Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Locality-Preserving Sketch Token Fusion for High-Fidelity Single-View 3D Reconstruction

Lei Qinchuan
Faculty of Innovation Engineering, Macau University of Science and Technology
Li Nannan
nnli@must.edu.mo
Faculty of Innovation Engineering, Macau University of Science and Technology
Zhong Jiaming
Guangdong University of Science and Technology
Luo XianZeng
Guangdong University of Science and Technology
Zhao Min
Guangdong University of Science and Technology

Article

Keywords: Sketch-based 3D reconstruction, sepSingle-view 3D modeling, sepLocal feature learning, sepToken fusion, sepCross-attention

Posted Date: October 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11022086/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Locality-Preserving Sketch Token Fusion for High-Fidelity Single-View 3D Reconstruction

Lei Qinchuan[1,2], Li Nannan[1*], Zhong Jiaming[3], Luo XianZeng[2], Zhao Min[3]

[1*]School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Taipa, 999078, Macau, China.
[2]School of Computer Science and Artificial Intelligence (Industrial Software), Guangdong University of Science and Technology, Dongguan, Dongguan, 523000, Guangdong, China.
[3]School of Intelligent Manufacturing, Guangdong University of Science and Technology, Dongguan, Dongguan, 523000, Guangdong, China.

*Corresponding author(s). E-mail(s): nnli@must.edu.mo;
Contributing authors: 3230002357@student.must.edu.mo;
zhongjiaming1@gdust.edu.cn; 404990026@qq.com;
zhaomin@gdust.edu.cn;

Abstract

Reconstructing 3D shapes from a single freehand sketch remains challenging due to the ambiguity of hand-drawn strokes and the loss of fine-grained geometric details during feature encoding. Existing approaches typically compress sketch features into global latent representations through flattening or pooling operations, which effectively capture coarse shape semantics but inevitably discard spatially localized stroke cues that are critical for recovering detailed geometry. In this work, we propose Locality-Preserving Sketch Token Fusion, a novel framework that maintains and exploits local sketch information for high-fidelity single-view 3D reconstruction. Instead of collapsing spatial features into a single vector, our method introduces a dual-path sketch representation that simultaneously learns a global latent representation for shape and viewpoint modeling and a spatially structured token representation for preserving local geometric patterns. Furthermore, we design a local-token cross-attention mesh sampler, where the global shape representation adaptively queries local sketch tokens during mesh decoding, enabling precise alignment between 2D stroke structures and 3D

vertex deformation. To handle inherent viewpoint ambiguity, we incorporate a view-disentanglement branch that improves robustness under varying observation conditions.Extensive experiments on ShapeNet-Synthetic and ShapeNet-Sketch demonstrate that our approach achieves superior reconstruction accuracy compared with existing methods, producing geometrically faithful 3D shapes with improved preservation of thin structures and local details.

Keywords: Sketch-based 3D reconstruction Single-view 3D modeling Local feature learning Token fusion Cross-attention

1 Introduction

With the rapid advancement of augmented reality (AR), virtual reality (VR), game development, product design, and digital manufacturing, the demand for high-quality 3D assets has grown exponentially [1–3]. However, traditional 3D assets creation pipelines based on computer-aided design (CAD) remain highly labor-intensive and require substantial expertise, limiting their accessibility to novice users [4, 5]. Existing 2D-to-3D modeling approaches have explored various paradigms, including retrieval-based methods, deep neural network-based reconstruction techniques, and generative models that synthesize 3D meshes or volumetric representations from images and text [6–13].Nevertheless, making high-quality 3D assets creation accessible to non-expert users remains challenging.Freehand sketching provides an intuitive and natural interface for 3D modeling, allowing users to express shape concepts with only a few strokes. Therefore, developing effective methods to transform hand-drawn sketches into complete 3D models has great potential to democratize 3D assets creation and bridge the gap between professional designers and novice users.

Despite recent advances, generating high-fidelity 3D models from freehand sketches remains a challenging problem due to the inherent ambiguity and incompleteness of sketch representations. Unlike photographs, sketches often lack accurate geometric boundaries, consistent perspective relationships, and reliable depth cues, making it difficult to infer the underlying 3D structure from sparse 2D observations. Existing sketch-based 3D reconstruction approaches typically extract geometric cues from one or multiple sketches and employ powerful deep learning architectures [14–22], including convolutional neural networks (CNNs) [23, 24], ResNet-based networks [25], Transformers [26], generative adversarial networks (GANs) [27], and diffusion models [28, 29], to recover 3D geometry.Recent studies have attempted to improve reconstruction performance by enhancing sketch feature representation from different perspectives. For example, viewpoint-aware encoding strategies have been introduced to alleviate the influence of viewpoint variations during feature extraction and reconstruction [14, 30–32]. Other approaches incorporate textual guidance and sketch-text multimodal fusion to enrich semantic understanding of sketches [33, 34]. Meanwhile, improving backbone architectures or strengthening contour information has also been explored to obtain more discriminative sketch representations [35, 36]. Although

these methods significantly improve global shape understanding, accurately recovering fine-grained geometric details from sparse sketch inputs remains difficult.

However, a fundamental limitation remains in the prevailing encoder-decoder paradigm. Most existing methods compress the extracted 2D sketch features into compact global latent representations through flattening or pooling operations to capture global context, such as shape and viewpoint information. However, this compression strategy inevitably weakens the spatial correspondence between image features and local sketch structures, causing the loss of fine-grained geometric and topological cues encoded by individual strokes, including detailed contours, local concavities, and thin protrusions. During decoding, mesh generation is typically performed by regressing vertex deformation from the global latent representation in a single feed-forward process, without an explicit mechanism to adaptively access local sketch evidence when reconstructing detailed geometry. Consequently, the reconstructed models often preserve the overall shape semantics but suffer from over-smoothed local regions and fail to fully capture the geometric intent conveyed by the original sketch.

To address this limitation, we propose Locality-Preserving Sketch Token Fusion (LP-STF), a novel framework for high-fidelity single-view sketch-based 3D reconstruction. The key idea is to preserve local sketch information while maintaining global shape understanding. Instead of collapsing the entire sketch representation into a single latent vector, our framework introduces a dual-path sketch representation that simultaneously learns a global latent representation for shape and viewpoint modeling and a spatially structured local token representation for preserving fine-grained stroke patterns.

Specifically, we design a Token Transformer Block that transforms the extracted sketch features into complementary global and local representations. The global representation captures overall shape semantics and viewpoint information, while the local tokens retain spatially distributed geometric cues from different sketch regions. Furthermore, we introduce a Local Token Cross-Attention Module that establishes an adaptive interaction between 2D sketch tokens and 3D mesh deformation. During decoding, the global shape representation queries local sketch tokens through an attention mechanism, enabling the reconstruction network to dynamically retrieve relevant local evidence for accurate vertex deformation and better alignment between sketch structures and reconstructed geometry. In addition, we incorporate a view-disentanglement branch to improve robustness against viewpoint ambiguity and enhance reconstruction consistency under varying observation conditions.

Our contributions are summarized as follows:

1. **A Token Transformer Block for locality-preserving dual-path sketch representation.** We introduce a novel sketch encoding strategy that avoids excessive global compression by jointly learning global shape semantics and spatially structured local tokens. This design preserves fine-grained stroke information and provides richer geometric cues for downstream 3D reconstruction.
2. **A Local Token Cross-Attention Module for geometry-aware sketch-to-3D alignment.** We propose an adaptive cross-attention-based mesh feature sampling mechanism that enables the decoder to selectively retrieve informative

local sketch regions during mesh generation. By establishing a more precise correspondence between 2D sketch structures and 3D vertex deformation, our approach improves the recovery of local geometric details.

3. **High-fidelity and robust sketch-based 3D reconstruction.** Extensive experiments on ShapeNet-Synthetic and ShapeNet-Sketch [11, 37] demonstrate that our method achieves superior reconstruction performance under both view-conditioned and fully automatic settings. In particular, it provides significant improvements for objects with thin structures and complex local geometries.

2 Related Work

2.1 Sketch-based 3D reconstruction

Sketch-based 3D reconstruction aims to generate three-dimensional models from intuitive hand-drawn sketches and has attracted increasing attention due to its potential to simplify and democratize 3D content creation. Compared with traditional computer-aided design (CAD) systems, sketch-based modeling provides a more natural interaction paradigm, allowing users to express their design concepts with only a few strokes [33]. However, reconstructing accurate 3D geometry from free-hand sketches remains challenging because sketches usually contain incomplete geometric information, ambiguous depth cues, and significant variations caused by different drawing styles.

Early sketch-based modeling approaches mainly relied on handcrafted geometric constraints, contour analysis, and interactive user operations to infer 3D structures from 2D strokes. For example, Sketch2Mesh [15] demonstrated the feasibility of reconstructing and editing 3D shapes from sketches by exploiting geometric relationships between drawn contours and 3D surfaces. Similarly, early studies on free-hand sketch reconstruction explored the use of stroke-based constraints and user guidance to recover plausible 3D structures [38, 39]. However, these approaches typically require substantial manual interaction and are difficult to generalize to diverse object categories.

With the development of deep learning, recent methods have shifted toward automatic sketch-to-3D reconstruction using neural networks. These approaches generally employ encoder-decoder architectures to extract discriminative sketch features and generate 3D representations, including voxels, point clouds, implicit fields, and polygonal meshes [40].

Sketch2Model [14] introduced a view-aware framework for reconstructing 3D models from a single free-hand sketch. By explicitly modeling viewpoint information and disentangling shape-related and view-related representations, the method effectively reduced ambiguity caused by different observation conditions. Deep3DSketch [16–19], further improved reconstruction quality by enhancing stroke-level feature representation and incorporating shape-aware constraints, enabling better recovery of geometric structures from sparse sketch inputs. More recently, SS2M [35] explored stronger feature extraction backbones and additional geometric constraints to improve the fidelity of reconstructed meshes.

Although these methods have achieved significant progress, most existing approaches still encode sketch information into compact global latent representations before mesh generation. Such representations are effective for capturing overall object semantics but may weaken the spatial correspondence between local sketch strokes and reconstructed geometry. Consequently, recovering fine-grained geometric details, such as thin structures, local concavities, and small-scale surface variations, remains challenging.

2.2 View-aware and multimodal sketch reconstruction

A fundamental challenge in sketch-based 3D reconstruction is the inherent ambiguity caused by missing depth information and uncertain viewpoints. Since a free-hand sketch represents only a partial projection of a 3D object, different viewpoints may result in significantly different interpretations of the same object [41, 42]. Therefore, incorporating viewpoint information and additional prior knowledge has become an important direction for improving reconstruction robustness.View-aware approaches attempt to explicitly model viewpoint variations during the reconstruction process. Sketch2Model [14] introduced a viewpoint-aware encoding strategy that separates object shape information from observation-related variations, allowing the network to better distinguish intrinsic geometry from camera effects. Similarly, SS2M [35] incorporated viewpoint prediction and geometric constraints to provide additional guidance for reconstructing 3D shapes from ambiguous sketch inputs.

Beyond viewpoint modeling, recent studies have explored multimodal learning strategies to compensate for the limited information contained in sketches. By introducing complementary semantic information, multimodal approaches can provide additional priors for shape understanding and generation. Magic3DSketch [20] incorporated textual descriptions and language-image pretrained models into sketch-based 3D modeling, leveraging semantic knowledge from large-scale multimodal models to improve the realism and completeness of generated 3D models.

Despite the effectiveness of viewpoint-aware and multimodal strategies, these approaches mainly focus on improving global shape understanding by introducing additional information sources. The problem of preserving the intrinsic spatial structure of sketch strokes and dynamically utilizing local geometric cues during mesh reconstruction remains insufficiently explored.

2.3 Local feature representation and attention mechanisms

The quality of sketch-based 3D reconstruction largely depends on how effectively 2D sketch information is represented and transferred into 3D geometry. Conventional convolutional neural networks typically extract hierarchical features and aggregate them into compact global representations through pooling or flattening operations [22, 43–45]. Although these strategies are effective for capturing high-level semantic information, they may discard spatially localized patterns that are critical for recovering detailed geometry.

Recent advances in transformer-based architectures have demonstrated the effectiveness of token-based representations for preserving structured visual information.

Unlike traditional feature aggregation methods, transformer models represent images as collections of spatially distributed tokens and enable flexible feature interaction through attention mechanisms. Vision Transformer (ViT) [46] introduced patch-level token representations and demonstrated the capability of transformers to model long-range dependencies while maintaining structured visual information. Furthermore, Perceiver [47] explored latent query-based attention mechanisms, enabling efficient interaction between latent representations and input tokens.

Attention mechanisms have also been widely explored in 3D reconstruction tasks. For example, Pixel2Mesh [48] established a direct connection between image features and mesh deformation by progressively refining mesh vertices according to image guidance. Transformer-based architectures such as DETR [49] further demonstrated the effectiveness of cross-attention mechanisms for learning adaptive correspondence between different feature spaces.

However, existing sketch-based 3D reconstruction methods generally compress sketch features into a single global latent representation before decoding. Once local spatial information is aggregated, the decoder lacks an explicit mechanism to revisit informative sketch regions during mesh generation. Motivated by token-based representation learning and attention-driven feature interaction, our work introduces a locality-preserving sketch token representation and an adaptive mesh feature sampling strategy to dynamically exploit local sketch cues for high-fidelity 3D reconstruction.

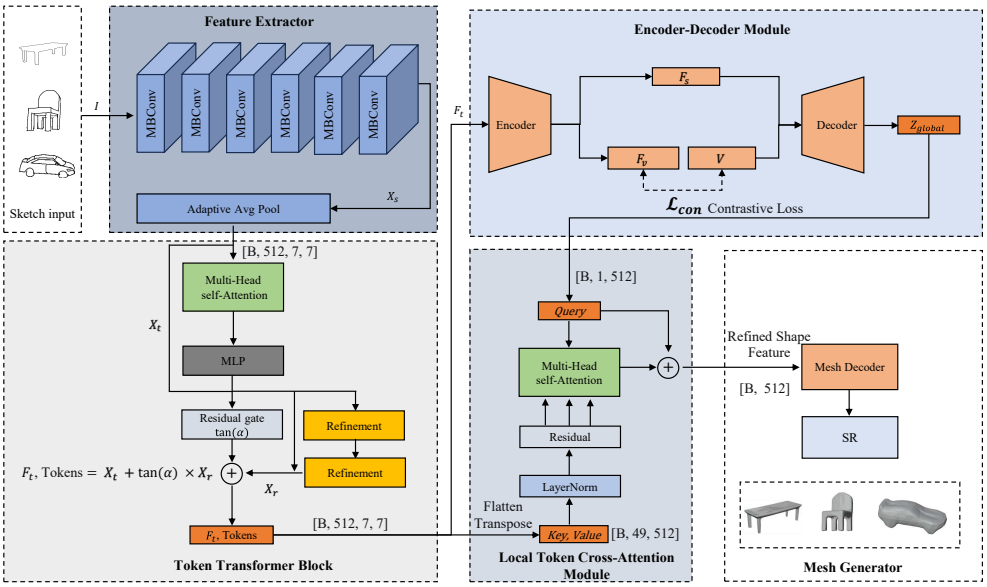
3 Method

3.1 Overview

Given a single free-hand sketch, our goal is to reconstruct a high-quality 3D mesh that preserves both global shape structures and fine-grained local geometries. However, existing sketch-based reconstruction methods often compress the input sketch into a compact global representation, which may weaken or discard spatially localized stroke information. Such information loss makes it difficult to recover fine geometric structures, particularly for objects with thin components and complex local details.

To address this limitation, we propose a Locality-Preserving Sketch Token Fusion framework, which maintains spatially structured sketch tokens and allows the reconstruction network to selectively exploit local geometric cues. As illustrated in Fig. 1, the framework consists of five main components: a Feature Extractor, a Token Transformer Block, an Encoder-Decoder Module, a Local Token Cross-Attention (LTCA) Module, and a Mesh Generator. Among them, the Token Transformer Block preserves and enhances spatially localized sketch information, the Encoder-Decoder Module learns disentangled shape- and viewpoint-related representations, and the LTCA Module dynamically retrieves local sketch features to complement the global shape representation. The resulting geometry-aware representation is then decoded by the Mesh Generator to produce the reconstructed 3D mesh.

Fig. 1 Overview of the proposed locality-preserving sketch-to-3D reconstruction framework. Given an input sketch, the sketch encoder with a Token Transformer produces a dual-path representation consisting of a global shape feature and a spatially structured local token representation $[B, 512, 7, 7]$. The dual-branch latent encoder factorizes the representation into shape- and viewpoint-related codes and produces a global shape latent Z_{global}. A Local Token Cross-Attention (LTCA) module uses Z_{global} as the query to dynamically retrieve informative local sketch tokens, which serve as the key and value features. The retrieved local cues are fused with the global shape representation to form a geometry-aware representation, which is subsequently decoded and refined to generate the final 3D mesh.



3.1.1 Feature Extractor

The input sketch I is first processed by a lightweight convolutional encoder to obtain an intermediate feature map. To preserve spatial structures, instead of performing global feature aggregation, we maintain the spatial feature layout and generate local sketch tokens.

Specifically, the extracted feature map is transformed into a fixed spatial resolution:

$$X_t = \mathcal{P}(X_s), X_t \in \mathbb{R}^{B\times 512\times 7\times 7}. \tag{1}$$

where $\mathcal{P}(\cdot)$ denotes adaptive pooling. Each spatial location corresponds to a local sketch region, providing explicit stroke-level geometric cues for subsequent reconstruction.

3.1.2 Token Transformer Block

Although local tokens preserve spatial information, convolutional features alone cannot sufficiently capture long-range structural relationships. Therefore, we introduce a Token Transformer Block to enhance global reasoning. The feature map is first flattened into a token sequence:

$$T = \{t_i\}_{i=1}^{49}, \quad t_i \in \mathbb{R}^{512}. \tag{2}$$

Multi-head self-attention is applied to model interactions among different sketch regions:

$$T_a = \mathrm{MSA}(T), \tag{3}$$

followed by an MLP refinement:

$$T_r = \mathrm{MLP}(T_a). \tag{4}$$

To adaptively balance original sketch information and transformed features, we introduce a learnable gated residual fusion:

$$F_{token} = X_t + \tan(\alpha) \cdot X_r, F_{token} \in \mathbb{R}^{B\times 512\times 7\times 7}. \tag{5}$$

where X_r represents the refined token feature and α is a learnable parameter. The proposed gated fusion preserves original stroke structures while enhancing discriminative geometric patterns, producing the final locality-preserving token representation F_{token}.

3.2 Encoder-Decoder Module

The locality-preserving tokens are further transformed into a compact latent representation. Rather than collapsing the sketch into a single embedding, we factorize the representation into two complementary codes, F_s and F_v, corresponding to intrinsic shape and viewpoint information, respectively. In particular, F_v is explicitly guided to encode the viewing configuration of the input sketch, while F_s is expected to preserve geometric characteristics that are less sensitive to viewpoint variations. The resulting factorized representation provides a suitable basis for subsequent 3D reconstruction and disentanglement learning.

To explicitly inject viewpoint supervision into the latent space, we impose a contrastive alignment objective between the predicted view representation F_v and the ground-truth view embedding V. Specifically, the ground-truth view is encoded into the same embedding space as F_v, providing a positive target, while the view embeddings of other samples serve as negative candidates. We formulate the contrastive objective as

$$\mathcal{L}_{\mathrm{con}} = -\log \frac{\exp\left(\mathrm{sim}(F_v, V)/\tau\right)}{\sum_j \exp\left(\mathrm{sim}(F_v, V_j)/\tau\right)}, \tag{6}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity, V is the embedding of the ground-truth viewpoint associated with the input sketch, V_j denotes the viewpoint embedding of the j-th candidate, and τ is a temperature parameter. This objective explicitly

encourages F_v to capture viewpoint-related information. Together with the subsequent disentanglement constraints, it further promotes the separation of view-dependent cues from intrinsic shape information. The decoder then aggregates the two complementary representations Z_{global}, which provides a global geometric prior for mesh reconstruction.

3.3 Local Token Cross-Attention Module

Although the global shape latent Z_{global} captures high-level semantic and geometric information, its spatially collapsed representation does not preserve explicit correspondence with local sketch structures. This limitation is particularly pronounced for fine-grained geometric elements, such as object contours, narrow parts, and small structural details. To recover such information during mesh reconstruction, we introduce a Local Token Cross-Attention Module that performs global-to-local cross-attention between the shape latent and the locality-preserving sketch tokens.

Specifically, we first project the global shape latent into a query token:

$$Q = W_q Z_{\text{global}}, \qquad Q \in \mathbb{R}^{B\times 1\times 512}. \tag{7}$$

Meanwhile, the locality-preserving sketch tokens are flattened along their spatial dimensions and projected into key and value representations:

$$K = W_k \,\text{Flatten}(F_{\text{token}}), \qquad V = W_v \,\text{Flatten}(F_{\text{token}}), \tag{8}$$

where

$$K, V \in \mathbb{R}^{B\times 49\times 512}. \tag{9}$$

The global shape query then attends to the local sketch tokens through scaled dot-product attention:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{10}$$

where d denotes the feature dimension of the attention space. The corresponding local geometric cues are retrieved by

$$Z_{\text{local}} = AV. \tag{11}$$

Unlike uniform aggregation of local features, the proposed query-driven sampling allows the global shape representation to adaptively identify sketch regions that are relevant to the current geometric reconstruction. This mechanism facilitates the retrieval of fine-grained local cues, including object boundaries, thin structures, and other structurally informative regions.

Finally, the retrieved local information is integrated with the global shape representation through residual fusion:

$$Z_f = Z_{\text{global}} + Z_{\text{local}}, \tag{12}$$

yielding a geometry-aware representation that combines global semantic structure with fine-grained local sketch evidence.

3.4 Mesh Generator

The fused representation Z_f is fed into the mesh generator:

$$M_0 = \mathcal{G}(Z_f), \tag{13}$$

where M_0 denotes the initial mesh prediction.

To further improve geometric continuity and surface quality, a mesh refinement module is applied:

$$M = \mathcal{R}(M_0). \tag{14}$$

The refinement stage optimizes local surface details and produces the final 3D mesh corresponding to the input sketch.

3.5 Training Objectives

The proposed framework is trained with complementary objectives that constrain the reconstructed geometry, regularize the generated mesh, and provide explicit supervision for the viewpoint representation. The complete training objective is formulated as

$$\mathcal{L}_{\text{total}} = \lambda_s \mathcal{L}_s + \lambda_r \mathcal{L}_r + \lambda_{\text{con}} \mathcal{L}_{\text{con}}, \tag{15}$$

where λ_s, λ_r, and λ_{con} are weighting coefficients that control the contributions of the individual objectives. The three components provide complementary supervision: $\mathcal{L}_s$ constrains the reconstructed geometry in the image space, $\mathcal{L}_r$ promotes geometrically plausible mesh surfaces, and $\mathcal{L}_{\text{con}}$ explicitly calibrates the viewpoint-related latent representation. Their joint optimization enables the proposed framework to reconstruct 3D geometry that is consistent with the observed sketch while maintaining an explicit representation of viewpoint information.

3.5.1 Silhouette Reconstruction.

Since explicit 3D annotations are not required, we constrain the predicted mesh through its rendered silhouette. Given the reconstructed mesh and its corresponding viewpoint, a differentiable renderer is used to obtain the silhouette S. We then measure its consistency with the target silhouette $\hat{S}$ extracted from the input sketch using a differentiable Intersection-over-Union (IoU) objective:

$$\mathcal{L}_s = 1 - \frac{|S \cap \hat{S}|}{|S \cup \hat{S}|}. \tag{16}$$

Here, S denotes the silhouette rendered from the predicted mesh under the corresponding viewpoint, while $\hat{S}$ represents the target silhouette. Minimizing $\mathcal{L}_s$ encourages the reconstructed geometry to preserve the visible shape evidence contained in the input sketch.

3.5.2 Mesh Regularization.

Silhouette supervision alone does not fully constrain the underlying surface geometry and may produce locally irregular or distorted mesh structures. To improve the geometric quality of the reconstructed mesh, we introduce surface regularization:

$$\mathcal{L}_r = \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{edge}}, \tag{17}$$

where $\mathcal{L}_{\text{lap}}$ promotes local surface smoothness by penalizing undesirable variations in neighboring vertex configurations, and $\mathcal{L}_{\text{edge}}$ encourages a regular distribution of edge lengths. These complementary constraints suppress geometric artifacts and promote coherent and plausible mesh surfaces during reconstruction.

3.5.3 View-Aware Contrastive Alignment.

The view branch is explicitly supervised to recover the viewpoint associated with the input sketch. Rather than relying solely on the reconstruction objective, we align the predicted view representation F_v with its corresponding ground-truth view embedding V. Specifically, the ground-truth view is mapped into the same representation space as F_v and serves as the positive target, whereas viewpoint embeddings associated with other samples provide negative candidates. We adopt cosine similarity to measure the correspondence between the predicted and target representations:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(\text{sim}(F_v, V)/\tau\right)}{\sum_j \exp\left(\text{sim}(F_v, V_j)/\tau\right)}, \tag{18}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, V is the ground-truth viewpoint embedding associated with the input sketch, V_j denotes the viewpoint embedding of the j-th candidate, and τ is a temperature parameter. This objective explicitly guides F_v toward discriminative viewpoint representations by increasing its similarity to the correct viewpoint while distinguishing it from competing view configurations. Together with the subsequent disentanglement constraints, it contributes to separating viewpoint-dependent cues from intrinsic shape information.

4 Experiments

We first conduct a case study to demonstrate the effectiveness of our Locality-Preserving Sketch Token Fusion framework and analyze how the proposed local-token cross-attention mesh sampler facilitates fine-grained sketch-to-geometry alignment. By enabling the global shape representation to adaptively query informative local sketch tokens during mesh decoding, our approach establishes a more precise correspondence between 2D stroke structures and 3D vertex deformations. Extensive quantitative and qualitative evaluations on both synthetic and hand-drawn sketches demonstrate that, compared with existing baselines, our method produces reconstructed shapes with higher geometric fidelity and stronger consistency with the input sketches.

4.1 Datasets

We evaluate on the two standard benchmarks introduced with Sketch2Model [14], covering 13 ShapeNet [37] categories: airplane, bench, cabinet, car, chair, display, lamp, loudspeaker, rifle, sofa, table, telephone and watercraft.

ShapeNet-Synthetic. We train our model on the ShapeNet-Synthetic [11] dataset, a synthetic benchmark constructed from ShapeNet containing 13 object categories, including airplanes, benches, cars, and other representative classes. Each 3D model is rendered from 20 different viewpoints with an image resolution of 224×224 pixels. Following Sketch2Model, we adopt the same training/testing split to enable a fair and consistent evaluation against existing sketch-based 3D reconstruction approaches.

ShapeNet-Sketch. To evaluate the generalization capability of our model on real hand-drawn sketches, we conduct cross-domain testing on the ShapeNet-Sketch dataset. This dataset is constructed by randomly selecting 100 rendered images from each of the 13 ShapeNet categories, with each rendered image paired with a corresponding hand-drawn sketch created by human participants with different levels of drawing proficiency. In total, the dataset contains 1,300 sketch-model pairs covering diverse object categories. Since ShapeNet-Sketch is designed exclusively for evaluation, none of its samples are used during the training process.

4.2 Implementation details

The backbone is EfficientNet-B0 [50] truncated at the seventh stage and projected to 512 channels. The Token Transformer and Local Token Cross-Attention Module each use 8 attention heads with a hidden width of 512. The mesh decoder deforms a fixed-topology template. silhouette supervision is obtained with a differentiable renderer [21, 22]. Input sketches are resized to 224×224.Our network is implemented in PyTorch and optimized using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. All experiments are performed on an NVIDIA GeForce RTX A6000 GPU. The loss weights are set as $\lambda_s = \lambda_r = \lambda_{\text{mse}} = 1$ and $\lambda_{\text{con}} = 0.05$. The model is trained for 2,000 epochs, with validation conducted every 500 epochs to select the best-performing checkpoint.

4.3 Comparisons and Evaluations

Quantitative Comparison. We compare our method with several state-of-the-art approaches, including retrieval-based methods, conventional encoder-decoder architectures equipped with the Soft Rasterizer (SoftRas) [22] differentiable renderer, and recent deep learning-based sketch reconstruction methods, including Sketch2Model [14], Deep3DSketch+ [17], Magic3DSketch [20], and SS2M [35].

The retrieval-based method extracts sketch features using a pretrained sketch classification network, which has demonstrated stronger representation capability than handcrafted descriptors in sketch-related tasks. Deep3DSketch+ improves reconstruction quality by enhancing sketch contour information, while SS2M introduces a more powerful feature extraction backbone and additional geometric constraints to improve reconstruction performance. These methods provide strong baselines for evaluating

Table 1 Quantitative comparison with state-of-the-art methods on the ShapeNet-Synthetic dataset. IoU scores are reported for each object category.

Dataset	Method	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp
Sketch-Synthetic	Retrieval	0.513	0.380	0.518	0.667	0.346	0.385	0.325
	SoftRas	0.576	0.467	0.663	0.769	0.496	0.541	0.431
	Sketch2Model (Pred)	0.618	0.477	0.667	0.746	0.515	0.550	0.463
	Sketch2Model (GT)	0.624	0.481	0.701	0.751	0.522	0.604	0.472
	Deep3DSketch+	0.632	0.510	0.699	0.782	0.525	0.588	0.466
	Magic3DSketch	0.637	0.477	0.701	0.772	0.530	0.593	0.459
	SS2M	0.637	**0.519**	**0.715**	**0.789**	0.541	**0.617**	0.469
	Ours (GT)	**0.660**	0.507	0.708	0.770	**0.560**	**0.617**	**0.505**
	Ours (Pred)	0.653	0.495	0.689	0.767	0.555	0.584	0.491

Dataset	Method	Loudspeaker	Rifle	Sofa	Table	Telephone	Watercraft	Sum	Avg
Sketch-Synthetic	Retrieval	0.468	0.475	0.483	0.311	0.622	0.422	5.915	0.455
	SoftRas	0.629	0.605	0.613	0.512	0.706	0.556	7.564	0.582
	Sketch2Model (Pred)	0.624	0.606	0.620	0.470	0.673	0.569	7.598	0.584
	Sketch2Model (GT)	0.641	0.612	0.622	0.478	0.719	0.586	7.813	0.601
	Deep3DSketch+	0.630	0.632	0.640	0.510	0.757	0.583	7.954	0.612
	Magic3DSketch	0.620	0.632	0.639	0.503	0.697	0.572	7.832	0.602
	SS2M	**0.694**	0.584	**0.642**	0.452	**0.754**	0.598	8.011	0.616
	Ours (GT)	0.672	**0.662**	0.639	**0.515**	0.739	**0.614**	**8.167**	**0.628**
	Ours (Pred)	0.630	0.649	0.638	0.486	0.707	0.592	7.935	0.610

the effectiveness of our proposed Locality-Preserving Sketch Token Fusion framework.Following prevIoU work [21, 22], we adopt the voxel Intersection-over-Union (IoU) as the evaluation metric to measure reconstruction accuracy. Higher IoU values indicate better geometric consistency between the reconstructed 3D shape and the ground-truth model.

Evaluation on ShapeNet-Synthetic Dataset. We first evaluate our method on the ShapeNet-Synthetic dataset, where all sketches are generated from rendered views with controlled viewpoints. The quantitative results are reported in Table 1. Compared with existing approaches, our method achieves the highest average IoU under both ground-truth viewpoint and predicted viewpoint settings, reaching 0.628 and 0.610, respectively. In particular, our method obtains superior performance on several categories with complex geometric structures, including Airplane, Chair, Lamp, Rifle, and Watercraft. These improvements indicate that the proposed locality-preserving representation effectively captures fine-grained stroke patterns and improves the correspondence between sketch structures and reconstructed 3D geometry. Compared with the previous state-of-the-art method SS2M [35], our approach improves the average IoU from 0.616 to 0.628 under the ground-truth viewpoint setting. The consistent improvements across different object categories demonstrate the effectiveness of preserving local sketch information rather than relying solely on globally compressed shape representations.

Evaluation on ShapeNet-Sketch Dataset. To further evaluate the generalization capability of our model, we conduct experiments on the ShapeNet-Sketch dataset, which contains human-created sketches with larger variations in drawing styles and stronger structural ambiguity. The quantitative results are summarized in Table 2. Our method achieves the highest average IoU among all compared approaches, obtaining 0.498 and 0.487 under ground-truth viewpoint and predicted viewpoint settings,

Table 2 Quantitative comparison with state-of-the-art methods on the ShapeNet-Sketch dataset. IoU scores are reported for each object category.

Dataset	Method	Airplane	Bench	Cabinet	Car	Chair	Display	Lamp
Sketches	Retrieval	0.411	0.219	0.409	0.626	0.238	0.338	0.223
	SoftRas	0.469	0.347	0.545	0.648	0.361	0.472	0.328
	Sketch2Model (Pred)	0.479	0.357	0.547	0.649	0.383	0.435	0.336
	Sketch2Model (GT)	0.487	0.366	0.568	0.659	0.393	0.479	0.338
	Deep3DSketch+	0.490	0.368	0.553	0.675	0.382	0.463	0.347
	Magic3DSketch	0.497	0.382	0.555	**0.688**	**0.413**	0.497	0.328
	Ours (GT)	**0.508**	0.382	**0.572**	0.674	0.406	**0.498**	**0.358**
	Ours (Pred)	0.506	**0.383**	0.549	0.660	0.392	0.462	0.353

Dataset	Method	Loudspeaker	Rifle	Sofa	Table	Telephone	Watercraft	Sum	Avg
Sketches	Retrieval	0.365	0.413	0.431	0.232	0.536	0.369	4.810	0.370
	SoftRas	0.533	0.541	0.534	0.359	0.537	0.456	6.130	0.472
	Sketch2Model (Pred)	0.526	0.510	0.528	0.361	0.551	0.450	6.112	0.470
	Sketch2Model (GT)	0.544	0.534	0.534	0.357	0.554	0.466	6.279	0.483
	Deep3DSketch+	0.514	0.543	0.534	0.370	0.576	0.467	6.282	0.483
	Magic3DSketch	0.523	0.550	0.531	**0.372**	0.517	0.459	6.312	0.486
	Ours (GT)	**0.556**	**0.565**	**0.559**	0.366	0.562	**0.474**	**6.480**	**0.498**
	Ours (Pred)	0.534	0.557	0.535	0.360	**0.568**	0.471	6.329	0.487

respectively. Although Magic3DSketch [20] achieves slightly better performance on several individual categories, including Car, Chair, and Table, our method provides better overall reconstruction accuracy across the entire dataset. This indicates that the proposed locality-preserving sketch representation is more robust when transferring from synthetic training sketches to real hand-drawn sketches with diverse drawing styles. The stable performance under predicted-view settings further demonstrates the capability of our framework to handle viewpoint uncertainty without requiring explicit viewpoint annotations.

Analysis of Improvement. The performance improvements mainly benefit from the complementary designs of the Token Transformer Block (TTB) and the Local Token Cross-Attention Module (LTCA). Unlike conventional approaches that compress sketch features into a single global latent representation, TTB preserves spatially structured local tokens while maintaining global shape semantics. This design enables the network to retain fine-grained stroke information, which is critical for recovering detailed geometric structures. Furthermore, LTCA introduces adaptive interactions between mesh queries and local sketch tokens during mesh decoding. By selectively retrieving informative local regions, LTCA establishes more accurate correspondence between 2D stroke structures and 3D vertex deformations, leading to improved reconstruction of local details while preserving global shape consistency. The consistent improvements observed on both synthetic and hand-drawn sketch datasets demonstrate that the proposed framework effectively alleviates the information loss caused by global feature compression and provides a more reliable solution for high-fidelity sketch-based 3D reconstruction.

Qualitative Results. Fig. 2 presents representative qualitative reconstruction results across multiple object categories, including airplanes, tables, chairs, cabinets, rifles, and sofas. The first row shows the input free-hand sketches, the second row

presents the corresponding ground-truth 3D models, and the third row illustrates the 3D shapes reconstructed by our method.Overall, our method produces reconstructed shapes that exhibit strong consistency with the input sketches in terms of both global structure and local geometric details. Despite the sparse and inherently ambiguous nature of free-hand sketches, the reconstructed models successfully recover the major structural components and overall shape configurations indicated by the input strokes. In particular, the results demonstrate that our method can preserve thin structures, elongated components, and distinctive local geometric patterns, which are often difficult to infer from limited 2D sketch information.

The reconstruction results can be attributed to the complementary roles of the proposed TTB and LTCA modules. TTB preserves spatially localized stroke information through the locality-preserving dual-path sketch representation, allowing fine-grained sketch cues to be retained alongside global shape semantics. Meanwhile, LTCA enables the mesh decoder to adaptively retrieve informative local sketch features during mesh generation, facilitating more precise correspondence between 2D stroke structures and 3D vertex deformations. These two components jointly provide richer geometric cues for recovering detailed 3D structures while maintaining global shape consistency.

Although minor discrepancies remain in some fine-scale regions, which can be attributed to the inherent ambiguity and incompleteness of free-hand sketches, the reconstructed shapes generally follow the structural and geometric cues conveyed by the input sketches. These qualitative results demonstrate the capability of our framework to transform sparse hand-drawn sketches into geometrically plausible and detailed 3D shapes.

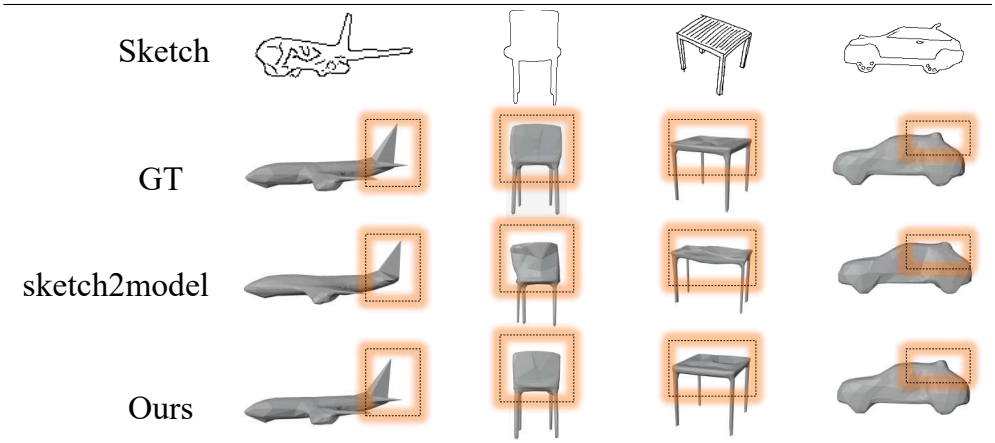
Fig. 2 Some representative results on our test dataset. The ground truth is provided at the bottom of each example for comparison. Our method is capable of generating promising shapes from hand-drawn sketches. Note that the results shown here are generated using predicted viewpoints.



We visualize selected failure-case samples from the ShapeNet-Sketch test set in Fig. 3. Although the overall global geometry of objects is largely reconstructed by all methods, fine-grained local details are prone to be lost in the baseline sketch2model. For the airplane sample, sketch2model suffers from distorted tail geometry, while our method faithfully preserves the fine tail details consistent with the ground-truth (GT). For the chair, the baseline produces a deformed backrest, whereas our approach recovers the original chair back structure accurately. For the bench, sketch2model yields an elongated rectangular bench, deviating from the square-shaped ground-truth; our model generates the correct square bench topology. For the car, although the baseline captures the coarse car silhouette, it fails to reconstruct the rear-wing detail. In contrast, our method reconstructs the car's rear-wing with clear fine-level geometry matching GT.

These examples highlight the necessity of hierarchical information for fine-detail reconstruction. Even for these challenging sketch inputs, our model maintains reasonable generalization capability and reconstructs objects from unseen training sketches, demonstrating that the learned shape priors can transfer to novel data. Nevertheless, these observed limitations reveal that improving fine-detail preservation and generalization is still a vital future research direction, which may be achieved via enhanced hierarchical feature extraction and more powerful representation learning.

Fig. 3 Qualitative comparison on challenging samples from ShapeNet-Sketch test set. Compared with sketch2model, our method better preserves fine-grained local details including airplane tail, chair backrest, square bench topology, and car rear-wing, while maintaining plausible global object shapes.



5 Ablation Experiments

We conduct ablation studies to investigate the individual and complementary contributions of the proposed Token Transformer Block (TTB) and Local Token

Table 3 Ablation study of each component on the ShapeNet-Synthetic test set. Values in parentheses indicate the results under predicted viewpoints. The best results are highlighted in bold.

Method	Airplane	Chair	Lamp
baseline	0.624 (0.618)	0.522 (0.515)	0.472 (0.463)
+ TTB	0.636 (0.630)	0.540 (0.521)	0.481 (0.473)
+ LTCA	0.650 (0.645)	0.551 (0.545)	0.490 (0.481)
Ours	**0.660 (0.653)**	**0.560 (0.555)**	**0.505 (0.491)**

Method	Rifle	Table	Watercraft
baseline	0.612 (0.606)	0.478 (0.470)	0.586 (0.569)
+ TTB	0.623 (0.612)	0.491 (0.473)	0.595 (0.578)
+ LTCA	0.645 (0.640)	0.512 (0.481)	0.612 (0.590)
Ours	**0.662 (0.649)**	**0.515 (0.486)**	**0.614 (0.592)**

Table 4 Ablation study of TTB and LTCA modules on the ShapeNet-Sketch test set in terms of IoU. The best results are highlighted in bold.

TTB	LTCA	Airplane	Chair	Lamp
		0.487 (0.479)	0.393 (0.383)	0.338 (0.336)
✓		0.492 (0.485)	0.398 (0.385)	0.346 (0.341)
	✓	0.505 (0.497)	0.402 (0.390)	0.354 (0.353)
✓	✓	**0.508 (0.506)**	**0.406 (0.392)**	**0.358 (0.353)**

TTB	LTCA	Rifle	Table	Watercraft
		0.534 (0.510)	0.357 (0.361)	0.466 (0.450)
✓		0.547 (0.521)	0.359 (0.357)	0.472 (0.454)
	✓	0.550 (0.545)	0.364 (0.358)	0.474 (0.468)
✓	✓	**0.565 (0.557)**	**0.366 (0.360)**	**0.474 (0.471)**

Cross-Attention (LTCA). Quantitative results on ShapeNet-Synthetic and ShapeNet-Sketch are reported in Table 3 and Table 4, respectively. For each reported IoU value, the number outside parentheses represents the ground-truth viewpoint (GT) setting, while the number inside parentheses represents the predicted viewpoint (Pred) setting.

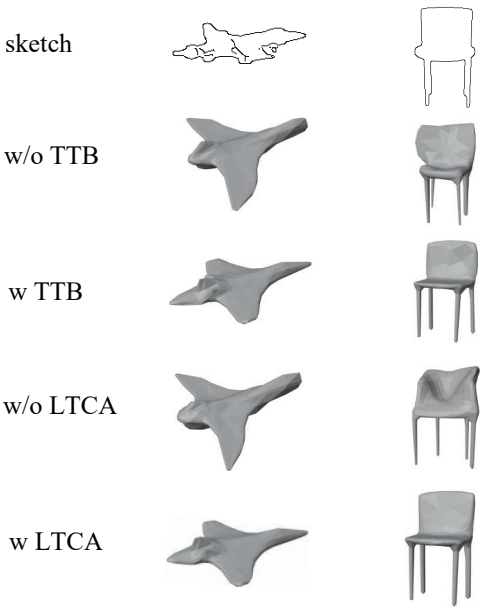
Effect of TTB. Without TTB, the reconstruction performance consistently decreases across all evaluated categories. On the ShapeNet-Synthetic dataset, the IoU drops from 0.660 to 0.624 for airplanes and from 0.560 to 0.522 for chairs. This performance degradation demonstrates the importance of the proposed token-based sketch representation for capturing discriminative and spatially structured features from free-hand sketches. By preserving localized stroke patterns while maintaining global shape semantics, TTB enables the network to establish more reliable correspondences between 2D sketch structures and 3D geometric representations.

Effect of LTCA. Introducing LTCA further improves the performance compared with the model equipped only with TTB. For example, on ShapeNet-Synthetic, the IoU increases from 0.636 to 0.650 for airplanes and from 0.540 to 0.551 for chairs. These improvements verify the effectiveness of explicitly leveraging local sketch tokens during

mesh reconstruction. Instead of relying solely on a globally compressed shape representation, LTCA enables mesh queries to adaptively attend to informative local sketch regions, providing additional geometric cues for recovering fine-grained structures and improving sketch-to-geometry alignment.

Complementary contributions. Combining TTB and LTCA achieves the best reconstruction performance on both ShapeNet-Synthetic and ShapeNet-Sketch datasets. Compared with the baseline model, the complete framework improves the IoU from 0.624 to 0.660 for airplanes and from 0.522 to 0.560 for chairs on ShapeNet-Synthetic. Similar improvements are consistently observed on ShapeNet-Sketch, where the full model surpasses variants containing only one of the proposed modules. These results demonstrate that TTB and LTCA provide complementary benefits: TTB enhances the quality of sketch token representations by preserving local spatial information, while LTCA enables the decoder to dynamically retrieve relevant local cues during mesh generation. Together, they contribute to more accurate reconstruction of detailed geometry from ambiguous sketch inputs.

Fig. 4 Qualitative ablation results on ShapeNet-Sketch. Removing TTB or LTCA causes the loss or distortion of fine-grained structures, such as the tail of the airplane and the backrest of the bench, and produces less accurate local geometry. In contrast, the full model better preserves distinctive local structures and yields sharper, more faithful reconstructions.



Qualitative analysis. To further examine the effect of the two modules, Fig. 4 presents qualitative comparisons on challenging sketches from ShapeNet-Sketch. Removing TTB or LTCA leads to noticeable degradation in structurally distinctive

regions that require fine-grained local cues. For example, the airplane reconstruction without the proposed modules tends to lose its tail structure, while the bench reconstruction exhibits a distorted backrest and missing local components. These failures are less pronounced in the full model, which produces sharper contours and better preserves small but geometrically important structures. The qualitative results are consistent with the quantitative ablation results, demonstrating that explicitly maintaining and exploiting local sketch tokens is critical for faithful detail reconstruction.

6 Conclusion

In this paper, we presented a Locality-Preserving Sketch Token Fusion framework for high-quality 3D mesh reconstruction from a single free-hand sketch. The proposed framework is designed to preserve both global shape structures and fine-grained local geometries that can be easily lost when sketch features are compressed into a compact global representation. Specifically, the Token Transformer Block (TTB) preserves and enhances spatially structured sketch tokens, while the Encoder-Decoder Module learns disentangled shape- and viewpoint-related representations. Building upon these representations, the Local Token Cross-Attention (LTCA) Module dynamically retrieves relevant local sketch features and complements the global shape representation during geometry reconstruction. The resulting geometry-aware representation is then decoded by the Mesh Generator to produce the final 3D mesh.

Extensive experiments on ShapeNet-Synthetic and ShapeNet-Sketch demonstrate the effectiveness of the proposed framework. The quantitative results show consistent improvements when TTB and LTCA are introduced, while qualitative comparisons further demonstrate their ability to preserve structurally distinctive local components and fine-grained geometric details. In particular, the proposed design helps alleviate the loss or distortion of localized structures, such as thin components and small geometric parts, that are difficult to recover from globally aggregated sketch representations. These results confirm that explicitly preserving and exploiting spatially localized sketch information is beneficial for faithful sketch-based 3D reconstruction.

In future work, we will explore more effective hierarchical token representations and adaptive local feature selection for highly ambiguous sketches and geometrically complex objects. We also plan to investigate additional complementary cues, such as depth or semantic information, to further constrain the reconstruction of geometric details that are difficult to infer from a single free-hand sketch.

Acknowledgements This work was partially supported by the Research Project of Guangdong Higher Education Institutions (2023KQNCX108), the Guangdong Province Key Construction Discipline Research Ability Enhancement Project (Grant No. 2024ZDJS071).

Author Contributions Q.L. designed the research and led the manuscript writing. M.J. assisted with the experiments and contributed to the analysis of results. N.L. provided research guidance and contributed to the overall study design. X.L. and M.Z. supervised the project and provided continuous support. All authors read and approved the final manuscript.

Data Availability The Shapenet Datasets are available at: https://shapenet.org. Further data will be made available from the corres ponding author on reasonable request.

Declarations Conflict of interest The authors declare no Conflict of interest.

References

Wang, M., Lyu, X.-Q., Li, Y.-J., Zhang, F.-L.: Vr content creation and exploration with deep learning: A survey. Computational Visual Media **6**(1), 3–28 (2020)

He, Y., Xie, H., Miyata, K.: Sketch2cloth: Sketch-based 3d garment generation with unsigned distance fields. In: 2023 Nicograph International (NicoInt), pp. 38–45 (2023). IEEE

Qi, A., Gryaditskaya, Y., Song, J., Yang, Y., Qi, Y., Hospedales, T.M., Xiang, T., Song, Y.-Z.: Toward fine-grained sketch-based 3d shape retrieval. IEEE transactions on image processing **30**, 8595–8606 (2021)

Cohen, J.M., Markosian, L., Zeleznik, R.C., Hughes, J.F., Barzel, R.: An interface for sketching 3d curves. In: Proceedings of the 1999 Symposium on Interactive 3D Graphics, pp. 17–21 (1999)

Deng, C., Huang, J., Yang, Y.-L.: Interactive modeling of lofted shapes from a single image. Computational Visual Media **6**(3), 279–289 (2020)

Colom, J.: Reconstruction from multi-view sketches: an inverse rendering approach (2023)

Delanoy, J., Aubry, M., Isola, P., Efros, A.A., Bousseau, A.: 3d sketching using multi-view deep volumetric prediction. Proceedings of the ACM on computer graphics and interactive techniques **1**(1), 1–22 (2018)

Li, K., Jin, W., Li, N., Huang, K.: Depth: Disentangled embeddings and priors via two-stage heterogeneous-fusion. Expert Systems with Applications, 131949 (2026)

Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., Wang, R.: 3d shape reconstruction from sketches via multi-view convolutional networks. In: 2017 International Conference on 3D Vision (3DV), pp. 67–77 (2017). IEEE

Zhou, J., Luo, Z., Yu, Q., Han, X., Fu, H.: Ga-sketching: Shape modeling from multi-view sketching with geometry-aligned deep implicit functions. In: Computer Graphics Forum, vol. 42, p. 14948 (2023). Wiley Online Library

Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. Advances in neural information processing systems **30** (2017)

Wang, F., Zhang, Z., Jiang, J., Zhao, B., Hao, Z., Luo, X.: Sketch-based feature fusion and complement for robust sketch-to-voxel reconstruction. Information Fusion, 103522 (2025)

Fu, L., Li, N., Hu, H., Chen, Z., Cao, Y.: Spatialfuse3d: Relational-aware multimodal framework for comprehensive 3d scene understanding. Neurocomputing, 132267 (2025)

Zhang, S.-H., Guo, Y.-C., Gu, Q.-W.: Sketch2model: View-aware 3d modeling from single free-hand sketches. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6000–6017 (2021). IEEE

Guillard, B., Remelli, E., Yvernay, P., Fua, P.: Sketch2mesh: Reconstructing and editing 3d shapes from sketches. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13003–13012 (2021). IEEE

Chen, T., Fu, C., Zhu, L., Mao, P., Zhang, J., Zang, Y., Sun, L.: Deep3dsketch: 3d modeling from free-hand sketches with view-and structural-aware adversarial training. arXiv preprint arXiv:2312.04435 (2023)

Chen, T., Fu, C., Zang, Y., Zhu, L., Zhang, J., Mao, P., Sun, L.: Deep3dsketch+: Rapid 3d modeling from single free-hand sketches. In: International Conference on Multimedia Modeling, pp. 16–28 (2023). Springer

Zang, Y., Ding, C., Chen, T., Mao, P., Hu, W.: Deep3dsketch+\+: High-fidelity 3d modeling from single free-hand sketches. arXiv preprint arXiv:2310.18178 (2023)

Chen, T., Cao, R., Li, Z., Zang, Y., Sun, L.: Deep3dsketch-im: rapid high-fidelity ai 3d model generation by single freehand sketches. Frontiers of Information Technology & Electronic Engineering **25**(1), 149–159 (2024)

Zang, Y., Han, Y., Ding, C., Zhang, J., Chen, T.: Magic3dsketch: Create colorful 3d models from sketch-based 3d modeling guided by text and language-image pre-training. Neurocomputing, 131925 (2025)

Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3907–3916 (2018). Ieee

Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7708–7717 (2019)

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015). pmlr

Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

Liu, F.-L., Fu, H., Lai, Y.-K., Gao, L.: Sketchdream: Sketch-based text-to-3d generation and editing. ACM Transactions on Graphics (TOG) **43**(4), 1–13 (2024)

Zheng, W., Xia, H., Chen, R., Sun, L., Shao, M., Xia, S., Ding, Z.: Sketch3d: Style-consistent guidance for sketch-to-3d generation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 3617–3626 (2024)

Chen, M., Yuan, W., Wang, Y., Sheng, Z., He, Y., Dong, Z., Bo, L., Guo, Y.: Sketch2nerf: multi-view sketch-guided text-to-3d generation. arXiv preprint arXiv:2401.14257 (2024)

Lei, Q., Li, N., Zhong, J., Zhao, M., Fan, M.: Hierarchical sketch encoding and text-guided 3d modeling: a novel approach for single free-hand sketch reconstruction. The Visual Computer **42**(7), 279 (2026)

Liu, G., Huo, C., Xu, Z., Zhao, J., Dong, Y., Wang, B., Sun, X., Jiang, Y., Zhao, Y.: Freehand sketch-based 3d reconstruction with contour constraints via elastic metrics. In: Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology, pp. 1–11 (2025)

Wu, Y., Wang, F., Zhao, P., Zhou, M., Geng, S., Zhang, D.: High-fidelity 3d mesh generation from a single sketch using shape constraints. Scientific Reports **16**(1), 1127 (2025)

Qiu, R., Miao, Y., Wang, S., Zhu, Y., Yu, L., Gao, X.-S.: Powermlp: An efficient version of kan. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 20069–20076 (2025)

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)

Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European Conference on Computer Vision, pp. 628–644 (2016). Springer

Wang, J., Lin, J., Yu, Q., Liu, R., Chen, Y., Yu, S.X.: 3d shape reconstruction from free-hand sketches. In: European Conference on Computer Vision, pp. 184–202 (2022). Springer

Li, X., Zhang, Q., Kang, D., Cheng, W., Gao, Y., Zhang, J., Liang, Z., Liao, J., Cao, Y.-P., Shan, Y.: Advances in 3d generation: A survey. arXiv preprint arXiv:2401.17807 (2024)

Horn, B.K.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view (1970)

Witkin, A.P.: Recovering surface shape and orientation from texture. Artificial intelligence **17**(1-3), 17–45 (1981)

Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3501–3512 (2020). IEEE

Tulsiani, S., Zhou, T., Efros, A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. corr abs/1704.06254 (2017). arXiv preprint arXiv:1704.06254

Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems **29** (2016)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint

arXiv:2010.11929 (2020)
Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International Conference on Machine Learning, pp. 4651–4664 (2021). PMLR
Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.-G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: European Conference on Computer Vision, pp. 55–71 (2018). Springer
Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer
Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PmLR