, Ding CT, Zhu LY, et al., 2023c. Reality3DSketch: rapid 3D modeling of objects from single freehand sketches. IEEE Trans Multim, early access. https://doi.org/10.1109/TMM.2023.3327533

Chen ZQ, Zhang H, 2019. Learning implicit fields for generative shape modeling. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5932-5941. https://doi.org/10.1109/CVPR.2019.00609

Chester I, 2007. Teaching for CAD expertise. *Int J Technol Des Educ*, 17:23-35. https://doi.org/10.1007/s10798-006-9015-z

Cohen JM, Markosian L, Zeleznik RC, et al., 1999. An interface for sketching 3D curves. Symp on Interactive 3D Graphics, p.17-21. https://doi.org/10.1145/300523.300655

Deng CY, Huang JH, Yang YL, 2020. Interactive modeling of lofted shapes from a single image. *Comput Visual Med*, 6(3):279-289. https://doi.org/10.1007/s41095-019-0153-0

Fu X, Zhang SZ, Chen TR, et al., 2022. Panoptic NeRF: 3D-to-2D label transfer for panoptic urban scene segmentation. Int Conf on 3D Vision, p.1-11. https://doi.org/10.1109/3DV57658.2022.00042

Gao CJ, Yu Q, Sheng L, et al., 2022. SketchSampler: sketch-based 3D reconstruction via view-dependent depth sampling. Proc 17th European Conf on Computer Vision, p.464-479. https://doi.org/10.1007/978-3-031-19769-7_27

Guillard B, Remelli E, Yvernay P, et al., 2021. Sketch2Mesh: reconstructing and editing 3D shapes from sketches. IEEE/CVF Int Conf on Computer Vision, p.13003-13012. https://doi.org/10.1109/ICCV48922.2021.01278

Huang SS, Wang YH, 2024. Controllable image generation based on causal representation learning. *Front Inform Technol Electron Eng*, 25(1):135-148. https://doi.org/10.1631/FITEE.2300303

Jo K, Shim G, Jung S, et al., 2023. CG-NeRF: conditional generative neural radiance fields for 3D-aware image synthesis. IEEE/CVF Winter Conf on Applications of Computer Vision, p.724-733. https://doi.org/10.1109/WACV56688.2023.00079

Kar A, Häne C, Malik J, 2017. Learning a multi-view stereo machine. Proc 31st Int Conf on Neural Information Processing Systems, p.364-375.

Kato H, Ushiku Y, Harada T, 2018. Neural 3D mesh renderer. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3907-3916. https://doi.org/10.1109/CVPR.2018.00411

Lei YM, Li JQ, 2024. Prompt learning in computer vision: a survey. *Front Inform Technol Electron Eng*, 25(1):42-63. https://doi.org/10.1631/FITEE.2300389

Li CJ, Pan H, Bousseau A, et al., 2020. Sketch2CAD: sequential CAD modeling by sketching in context. *ACM Trans Graph*, 39(6):164. https://doi.org/10.1145/3414685.3417807

Lin CH, Wang CY, Lucey S, 2020. SDF-SRN: learning signed distance 3D object reconstruction from static images. Proc 34th Int Conf on Neural Information Processing Systems, Article 961.

Lin GY, Yang L, Zhang CY, et al., 2023. Patch-Grid: an efficient and feature-preserving neural implicit surface representation. https://arxiv.org/abs/2308.13934

Liu SC, Saito S, Chen WK, et al., 2019a. Learning to infer implicit surfaces without 3D supervision. Proc 33rd Int Conf on Neural Information Processing Systems, Article 32.

Liu SC, Chen WK, Li TY, et al., 2019b. Soft rasterizer: a differentiable renderer for image-based 3D reasoning. IEEE/CVF Int Conf on Computer Vision, p.7707-7716. https://doi.org/10.1109/ICCV.2019.00780

Mahapatra C, Jensen JK, McQuaid M, et al., 2019. Barriers to end-user designers of augmented fabrication. CHI Conf on Human Factors in Computing Systems, Article 383. https://doi.org/10.1145/3290605.3300613

Metzer G, Richardson E, Patashnik O, et al., 2022. Latent-NeRF for shape-guided generation of 3D shapes and textures. https://arxiv.org/abs/2211.07600

Michel O, Bar-On R, Liu R, et al., 2022. Text2Mesh: text-driven neural stylization for meshes. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.13482-13492. https://doi.org/10.1109/CVPR52688.2022.01313

Park JJ, Florence P, Straub J, et al., 2019. DeepSDF: learning continuous signed distance functions for shape representation. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.165-174. https://doi.org/10.1109/CVPR.2019.00025

Reddy EJ, Rangadu VP, 2018. Development of knowledge based parametric CAD modeling system for spur gear: an approach. *Alex Eng J*, 57(4):3139-3149. https://doi.org/10.1016/j.aej.2018.07.010

Seufert M, 2019. Fundamental advantages of considering quality of experience distributions over mean opinion scores. Proc 11th Int Conf on Quality of Multimedia Experience, p.1-6. https://doi.org/10.1109/QoMEX.2019.8743296

Tong X, 2022. Three-dimensional shape space learning for visual concept construction: challenges and research progress. *Front Inform Technol Electron Eng*, 23(9):1290-1297. https://doi.org/10.1631/FITEE.2200318

Tong YZ, Yuan JK, Zhang M, et al., 2023. Quantitatively measuring and contrastively exploring heterogeneity for domain generalization. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.2189-2200. https://doi.org/10.1145/3580305.3599481

Wang F, Kang L, Li Y, 2015. Sketch-based 3D shape retrieval using convolutional neural networks. IEEE Conf on Computer Vision and Pattern Recognition, p.1875-1883. https://doi.org/10.1109/CVPR.2015.7298797

Wang WY, Xu QG, Ceylan D, et al., 2019. DISN: deep implicit surface network for high-quality single-view 3D reconstruction. Proc 33rd Int Conf on Neural Information Processing Systems, Article 45.

Xu R, Wang ZX, Dou ZY, et al., 2022. RFEPS: reconstructing feature-line equipped polygonal surface. *ACM Trans Graph*, 41(6):228. https://doi.org/10.1145/3550454.3555443

Xu R, Dou ZY, Wang NN, et al., 2023. Globally consistent normal orientation for point clouds by regularizing the winding-number field. *ACM Trans Graph*, 42(4):111. https://doi.org/10.1145/3592129

Yang L, Liang YQ, Li X, et al., 2023. Neural parametric surfaces for shape modeling. https://arxiv.org/abs/2309.09911

Yao SY, Zhong RZ, Yan YC, et al., 2022. DFA-NeRF: personalized talking head generation via disentangled face attributes neural rendering. https://arxiv.org/abs/2201.00791

Yu A, Ye V, Tancik M, et al., 2021. pixelNeRF: neural radiance fields from one or few images. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4576-4585. https://doi.org/10.1109/CVPR46437.2021.00455

Zang Y, Fu CL, Chen TR, et al., 2023. Deep3DSketch+: obtaining customized 3D model by single free-hand sketch through deep learning. https://arxiv.org/abs/2310.18609

Zhang SH, Guo YC, Gu QW, 2021. Sketch2Model: view-aware 3D modeling from single free-hand sketches. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6000-6017. https://doi.org/10.1109/CVPR46437.2021.00595

Zhang SZ, Peng SD, Chen TR, et al., 2023. Painting 3D nature in 2D: view synthesis of natural scenes from a single semantic mask. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8518-8528. https://doi.org/10.1109/CVPR52729.2023.00823

Zhong Y, Gryaditskaya Y, Zhang HG, et al., 2020. Deep sketch-based modeling: tips and tricks. Int Conf on 3D Vision, p.543-552. https://doi.org/10.1109/3DV50981.2020.00064

Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2300089

Zhu DD, Li YC, Zhang M, et al., 2023a. Bridging the gap: neural collapse inspired prompt tuning for generalization under class imbalance. https://arxiv.org/abs/2306.15955v2

Zhu DD, Li YC, Shao YF, et al., 2023b. Generalized universal domain adaptation with generative flow networks. Proc 31st ACM Int Conf on Multimedia, p.8304-8315. https://doi.org/10.1145/3581783.3612225

Zhu DD, Li YC, Yuan JK, et al., 2023c. Universal domain adaptation via compressive attention matching. IEEE/CVF Int Conf on Computer Vision, p.6974-6985.