Computer-Aided Design 185 (2025) 103896

Contents lists available at ScienceDirect

Computer-Aided Design

journal homepage: www.elsevier.com/locate/cad

Research Paper

Sketch123: Multi-spectral channel cross attention for sketch-based 3D generation via diffusion models

Zhentong Xu [a], Long Zeng [b], Junli Zhao [a],*, Baodong Wang [a], Zhenkuan Pan [a], Yong-Jin Liu [c]

[a] College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China
[b] Graduated School at Shenzhen, Tsinghua University, Shenzhen, 518055, China
[c] MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Keywords:
Sketch-based generation
Diffusion model
Attention mechanism
Spectral channel attention

ABSTRACT

With the development of generative techniques, sketch-driven 3D reconstruction has gained substantial attention as an efficient 3D modeling technique. However, challenges remain in extracting detailed features from sketches, representing local geometric structures, and ensuring generated fidelity and stability. To address these issues, in this paper we propose a multi-spectral channel cross-attention model for sketch reconstruction, which leverages the complementary strengths of frequency and spatial domains to capture multi-level sketch features. Our method employs a two-stage diffusion generation mechanism, additionally, a Sparse Feature Enhancement Module (SFE) replaces traditional down-sampling, reducing feature loss and enhancing detail preservation and noise suppression through a Laplace voxel smoothing operator. The Wasserstein distance introduced and integrated as part of the loss function, stabilizes the generative process using optimal transport theory to support high-quality 3D model reconstruction. Extensive experiments verify that our model surpasses state-of-the-art methods in terms of generation accuracy, local control, and generalization ability, providing an efficient, precise solution for transforming sketches into 3D models.

1. Introduction

3D models lay an important foundation for the construction of digital world, which is widely applied in virtual reality, meta-universe, digital entertainment and other industries. Rapid and efficient 3D model automatic generation technology, as an important research direction, has been drawing increasing attention. As an intuitive and concise approach, sketch-based generation offers a highly creative 3D model generation research field, which focuses on generating imaginative sketches drawn by designers or unprofessional users into 3D models. It significantly lowers the barrier to 3D modeling, users only need to simply sketch the outline of the object to generate a creative and personalized 3D model. By capturing the user's intuitive design intent, sketch-based 3D modeling technology not only simplifies the generation process but also makes the design process more efficient and intuitive. Therefore, it serves as a pivotal tool in broadening access to 3D design practices.

In recent years, with the rapid development of artificial intelligence technologies, generative tasks have made significant progress. Deep learning models, particularly Generative Adversarial Networks [1], Variational Auto-Encoders [2], Diffusion models [3], and Flow-based models [4], have demonstrated remarkable capabilities in tasks such as image, audio, and 3D model generation. However, sketch-based 3D model generation is a challenging task. The inherently abstract and information-sparse nature of hand-drawn sketches, which lack critical information such as depth, surface texture, and high-frequency geometric details, poses substantial difficulties in the precise reconstruction of complex three-dimensional features. Additionally, when sketches contain intricate details or complex intersections, these models often fail to segment and extract key features effectively, resulting in poor generation quality.

Early approaches struggle to represent models with intricate geometric structures effectively. Tanaka et al. [5] restricted the sketch to straight lines, ellipses, and elliptical arcs drawn using 2D Computer Aided Design (CAD) systems, meaning that the reconstructed 3D models were mainly suitable for regular cubic shapes. On the other hand, the sketch retrieval methods [6–8], overlooked the variability in individual sketching styles and user creativity, which often resulted in out of accordance with the user's design intent. Models [9,10] learn latent representations of the input sketch using encoder–decoder architectures to generate and optimize 3D meshes during inference according to the

* Corresponding author.
E-mail addresses: xuzhentong@qdu.edu.cn (Z. Xu), zenglong@sz.tsinghua.edu.cn (L. Zeng), zhaojl@yeah.net (J. Zhao), bdwang@qdu.edu.cn (B. Wang), zkpan@126.com (Z. Pan), liuyongjin@tsinghua.edu.cn (Y.-J. Liu).

https://doi.org/10.1016/j.cad.2025.103896
Received 23 January 2025; Received in revised form 11 April 2025; Accepted 17 May 2025
Available online 24 May 2025

user's design intent to a certain extent. However, these methods are mainly dependent on the contour information of the sketch and the results generated are often rough or lack internal details, producing significant discrepancies compared to real-world objects.

To address the above problems, we propose a method named Sketch123, which represents 3D models with a signed distance field, and design a 3D model reconstruction architecture based on cross-attention of multiple spectral channels. The framework takes sketch information as constraints and interacts with a 3D diffusion model with learned attention to achieve local controllability and better generalization, generating highly smooth 3D models. The main contributions of this work are as follows.

1. We design a sketch reconstruction architecture of a two-stage diffusion model based on optimal transport (OT) guidance to effectively measure the generated 3D data distribution and the real 3D model distribution. This combination ensures that the generated model maintains a high degree of consistency and fidelity of details and global structures. We achieve a stable and accurate sketch reconstruction effect.

2. We propose a multi-spectral channel cross-attention mechanism to adeptly capture hierarchical features across diverse spectral ranges. This mechanism maximizes the complementary strengths of spatial and frequency-based representations by optimizing feature extraction via the synergistic interplay of spatial and frequency characteristics.

3. We introduce a Sparse Feature Enhancement Module (SFE) to replace traditional down-sampling, effectively reducing feature loss. This module balances model performance and feature integrity by preserving crucial information while incorporating Laplace voxel smoothing to maintain detail and reduce noise, ensuring seamless voxel boundaries. Extensive experiments verify the effectiveness of our method.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 proposes a sketch-based 3D model generation method named Sketch123. 3D data representation and two network architectures of Sketch123 are introduced, respectively. Section 4 presents experiments and discussion, and Section 5 concludes.

2. Related works

2.1. Single-view 3D reconstruction

Single-View 3D Reconstruction (SV3D) has been a hot research topic in both computer graphics and computer vision. With the emergence of 3D model datasets like ShapeNetV1 [11], deep learning-based methods for 3D reconstruction have become mainstream [12–15]. In particular, [16–20] based on autoencoder structures train models using 2D images and voxel models as inputs. [19,20] stack two autoencoders and introduce a perceptual fusion module to fuse multi-view information for 3D reconstruction. Other methods [21–23] leverage differentiable rendering techniques. [24,25] build on these approaches by incorporating 2D observations like depth maps, surface normals, and object contours. Furthermore, Duggal et al. [26] uses a depth encoder to predict an initial encoding and applies a depth discriminator for regularization.

The above methods mainly focus on generating 3D models from 2D color images, which have many advantages, such as rich color textures and higher information density. These features provide more clues and details for inference and generation. However, these advantages do not exist in the sketch. Sketch-based 3D reconstruction faces the challenges of lack of geometric perspective reference and information sparsity.

2.2. Sketch-based shape reconstruction

Early research is focused mainly on geometry-based modeling methods [27,28], using operations such as inflation and extrusion to generate 3D models. Optimization-based methods [29–31] further improve the quality of generation and realize interactive operation, but can only represent regular shapes. With the rapid advancement of AI technologies, deep learning-based 3D reconstruction techniques have become a research hotspot. Retrieval-based methods [6–8,32,33] are inherently influenced by training data, resulting in suboptimal generalization capabilities. Generation-based methods [34–40] are becoming mainstream. Chen et al. [34,35] proposed the Deep3DSketch and Deep3DSketch+ methods, which use adversarial learning to reconstruct 3D models from sparse sketch information by randomly sampling 3D shapes and 2D contours. However, GAN models are susceptible to mode collapse. Sketch2Mesh [36] designed an encoder/decoder architecture to convert sketches to 3D models. However, it focuses on matching the external contours of the sketch without paying much attention to the internal details. Sketch3D [40] obtains the image through the control [41] and indirectly generates the mesh. Implicit representation methods [42–44] can be applied to large resolution scenes. SketchDream [44] proposed a text-driven 3D content generation method to generate NeRF from a given sketch. While these methods offer a range of innovative approaches for sketch-to −3D reconstruction, the inherent information sparsity in sketch remains a challenge that needs further exploration.

We propose a novel method for sketch-based 3D model reconstruction that effectively addresses the issue of insufficient information in a single sketch. Combining spatial and frequency domain information, our method efficiently extracts sketch features and reconstructs high-quality 3D mesh, providing an accurate and efficient solution for converting sketches into 3D models.

3. Method

3.1. Overview

Our approach treats the sketch reconstruction task as a conditional generation problem, where a diffusion model is used as the generation component, and sketch features extracted using multi-spectral channel cross-attention are used to guide model generation. To accurately represent the details and shapes of objects, we introduce a high-resolution voxel grid based on the Discrete Signed Distance Function (DSDF). In order to generate a more detailed 3D model, we adopt a two-stage diffusion generation process. The first stage is named the 'Rough Prediction Diffusion Module', which predicts a rough shape shell from a voxel grid initialized with Gaussian noise. The second stage is called the 'Detail Refinement Diffusion Module', which acts as a refinement generator, converting the shape shell obtained in the first step into a high-resolution SDF. Fig. 1 illustrates the overall framework of our sketch-based 3D reconstruction.

3.2. 3D shape representations based on DSDF

The discrete signed distance function (DSDF) is used to represent 3D model in our method. Given a 3D model $\Omega \subset R^3$, the DSDF encodes the closest distance from each point to the model's surface and uses a sign to indicate whether the point lies inside or outside.

According to the properties of DSDF, 3D data represented by DSDF can be converted into a 3D model shell according to the SDF Boundary Function (SBF). It can be defined as:

$$T(z) = \begin{cases} 1, & |d(z)| \leq \Gamma, \\ 0, & otherwise. \end{cases} \tag{1}$$

here, Γ is a threshold, $d(z)$ is the discrete signed distance function. So the set $\Omega_o = \{z \epsilon \mathcal{Z} : T(z) = 1\}$ represents voxel whose center points are within Γ from the surface. Therefore Ω_o serves as an approximation of a thin shell surrounding the 3D model.

For each 3D data, we first normalize the data to lie within the spatial range of $[-0.8, 0.8]^3$. Then, within the space range of $[-1, 1]^3$, we compute a discrete SDF function at a resolution of 128^3 referencing the algorithm in [45]. In order to adapt to the first stage of training, we downsample it to obtain data with a resolution of 64^3. In the 64^3 voxel grid, each voxel inherently contains eight sub-voxels from the 128^3 grid. To achieve a balance between spatial resolution and surface reconstruction precision, we define $\Gamma = 1/32$.

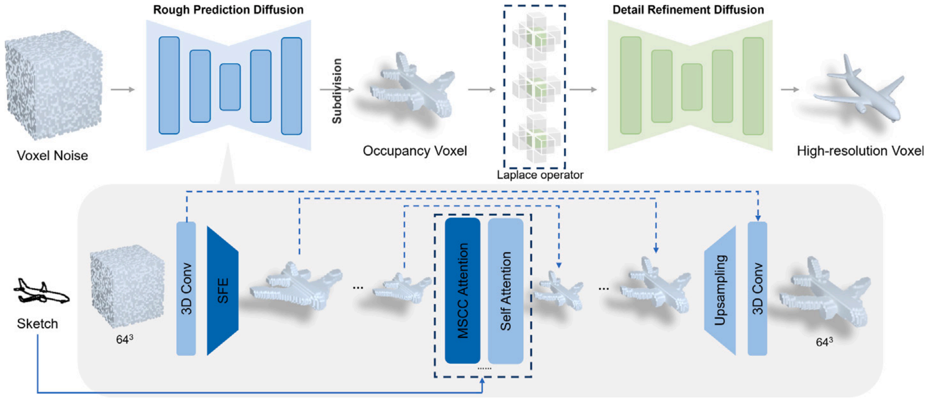


Fig. 1. Pipeline Overview. Sketch123 model includes two sub-processes: “Rough Prediction Diffusion” and “Detail Refinement Diffusion”, which share similar structures and the detailed information is in the gray area. Rough Prediction Diffusion incorporates sketch features into the network through multi-spectral channel cross-attention, enabling interaction with voxel data to generate a coarse voxel grid. The voxel positions are then refined using the Laplacian voxel smoothing operator while introducing random noise. This processed voxel grid serves as the input for Detail Refinement Diffusion, which ultimately predicts the SDF values. In this process, Sparse Feature Enhancement Module (SFE) is employed in place of conventional downsampling to minimize feature loss. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Sketch-based 3D model generation by diffusion model with OT-guidance

We employ the diffusion model as the core framework to generate high-quality 3D models from sketch. The diffusion models consist of the forward process and the reverse process. The forward process gradually adds Gaussian noise to the 3D data, transitioning it to noise. For a 3D model X_0 represented by DSDF, the forward process is simulated by perturbing the recorded SDF values, which is described as a Markov chain. At each time step t, Gaussian noise is added to the 3D model data:

$$q(X_t|X_{t-1}) = \mathcal{N}\left(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t \mathrm{I}\right). \tag{2}$$

Here X_t is the noisy voxel at time step t, and X_{t-1} is the noisy voxel at the previous moment. β_t is the noise scheduling parameter, $\mathbf{I}$ is the identity matrix, and $\mathcal{N}$ denotes the Gaussian distribution. Following the configuration proposed by [46], we define $\beta_t = e^{-10t^2-10^{-4}}$. By recursively applying the above transition probabilities, the joint distribution from X_0 to X_T can be expressed as:

$$q(X_T|X_0) = \prod_{t=1}^{T} q(X_t|X_{t-1}). \tag{3}$$

This process ultimately converts X_0 into standard normal distribution noise. The reverse process aims to reconstruct the SDF values of the 3D data from the noise X_T. This process is also modeled as a Markov chain:

$$p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_T, t), \Sigma_\theta(X_t, t)). \tag{4}$$

Here μ_θ and Σ_θ denote the mean and covariance, respectively, which are typically predicted by a neural network model. To ensure the predicted model converges toward the target 3D model, we incorporate optimal transport theory to quantify the discrepancy between the predicted and real 3D data distributions. The “transport cost” between two distributions is evaluated using the Wasserstein distance. Details are provided in Section 3.5.2.

Our proposed sketch-based 3D model generation network is constructed based on the optimal transport-guided diffusion model as the core and consists of the Rough Prediction Diffusion and the Detail Refinement Diffusion. Implementation details will be introduced below.

3.4. Rough prediction diffusion module

The rough prediction diffusion module is aimed at generating a coarse shape shell of the object from a Gaussian-distributed voxel grid. The detailed implementation is given below.

3.4.1. Network architecture

We employ the U-Net [47] architecture to implement the diffusion model. The U-Net architecture is designed based on a standard 3D convolutional neural network, comprising five layers and incorporating residual connections. Model training is performed using Eq. (4). The specific network architecture is detailed in Fig. 2.

Network Inference: Initially, a 64^3 voxel grid is initialized with Gaussian noise, and the Denoising Diffusion Implicit Models (DDIM) [48] sampling strategy is employed to remove noise over a finite number of steps. In this process, the diffusion model gradually approaches the target 3D model by denoising. To ensure the accuracy of the generated results, a voxel with predicted surface occupancy values greater than 0.5 is retained and then subdivided to achieve a resolution of 128^3.

3.4.2. Sparse feature enhancement of voxel

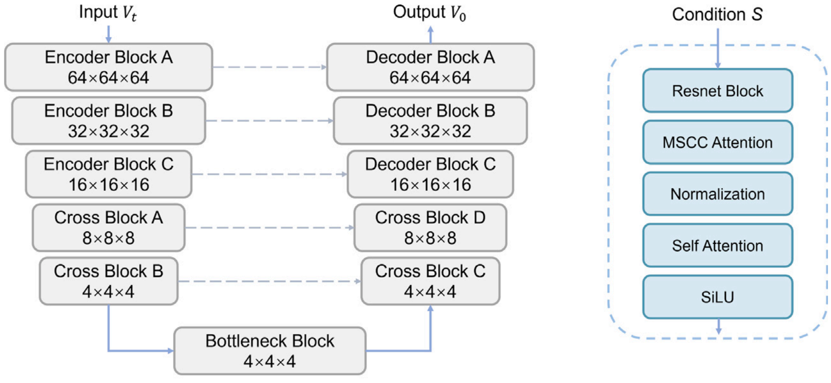
The downsampling process in U-Net often leads to the loss of feature information, which brings more challenges to the voxel features after combining them with a sketch. To tackle this challenge, we propose a novel Sparse Feature Enhancement Module (SFE) to replace classical downsampling operations. The SFE leverages the strengths of depthwise separable convolution and dilated convolution to effectively compress input features while retaining important local and global information.

In the generation process, the voxel f_V is processed in parallel by two branches for feature fusion, followed by feature compression through a Max pooling layer, thereby effectively preserving the sparse features of the voxel. Eq. (5) describes the processing process, *DSConv* represents depthwise separable convolution, *DConv* represents dilated convolution, *MaxPooling* represents maximum pooling processing, and ⊕ represents summation. Fig. 3 illustrates the structure of the SFE module in detail.

$$\tilde{V} = MaxPooling(DSConv(f_V) \oplus DConv(f_V)). \tag{5}$$

3.4.3. Multi-spectrum channel cross attention of sketch and 3D shape

A fundamental challenge in 3D model reconstruction from sketch lies in the sparsity and abstractness of sketch information. Sketch typically contains only contour information, lacking detailed texture and depth information. As a result, the model must efficiently extract core geometric features from the sketch. Unlike previous methods focusing solely on spatial and channel attention, we propose a novel approach exploring feature extraction from the frequency domain. Specifically, we introduce multiple frequency components of the Discrete Cosine Transform (DCT) into channel attention to enhance the expressive power of sketch feature compression, which we call the Multi-Spectral



(a)Diffusion's U-net architecture

(b)Cross Block architecture

Fig. 2. U-Net architecture. (a) For a noise voxel with a resolution of 64, the output resolutions after each layer are 64^3, 32^3, 16^3, 8^3 and 4^3. The corresponding feature dimensions for each layer are 32, 64, 128, 256 and 256. (b) The attention mechanism we proposed is used at the bottom layer of U-Net.

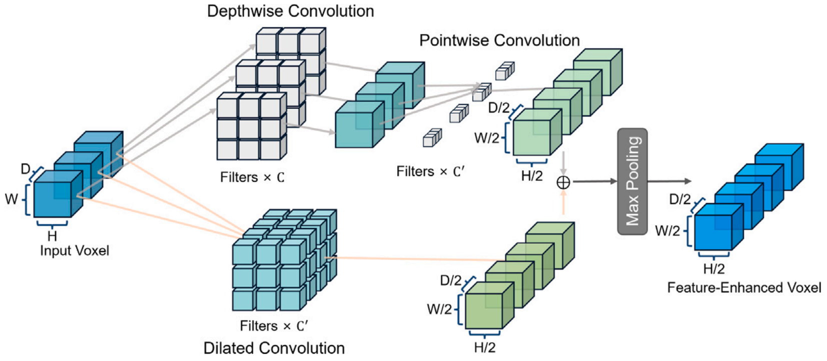


Fig. 3. Sparse Feature Enhancement Module (SFE). The SFE consists of two branches: Depthwise separable convolution to extract local features of the voxel, and Dilated convolution to capture global structural information by expanding the receptive field.

Channel Cross-Attention (MSCC-Attention). As shown in Fig. 4, for any voxel feature f_V, its center point is projected into the camera coordinate system based on the sketch view information. To eliminate ambiguities in projection point, a perspective projection normalization operation is applied, guided by voxel depth information enabling the interaction between 3D voxel features and 2D pixel features. The proposed Multi-Spectral Channel Cross-Attention Mechanism consists of three core components:

Sketch Spatial Attention Module. This module is used to accurately extract location information. Focus the model's attention on the most important areas in the sketch. The spatial attention can be formulated as:

$$att = sigmoid\big(conv(S)\big), \tag{6}$$

where *att* is the spatial attention vector, S represents the sketch, which are resized to 224 × 224 pixels before input, $conv(\bullet)$ denotes the convolution operation, and $sigmoid(\bullet)$ is the Sigmoid function. We input S into the module. After obtaining the attention vector, S is scaled by the corresponding attention value, result in the scaled output $\widetilde{S}$:

$$\tilde{S} = att \cdot S. \tag{7}$$

Spectral Channel Attention Module. This module aims to capture the response components of various channels to the salient regions of the sketch feature. As demonstrated in [49], Global Average Pooling (GAP) is, in fact, a special case of the DCT, with its output proportional to the lowest-frequency components of a 2D DCT. Therefore, the GAP operation used in conventional channel attention utilizes only the lowest frequency component of DCT. We compress information from the sketch through multiple frequency components of the 2D discrete cosine transform. This module is designed to capture the characteristic response components in salient areas at different channel frequencies. The basis function of 2D DCT is:

$$B^{i,j}_{h,w} = COS\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)COS\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right). \tag{8}$$

Then the 2D DCT can be written as:

$$\begin{aligned} DCT(S) &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} S \cdot B^{i,j}_{h,w}; \\ s.t.\ h &\in \{0,1,\ldots,H-1\}\ ;\ w \in \{0,1,\ldots,W-1\}; \\ i &\in \{0,1,\ldots,H-1\}\ ;\ j \in \{0,1,\ldots,W-1\}. \end{aligned} \tag{9}$$

Where H and W are the height and width of the input sketch S, i, j represent the pixel positions in the image space, and h, w represent the frequency indices of the basis functions, respectively. S is divided into multiple parts along the channel dimension, each part will be assigned a DCT frequency component, and the calculated result is used as the attention of the channel. The Spectral Channel Attention can be described as:

$$SCA(S) = sigmoid(Liner(ReLU(Liner(DCT(S))))) \otimes S, \tag{10}$$

where *sigmoid* is the Sigmoid function, *Liner* represents the linear layer, *ReLU* is the ReLU activation function, and *DCT* is the 2D DCT.

Perceptual Feature Cross Attention. This module aims to combine the outputs of the two preceding sub-modules into the final sketch feature representation. We employ a multi-head attention mechanism [50] to facilitate interaction between the voxel features and the sketch image patch features based on the known view projection relationship. We consider that the SDF value recorded at any voxel position is only affected by a small range of the corresponding sketch image, so during the interaction, the image features are divided into feature blocks with

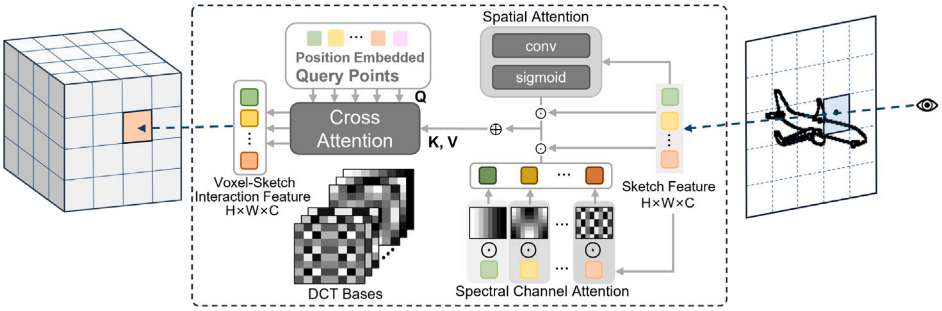


Fig. 4. Illustration of our Multi-Spectrum Channel Cross Attention mechanism. Each voxel corresponds to a specific image feature region, and the proposed attention mechanism facilitates interaction between voxel and image features.

a patch width of 14. The combined sketch features can be expressed as:

$$f_S = \tilde{S} \oplus SCA(S). \tag{11}$$

Finally, we compute the query Q, key K, and value V as follows. The voxel features f_v^{new} are obtained through Eq. (13), where M denotes the mask for attention computation introduced by view projection.

$$Q = f_V W^Q, K = f_S W^K, V = f_S W^V, \tag{12}$$

$$f_v^{new} = Attention(Q, K, V, M). \tag{13}$$

3.5. *Detail refinement diffusion module*

The detail refinement diffusion module focuses on refining the rough model generated in the first stage. Its structure is similar to that of rough prediction diffusion. The process begins from initializing the rough prediction diffusion output with Gaussian noise. Next, the DDIM sampling strategy is applied to iteratively denoise the voxel data. After completion of the denoising process, the Marching Cubes algorithm [51] is used to convert the voxel representation into a mesh output.

3.5.1. *Laplace operator smoothing of voxel model*

Laplace smoothing is applied to voxel data. This reduces noise and irregularities by smoothing the SDF values. Given a vertex v_i, its Laplace voxel smoothing operator Δ_{v_i} can be defined as:

$$\Delta_{v_i} = \sum_{j\in N(i)} \frac{1}{|N(i)|}(v_j - v_i), \tag{14}$$

where *N(i)* represents the set of one-neighboring voxel vertices of v_i, v_j is a neighboring voxel vertex, and $|N(i)|$ is the number of neighboring voxel vertices, the size of the set *N(i)*.

The Laplace voxel smoothing operator calculates the average difference between a voxel vertex and its neighbors. Similar to smoothing on 3D meshes, this operator smooths the voxel space by adjusting each vertex's position toward the average of its neighbors, effectively reducing noise while preserving the geometric consistency of the voxel model.

3.5.2. *Wasserstein distance constraint of generated shape*

As introduced in Section 3.3, We regard the predicted SDF and the true SDF as two distributions $X, Y \subset R^3$, with probability densities μ and ν respectively, and the total measure is equal:

$$\int_{\mathrm{X}} \mu = \int_{\mathrm{Y}} \nu. \tag{15}$$

By defining a cost function $c(x, y)$, $x \in X$, $y \in Y$, we hope to find a transport mapping that minimizes the cost of transport. That is the optimal transport mapping $\mathrm{T} : \mathrm{X} \rightarrow \mathrm{Y}$:

$$T = {}^{min}_{\mathrm{T}_{\#}\mu=\nu} E_{X\sim\mu} c(X, T(X)), \tag{16}$$

here, $\mathrm{T}_{\#}\mu = \nu$ indicates $\int_X \mu(x)dx = \int_{T(X)} \nu(y)dy$ for every measurable X, Y. We choose the cost as $c(x, y) = \|x - y\|^2$, then the Wasserstein Distance is defined as:

$$W(\mu, \nu) = [inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)d\gamma(x, y)]^{\frac{1}{2}}, \tag{17}$$

here, $\gamma \in \prod(\mu, \nu)$ represents the joint probability density of μ and ν. We use the Sinkhorn Algorithm [52] to calculate the Wasserstein Distance. First, the transition matrix P is initialized using a uniform distribution. The row and column scaling factors a and b are updated alternately. Finally, the transition matrix is updated by $P = diag(a)Mdiag(b)$ until convergence or the predetermined number of iterations is reached, here $M = c(x, y)$, $diag()$ indicate as Diagonal Matrix. Then the Wasserstein loss can be defined as:

$$\sum_{\forall i,j} M_{ij} P_{ij} + \lambda P_{ij} \log P_{ij}, \tag{18}$$

here λ is the Entropy regularization parameter, we set $\lambda = 7$. The introduction of Wasserstein Distance enables the network to consider the prediction of SDF from a global perspective.

3.6. *LOSS function*

To obtain a realistic 3D model, we designed a novel loss function that consists of three components:

SDF Loss: This loss directly constrains the generation of SDF values. To ensure that the generated model closely aligns with the expected geometry, we define the SDF loss in Eq. (19), where *MSE* refers to the Mean Squared Error between the $pred_{sdf}$ and the $true_{sdf}$.

$$SDF_{loss} = MSE\left(pred_{sdf}, true_{sdf}\right). \tag{19}$$

Wasserstein Loss: This loss is based on Optimal Transport. As described in Section 3.5.2, minimizing the Wasserstein distance between the SDF distributions of the target and source domains enhances the global consistency and fidelity of the generated 3D model. It effectively removes connectivity issues inherent in voxel-based generation. We implement it in the network, we define:

$$Wasserstein_{loss} = sinkhorn\left(pred_{sdf}, true_{sdf}, M, \lambda\right). \tag{20}$$

Boundary Loss: A key property of SDF is that zero values indicate the surface of the object. To reinforce this characteristic, we designed a boundary loss to improve prediction accuracy near the object's surface. The values within $threshold = 0.1$ are considered part of the surface boundary. We define:

$$mask = ABS\left(true_{sdf}\right) < threshold, \tag{21}$$

$$Boundary_{loss} = MSE\left(pred_{sdf}(mask), true_{sdf}(mask)\right). \tag{22}$$

The overall loss for the model is a weighted combination of the three components:

$$loss = \lambda_1 SDF_{loss} + \lambda_2 Wasserstein_{loss} + \lambda_3 Boundary_{loss}. \tag{23}$$

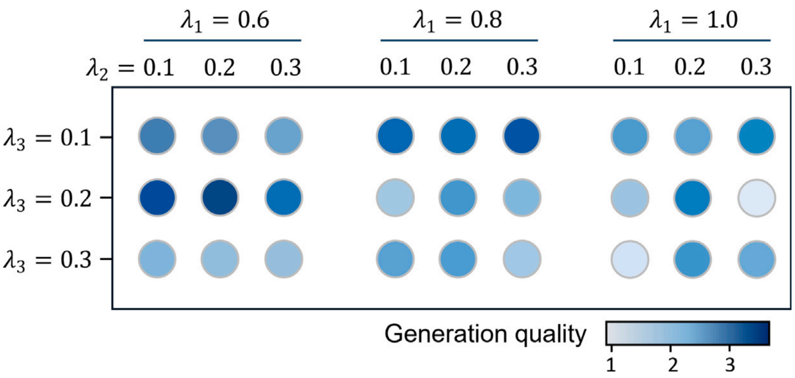


Fig. 5. The evaluation of loss weight coefficients is conducted based on the generation quality, which is measured using the IoU metric. The color intensity is positively correlated with the generation quality.

To set the loss weight coefficients reasonably, we employed a grid search strategy to define the value sets for the 3 weights as $\lambda_1 \in \{0.6, 0.8, 1.0\}$, $\lambda_2 \in \{0.1, 0.2, 0.3\}$, and $\lambda_3 \in \{0.1, 0.2, 0.3\}$, generating 27 combinations. We compared the generated quality one by one and selected the best weight combination. The results are shown in Fig. 5. Therefore, we set $\lambda_1 = 0.6$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.2$.

4. Experiments

In this section, we exhibit and validate the capability of Sketch123 for sketch-conditioned shape generation.

Data Preparation: We use ShapeNetV1 [11] and our craniofacial dataset to train Sketch123. The following four categories are incorporated into the analytical framework: chair, car, table, and rifle. For each data, we rendered projection images from five predefined different viewpoints and used a Canny edge detector [53] to extract contours, which served as the synthetic sketch. The viewpoints are set at five angles around the model's center on a horizontal plane: −90°, −45°, 0°, 45° and 90°.

Training Details: For the Rough Prediction Diffusion, we used a learning rate of 1.25×10^{-4} and trained for 300 epochs. For Detail Refinement Diffusion, we used the AdamW optimizer [54] with a fixed learning rate of 10^{-4} over 500 epochs. The data split followed a ratio of 7:1:2 for training, validation and testing, respectively.

Inference Efficiency: Experimental measurements conducted on an NVIDIA GeForce RTX 4090 demonstrate that the implemented Denoising Diffusion Implicit Models (DDIM) sampling strategy with 50 iterative steps achieves an average inference time of 10 s.

4.1. Model evaluation

Both the qualitative and quantitative assessment protocols were systematically employed. The evaluation metrics used include Chamfer Distance (CD), Normal Consistency (NC), F1 Score, voxel-Intersection over Union (IoU), and Contrastive Language-Image Pre-training Score (CLIP-Score) [55]. Here, we employ CLIP-Score to evaluate the cosine similarity of sketch features. For the sketch S_{in}, we render the sketch S_{out} from the same perspective using the mesh obtained from it. Consequently, it can be expressed as follows:

$$\text{CLIP-Score} = CLIP(S_{in}, S_{out}). \tag{24}$$

4.1.1. Evaluation details

We compared our method with the following approaches: Sketch2Model [56], Sketch2Mesh [9], SketchSampler [57], LAS-Diffusion [39] and SENS [38]. To ensure accurate evaluation, the predicted mesh was scaled to align with the size of the target mesh. Sketch2Model, Sketch2Mesh and SENS were provided with appropriate sketches that match their respective training style. LAS-Diffusion was given accurate sketch viewpoint information. Since SketchSampler only generates point clouds, its output was converted to mesh using Shape-As-Points [58] for evaluation.

Table 1
Quantitative evaluations on chairs and cars. The units of CD, NC, F1 Score, and IoU are 10^{-3}, 10^{-2}, 10^{-2}, and 10^{-2}, respectively.

Method	Chairs				
	CD↓	NC↑	F1 Score↑	IoU↑	CLIP-Score↑
Sketch2Model [56]	20.5	62.3	45.7	30.6	85.5
Sketch2Mesh [9]	48.7	54.1	63.2	47.2	91.5
SketchSampler [57]	32.0	59.5	53.9	48.7	92.4
LAS-Diffusion [39]	13.1	**77.8**	63.0	49.2	94.6
SENS [38]	11.0	63.6	46.6	32.2	87.9
ours	**10.7**	76.3	**68.2**	**58.4**	**96.3**
	Cars				
Sketch2Model [56]	18.1	66.5	47.3	49.9	88.3
Sketch2Mesh [9]	54.1	58.6	69.4	53.1	88.7
SketchSampler [57]	38.6	57.2	59.7	51.6	94.7
LAS-Diffusion [39]	12.7	72.6	67.4	49.8	96.4
SENS [38]	11.4	67.9	44.5	38.6	85.0
ours	**11.1**	**82.4**	**72.6**	**57.0**	**97.1**

As quantitatively demonstrated in Table 1, the proposed method exhibits superior performance across multiple evaluation metrics. It is noteworthy that the normal consistency metric in chairs remains suboptimal compared to LAS-Diffusion. This limitation is attributed to the Laplacian voxel smoothing Operator's inherent trade-off: it perturbs surface normal vectors through a smoothing kernel. The qualitative assessment in Fig. 7 provides visual validation of these findings. When reconstructing chair geometries, our framework demonstrates preservation of structural integrity while achieving superior detail fidelity. Fig. 6 presents a computational efficiency analysis to complement these results. Our proposed method achieves an inference time of approximately 10 s, in the middle range among the compared models. Notably, despite its moderate inference time, our method consistently generates high-fidelity 3D models.

4.1.2. Detail controllability test

Our method demonstrates excellent detail extraction capabilities through the proposed multi-spectral channel attention mechanism. This mechanism shows a strong adaptability to small structural changes in the input sketch. Fig. 8 specifically demonstrates this effect, where even subtle modifications in the sketch lead to corresponding adjustments in the generated 3D model. The results emphasize the robustness of our approach in accurately reflecting intricate sketch details while maintaining overall structural consistency.

4.1.3. Model generalization test

In this test, the generalizability of the model is evaluated from two perspectives: hand-drawn sketch and sparse sketch.

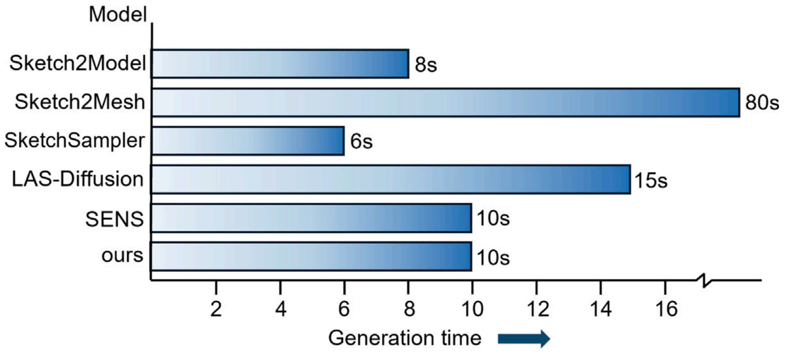


Fig. 6. Time Comparison. The proposed method demonstrates the capability to synthesize realistic 3D models while maintaining computational efficiency. Notably, the computational time metric for Sketch2Model incorporates both the iterative optimization and the mesh generation.

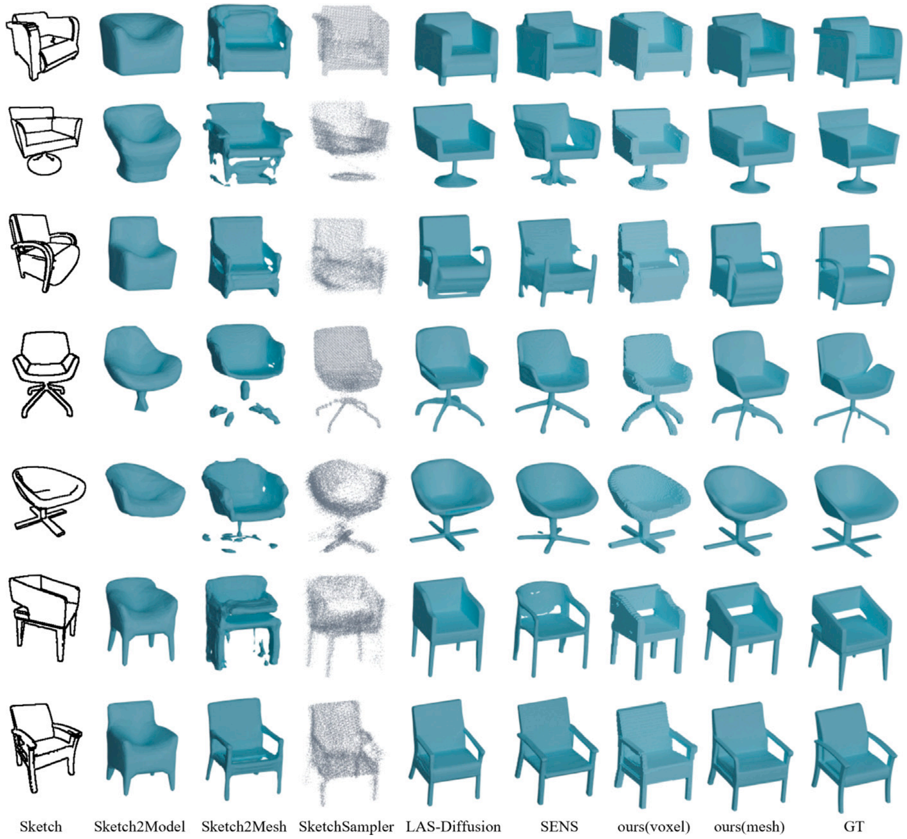


Fig. 7. Generation result quality comparison. From left to right are: Sketch, Sketch2Model, Sketch2Mesh, SketchSampler, LAS-Diffusion, ours(voxel), ours(mesh), and the ground truth. Ours(voxel) represents the voxel data without post-processing.

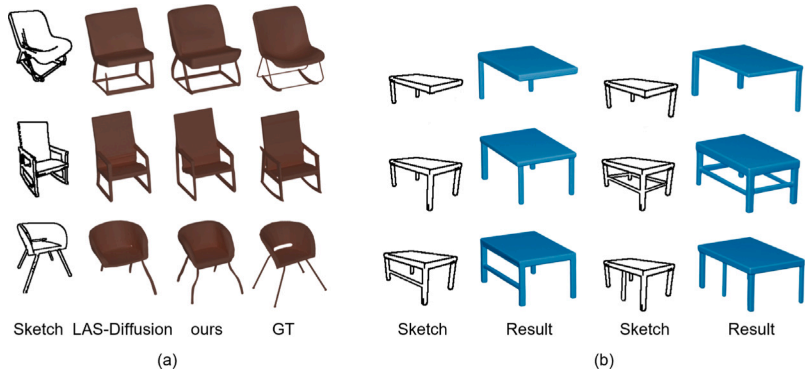


Fig. 8. Detail controllability test. (a) Our method can generate good results for easily overlooked and complex lines in sketch. (b) Adding different lines to a table sketch, our results demonstrate the accuracy of detail changes.

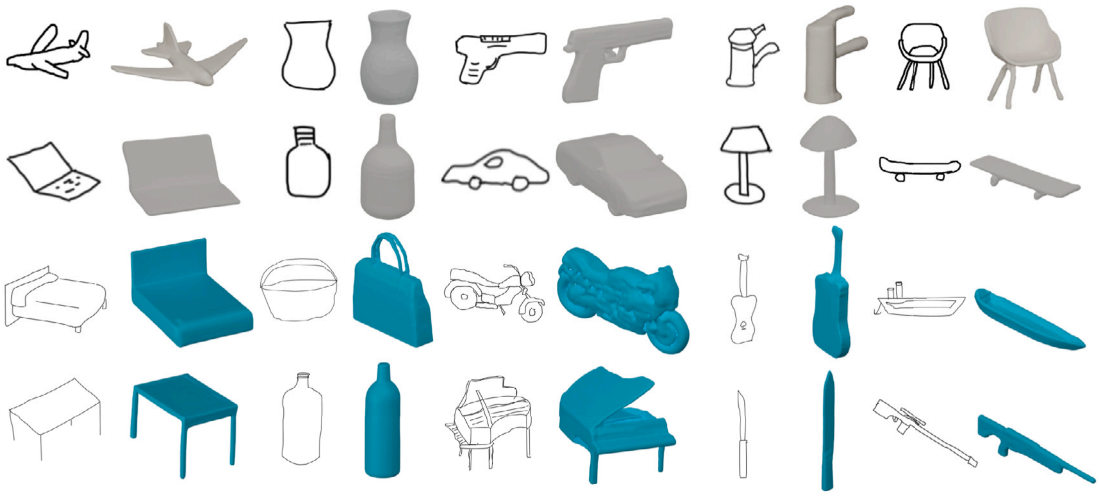


Fig. 9. Our Sketch123 has the ability to adapt to hand-drawn sketch. The first two rows of hand-drawn sketches are from the manual collection and the rest are from the TU Berlin Sketch Dataset [59].

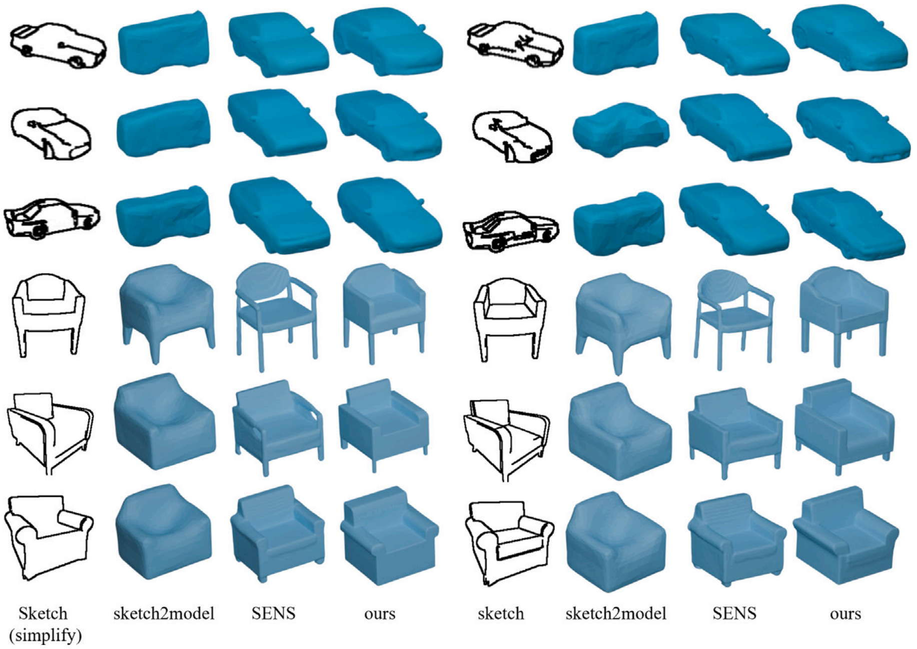


Fig. 10. Sparse Sketch test. The test samples feature simplified lines, and our results demonstrate minimal deviation from those generated with the original sketch.

Hand-drawn Sketch: Through our proposed multi-spectral channel cross-attention mechanism, features in the frequency domain and spatial domain are extracted, which effectively reduces the dependence on line smoothness and enables the generation of satisfactory 3D models from hand-drawn sketch. Fig. 9 illustrates the results generated for the hand-drawn sketch.

Sparse Sketch: Sketch lines are categorized into two types: structural lines, which define the skeletal framework of the object, and non-structural lines, which represent decorative or supplementary elements. Random line removal strategies are considered problematic, as they may eliminate critical structural lines that encode essential object features. Therefore, we strategically remove non-structural lines to get sparse sketches.

To evaluate the robustness of our method under informationally constrained conditions, we conducted a controlled study by strategically removing critical lines from the original sketches. As demonstrated in Fig. 10, the proposed method exhibits remarkable structural preservation capabilities.

4.1.4. Model scalability test

A single sketch123 model was trained on 50 categories, utilizing data from the ShapeNetV1 dataset and a custom craniofacial dataset. Despite the limited data available in certain categories, the model consistently exhibited robust performance. This indicates that the model possesses strong learning and generalization capabilities, even when dealing with small sample sizes. As shown in Fig. 11, the generation results for several categories highlight the ability of the model to maintain high-quality output across diverse and complex 3D model types.

4.2. Ablation studies

In this section, in order to more clearly understand the influence of different parts of our model on sketch reconstruction. We performed ablation studies to validate the effectiveness of the proposed method. We successively disabled or replaced key components such as the

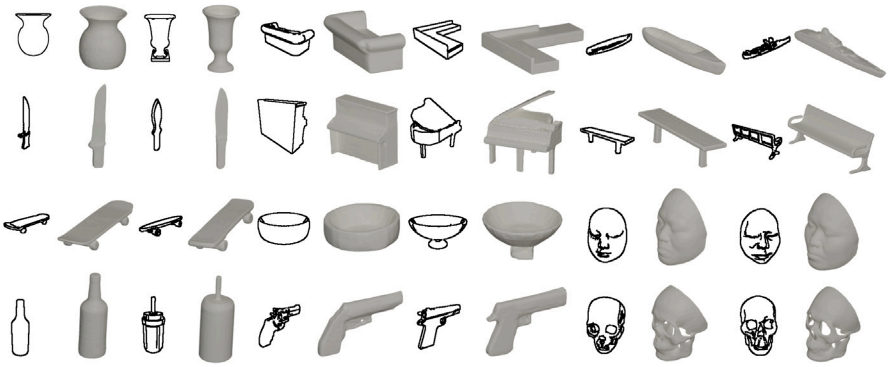


Fig. 11. Visualization results of some training categories.

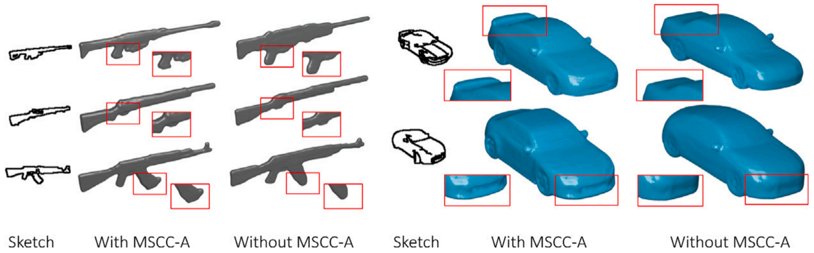


Fig. 12. Ablation study of our proposed MSCC-Attention module.

Table 2
An ablation study on the impact of different components of our method on generation.

	CD↓	NC↑	F1 Score↑	IoU↑
Without MSCC-Attention	13.1	67.8	63.0	49.2
Without Laplace	**10.7**	76.2	67.5	52.2
Without SFE	12.7	71.9	66.3	49.9
Without Boundary Loss	10.8	68.6	64.3	48.1
Without Wasserstein Loss	**10.7**	**76.3**	64.1	50.5
Full Model	**10.7**	**76.3**	**68.2**	**58.4**

Table 3
The performance of the multi-spectral channel attention mechanism is evaluated by configuring different frequency components.

	CD↓	F1 Score↑	IoU↑
Low-Frequency Dominance	**10.7**	**76.2**	**58.4**
Mid-Frequency Dominance	11.6	68.2	53.3
High-Frequency Dominance	11.3	65.1	50.1
GAP	**10.7**	66.7	50.6

MSCC-Attention mechanism, the Laplace voxel smoothing operator, and the SFE. Notably, the Laplace voxel smoothing operator analyzed in Section 4.1.1 may introduce negative artifacts. As evidenced in Table 2, the operator significantly improves the IoU score, indicating its strong positive influence on voxel distribution. Laplacian voxel smoothing operator effectively enhances spatial coherence and overall reconstruction quality.

In Fig. 12, we report the performance without using MSCC-Attention and instead use a normal attention mechanism. The results indicate a decline in generation quality. This validates the significant advantages of the proposed multi-spectral strategy.

We defined three frequency bands: Low-Frequency Dominance, Mid-Frequency Dominance, and High-Frequency Dominance. Additionally, we replaced the component with GAP to comprehensively evaluation its contributions to model performance across different frequency domains. From Table 3 we can see that Low-Frequency Dominance achieve the best result, aligning with the principle that GAP functions as a DCT low-frequency information extractor, as described in Section 3.4.3.

5. Conclusion

In this paper, we proposed a 3D reconstruction method that integrates a multi-spectral channel cross-attention mechanism with a diffusion model, enabling the generation of 3D models from sketches. The multi-spectral channel cross-attention mechanism facilitates the extraction of sketch features in both frequency and spatial domains, improving the model's capacity to interpret abstract sketch information. The Sketch123 effectively addresses issues of incomplete sketch information and missing details, resulting in high-quality, shape-controllable 3D model generation. Experimental results indicate that our method achieves superior performance across various sketch input tasks, surpassing existing mainstream methods in generation accuracy, Sketch perception, and generalization capability.

A significant limitation of this study is that the performance of the model lies in its dependency on the consistency of the sketch style. We argue that optimizing the sketch lines in future work could help bridge the gap between the training sketch style and the real-world sketches. Developing a Style-Invariant Feature Extraction module capable of isolating structural primitives (e.g., edge connectivity) while suppressing stylistic artifacts (e.g., stroke thickness variations) through feature disentanglement is crucial. On the other hand, an implementation of a bidirectional interface enables real-time user corrections to both input sketches and output geometries, thereby iteratively aligning the latent space with human perceptual priors.

CRediT authorship contribution statement

Zhentong Xu: Writing – original draft, Validation, Software. **Long Zeng:** Writing – review & editing. **Junli Zhao:** Writing – review & editing, Resources, Project administration, Methodology. **Baodong Wang:** Writing – review & editing. **Zhenkuan Pan:** Writing – review & editing, Supervision. **Yong-Jin Liu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully appreciated the anonymous reviewers for all of their helpful comments. This research is supported in parts by the National Natural Science Foundation of China (Nos.62172247, 62461160309, and 61772294), National Key Research and Development Program of China (Grant No. 2022YFB3303101), NSFC-RGC Joint Research Scheme N_HKU705/24, and Natural Science Foundation of Shandong Province (No. ZR2024MF087).

Data availability

Data will be made available on request.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014;27.

[2] Kingma DP. Auto-encoding variational bayes. Int Conf Learn Represent 2014.

[3] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 2020;33:6840–51.

[4] Rezende D, Mohamed S. Variational inference with normalizing flows. In: International conference on machine learning. PMLR; 2015, p. 1530–8.

[5] Tanaka M, Terano M, Higashino C, Asano T, Takasugi K. A learning method for reconstructing 3D models from sketches. Computer- Aided Des Appl 2019;16(6):1158–70.

[6] Zeng L, Dong Z-k, Yu J-y, Hong J, Wang H-y. Sketch-based retrieval and instantiation of parametric parts. Computer- Aided Des 2019;113:82–95.

[7] Yang H, Tian Y, Yang C, Wang Z, Wang L, Li H. Sequential learning for sketch-based 3D model retrieval. Multimedia Syst 2022;1–18.

[8] Kosalaraman KK, Kendre PP, Manilal RD, Muthuganapathy R. SketchCleanGAN: A generative network to enhance and correct query sketches for improving 3D CAD model retrieval systems. Comput Graph 2024.

[9] Guillard B, Remelli E, Yvernay P, Fua P. Sketch2mesh: Reconstructing and editing 3d shapes from sketches. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 13023–32.

[10] Chen T, Ding C, Zhu L, Zang Y, Liao Y, Li Z, et al. Reality3dsketch: Rapid 3d modeling of objects from single freehand sketches. IEEE Trans Multimed 2023.

[11] Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, et al. Shapenet: An information-rich 3d model repository. 2015, arXiv preprint arXiv:1512.03012.

[12] Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang Y-G. Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European conference on computer vision. 2018, p. 52–67.

[13] Xu Q, Wang W, Ceylan D, Mech R, Neumann U. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. Adv Neural Inf Process Syst 2019;32.

[14] Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 165–74.

[15] Chen Z, Zhang H. Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 5939–48.

[16] Yan X, Yang J, Yumer E, Guo Y, Lee H. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Adv Neural Inf Process Syst 2016;29.

[17] Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. Adv Neural Inf Process Syst 2016;29.

[18] Smith EJ, Meger D. Improved adversarial systems for 3d object generation and reconstruction. In: Conference on robot learning. PMLR; 2017, p. 87–96.

[19] Xie H, Yao H, Sun X, Zhou S, Zhang S. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 2690–8.

[20] Xie H, Yao H, Zhang S, Zhou S, Sun W. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. Int J Comput Vis 2020;128(12):2919–35.

[21] Henderson P, Ferrari V. Learning to generate and reconstruct 3d meshes with only 2d supervision. 2018, arXiv preprint arXiv:1807.09259.

[22] Lin C-H, Wang C, Lucey S. Sdf-srn: Learning signed distance 3d object reconstruction from static images. Adv Neural Inf Process Syst 2020;33:11453–64.

[23] Liu S, Li T, Chen W, Li H. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 7708–17.

[24] Liu S, Zhang Y, Peng S, Shi B, Pollefeys M, Cui Z. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 2019–28.

[25] Wu J, Wang Y, Xue T, Sun X, Freeman B, Tenenbaum J. Marrnet: 3d shape reconstruction via 2.5 d sketches. Adv Neural Inf Process Syst 2017;30.

[26] Duggal S, Wang Z, Ma W-C, Manivasagam S, Liang J, Wang S, et al. Mending neural implicit modeling for 3d vehicle reconstruction in the wild. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022, p. 1900–9.

[27] Igarashi T, Matsuoka S, Tanaka H. Teddy: a sketching interface for 3D freeform design. In: ACM SIGGRAPH 2006 courses. 2006, p. 11–es.

[28] Dvorožňák M, Sỳkora D, Curtis C, Curless B, Sorkine-Hornung O, Salesin D. Monster mash: a single-view approach to casual 3D modeling and animation. ACM Trans Graph (ToG) 2020;39(6):1–12.

[29] Bobenrieth C, Cordier F, Habibi A, Seo H. Descriptive: Interactive 3D shape modeling from a single descriptive sketch. Computer- Aided Des 2020;128:102904.

[30] Tian C, Masry M, Lipson H. Physical sketching: Reconstruction and analysis of 3D objects from freehand sketches. Computer- Aided Des 2009;41(3):147–58.

[31] Zeng L, Liu Y-J, Lee SH, Yuen MM-F. Q-complex: efficient non-manifold boundary representation with inclusion topology. Computer-Aided Des 2012;44(11):1115–26.

[32] Sangkloy P, Burnell N, Ham C, Hays J. The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans Graph (ToG) 2016;35(4):1–12.

[33] Ban S, Hyun KH. 3D computational sketch synthesis framework: Assisting design exploration through generating variations of user input sketch and interactive 3D model reconstruction. Computer- Aided Des 2020;120:102789.

[34] Chen T, Fu C, Zhu L, Mao P, Zhang J, Zang Y, et al. Deep3dsketch: 3D modeling from free-hand sketches with view-and structural-aware adversarial training. 2023, arXiv preprint arXiv:2312.04435.

[35] Chen T, Fu C, Zang Y, Zhu L, Zhang J, Mao P, et al. Deep3dsketch+: Rapid 3d modeling from single free-hand sketches. In: International conference on multimedia modeling. Springer; 2023, p. 16–28.

[36] Guillard B, Remelli E, Yvernay P, Fua P. Sketch2mesh: Reconstructing and editing 3d shapes from sketches. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 13023–32.

[37] He Y, Xie H, Miyata K. Sketch2cloth: Sketch-based 3d garment generation with unsigned distance fields. In: 2023 nicograph international (NicoInt). IEEE; 2023, p. 38–45.

[38] Binninger A, Hertz A, Sorkine-Hornung O, Cohen-Or D, Giryes R. SENS: Part-aware sketch-based implicit neural shape modeling. Comput Graph Forum (Proc EUROGRAPHICS 2024) 2024;43(2).

[39] Zheng X-Y, Pan H, Wang P-S, Tong X, Liu Y, Shum H-Y. Locally attentional sdf diffusion for controllable 3D shape generation. ACM Trans Graph (ToG) 2023;42(4):1–13.

[40] Zheng W, Xia H, Chen R, Sun L, Shao M, Xia S, et al. Sketch3D: Style-consistent guidance for sketch-to-3D generation. In: Proceedings of the 32nd ACM international conference on multimedia. 2024, p. 3617–26.

[41] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 3836–47.

[42] Deng K, Yang G, Ramanan D, Zhu J-Y. 3D-aware conditional image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 4434–45.

[43] Gao L, Liu F-L, Chen S-Y, Jiang K, Li C, Lai Y, et al. SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields. ACM Trans Graph 2023;42(4).

[44] Liu F-L, Fu H, Lai Y-K, Gao L. Sketchdream: Sketch-based text-to-3D generation and editing. ACM Trans Graph (TOG) 2024;43(4):1–13.

[45] Xu H, Barbič J. Signed distance fields for polygon soup meshes. In: Graphics interface 2014. AK Peters/CRC Press; 2020, p. 35–41.

[46] Kingma D, Salimans T, Poole B, Ho J. Variational diffusion models. Adv Neural Inf Process Syst 2021;34:21696–707.

[47] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015, p. 234–41.

[48] Song J, Meng C, Ermon S. Denoising diffusion implicit models. 2020, arXiv preprint arXiv:2010.02502.

[49] Qin Z, Zhang P, Wu F, Li X. Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 783–92.

[50] Vaswani A. Attention is all you need. Adv Neural Inf Process Syst 2017.

[51] Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field. 1998, p. 347–53.

[52] Bregman LM. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput Math Math Phys 1967;7(3):200–17.

[53] Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 1986;(6):679–98.

[54] Loshchilov I. Decoupled weight decay regularization. 2017, arXiv preprint arXiv:1711.05101.

[55] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021, p. 8748–63.

[56] Zhang S-H, Guo Y-C, Gu Q-W. Sketch2model: View-aware 3d modeling from single free-hand sketches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 6012–21.

[57] Gao C, Yu Q, Sheng L, Song Y-Z, Xu D. Sketchsampler: Sketch-based 3d reconstruction via view-dependent depth sampling. In: European conference on computer vision. Springer; 2022, p. 464–79.

[58] Peng S, Jiang C, Liao Y, Niemeyer M, Pollefeys M, Geiger A. Shape as points: A differentiable poisson solver. Adv Neural Inf Process Syst 2021;34:13032–44.

[59] Eitz M, Hays J, Alexa M. How do humans sketch objects? ACM Trans Graph (Proc SIGGRAPH) 2012;31(4):44:1–44:10.