TopoMesh: High-Fidelity Mesh Autoencoding via Topological Unification

Guan Luo[1,2] Xiu Li[2] Rui Chen[2,3] Xuanyu Yi[2] Jing Lin[2]
Chia-Hao Chen[1] Jiahang Liu[2] Song-Hai Zhang[1†] Jianfeng Zhang[2†]
[1] Tsinghua University [2] ByteDance Seed [3] HKUST

https://logan0601.github.io/projects/topomesh/index.html

Figure 1. High-fidelity mesh at 1024^3 resolution reconstructed by TopoMesh within 5s.

Abstract

*The dominant paradigm for high-fidelity 3D generation relies on a VAE-Diffusion pipeline, where the VAE's reconstruction capability sets a firm upper bound on generation quality. A fundamental challenge limiting existing VAEs is the representation mismatch between ground-truth meshes and network predictions: GT meshes have arbitrary, variable topology, while VAEs typically predict fixed-structure implicit fields (*e.g.*, SDF on regular grids). This inherent misalignment prevents establishing explicit mesh-level correspondences, forcing prior work to rely on indirect supervision signals such as SDF or rendering losses. Consequently, fine geometric details, particularly sharp features, are poorly preserved during reconstruction. To address this, we introduce TopoMesh, a sparse voxel-based VAE that unifies both GT and predicted meshes under a shared Dual Marching Cubes (DMC) topological framework. Specifically, we convert arbitrary input meshes into DMC-compliant representations via a remeshing algorithm that preserves sharp edges using an $L\infty$ distance metric. Our decoder outputs meshes in the same DMC format, ensuring that both predicted and target meshes share identical topological structures. This establishes explicit correspondences at the vertex and face level, allowing us to derive explicit mesh-level supervision signals for topology, vertex positions, and face orientations with clear gradients. Our sparse VAE architecture employs this unified framework and is trained with Teacher Forcing and progressive resolution training for stable and efficient convergence. Extensive experiments demonstrate that TopoMesh significantly outperforms existing VAEs in reconstruction fidelity, achieving superior preservation of sharp features and geometric details, as shown in Fig. 1.*

[†]Corresponding authors.

1. Introduction

Deep generative models are rapidly transforming 3D content creation, with applications spanning gaming, virtual reality, robotics, and computer-aided design. Among various approaches, 3D native diffusion models [55, 61] have emerged as a leading paradigm, attributed to their superior generation quality, strong generalization, and scalability. These models operate in a latent space, requiring a powerful Variational Autoencoder (VAE) [19] to compress irregular, topologically-varying meshes into regular latent representations and reconstruct them reliably. Consequently, the VAE's reconstruction fidelity acts as the primary bottleneck, setting a firm upper bound on the generation quality.

A fundamental challenge limiting existing VAEs is the *representation mismatch* between ground-truth meshes and network predictions. Ground-truth meshes exhibit arbitrary topology, such as irregular connectivity and varying vertex counts, while VAEs typically predict fixed-structure representations. This inherent structural gap prevents establishing explicit mesh-level correspondence, forcing previous methods to rely on indirect supervision signals, such as SDF or rendered images, which introduce distinct limitations. Early methods like 3DShape2VecSet [61], Clay [63], and TripoSG [22] predict implicit SDF and extract meshes via Marching Cubes (MC) [30], which constrain vertices to lie on grid edges, inevitably smoothing sharp edges and corners. Recent works like Trellis [55] and SparseFlex [13] replace MC with the more expressive FlexiCubes [39] decoder but pivot to rendering-based supervision, which introduces supervisory ambiguity. For example, gradients for fine geometric details are often lost due to limited resolution, occlusion, and sparse viewpoints. Other methods like Direct3D-S2 [54] and Sparc3D [23] return to SDF supervision and MC-based extraction, thus still suffering from the grid-alignment constraints. Consequently, a critical question remains: **How can we design a VAE that possesses both the expressive power to faithfully represent sharp features and the structural alignment necessary to enable direct, unambiguous supervision on mesh?**

To answer this question, we introduce TopoMesh, a novel framework that resolves the representation mismatch through Topological Unification. By ensuring both network predictions and ground-truth geometry share a unified Dual Marching Cubes (DMC) [34] framework, our method establishes explicit correspondence at the vertex and face level. This enables, for the first time, direct supervision on mesh topology, vertex positions, and face orientations.

To realize topological unification on both ground-truth geometry and network outputs, TopoMesh introduces two core components. First, Topo-Remesh, a robust, fully GPU-accelerated algorithm designed to convert arbitrary input meshes into feature-preserving, DMC-compliant representations. Unlike traditional methods [2, 63] that use L2 distance, a point-to-point metric that rounds sharp corners, we introduce an $L\infty$ distance that incorporates local surface structure to preserve sharp features. Combined with a Manifold Dual Contouring extractor [16], Topo-Remesh produces high-fidelity watertight outputs at 1024^3 resolution in approximately 15 seconds. Second, Topo-VAE operates on the unified DMC topology with a sparse encoder and a decoupled decoder. Our encoder employs sparse voxel-point cross-attention mechanism, where each voxel attends only to points within it, enabling efficient processing of millions of points. Our decoder builds upon FlexiCubes but decouples the parameters into topology and geometry components. Combined with the DMC framework that establishes correspondences, this enables direct supervision on topology, positions, and orientations. These two components synergistically enable high-fidelity mesh autoencoding with direct supervision.

With the unified architecture in place, a principled training strategy is essential for realizing the full potential of our VAE. In preliminary experiments, we observed severe instability when applying geometry losses only upon correct topology prediction. The root cause is a tug-of-war: when topology becomes correct, its loss vanishes while large geometry losses suddenly activate, introducing gradients that often flip the topology back to incorrect and prevent convergence. To break this cycle, we introduce Teacher Forcing: during training, we provide the decoder with ground-truth topology, allowing geometry components to learn under stable, correct topological configurations. At test time, the decoder independently predicts both. The strategy is complemented by ground-truth guided voxel pruning and progressive resolution training to accelerate overall convergence.

Our main contributions are: **(1)** A novel paradigm for VAE-based mesh autoencoding that unifies predictions and ground truth under a shared DMC topological framework, enabling direct supervision on fundamental mesh attributes (topology, positions, orientations). **(2)** A robust, fully GPU-accelerated remeshing algorithm that converts arbitrary input meshes into DMC-compliant representations while preserving sharp geometric features. **(3)** A sparse voxel-based VAE with efficient voxel-point cross-attention encoding and a FlexiCubes-based decoder with topology-geometry decoupling. **(4)** A comprehensive training strategy with Teacher Forcing, voxel pruning, and progressive resolution training. **(5)** Extensive experiments demonstrating state-of-the-art reconstruction fidelity, achieving 8% improvement in F-Score and superior sharp feature preservation compared to existing methods.

2. Related Work

3D Generation. Existing 3D generation methods roughly fall into three categories. DreamFusion [36] and subsequent

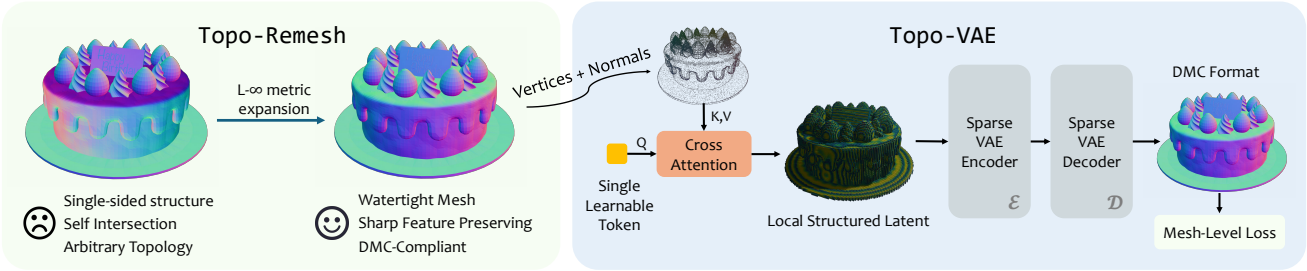

Figure 2. TopoMesh comprises two core modules. **Topo-Remesh** converts imperfect, real-world meshes into clean, DMC-compliant representations while preserving sharp features. **Topo-VAE** takes vertices with normals as input and reconstructs the mesh in the same DMC format. With topological unification, we apply direct supervision on mesh attributes: topology, vertex positions, and face orientations.

works [1, 24, 27, 32, 37, 44, 47, 49] leverage a pre-trained 2D diffusion model to optimize the radiance fields [18, 33] via Score Distillation Sampling (SDS). Multi-view generation [28, 29, 40, 41, 46] followed by sparse-view reconstruction [14, 20, 26, 43, 50, 51, 57, 58, 62, 66] forms a fast pipeline, but is limited in geometric fidelity and generation stability. 3D native diffusion models [3, 6, 13, 22, 23, 25, 38, 45, 53–56, 61, 63] have emerged as a leading paradigm for their high quality and strong generalization. Crucially, these models operate in the latent space, requiring a powerful VAE for high-fidelity mesh reconstruction.

3D Shape VAEs. Modern 3D VAEs can be broadly categorized by latent representation. VecSet-based VAEs [2, 21, 22, 45, 60, 61, 63, 64] represent shapes as a global set of latent vectors. While effective at capturing overall shape, their global nature limits the modeling of fine-grained details. Sparse Voxel-based VAEs [13, 23, 38, 52, 54, 55, 59], which our work builds upon, have emerged as the state-of-the-art for high-resolution modeling. These methods leverage sparse voxel grids to represent complex shapes. However, they are caught in a difficult trade-off regarding supervision. On one hand, methods like Trellis [55] and SparseFlex [13] employ rendering-based supervision on an expressive FlexiCubes [39] decoder, thus suffer from ambiguous gradients for fine geometric details. On the other hand, methods like Direct3D-S2 [54] and Sparc3D [23] revert to SDF supervision but use MC-based decoders, re-encountering sharp feature loss during the extraction stage.

Isosurface Extraction. Marching Cubes [30] generates vertices via linear interpolation on grid edges, which inherently prevents the representation of sharp features. To overcome this, Dual Contouring [17] and Dual Marching Cubes [34] generate a topologically dual mesh with vertices freely placed inside each grid cell, dramatically enhancing expressive power for sharp features. This dual paradigm has been made fully differentiable for end-to-end learning in FlexiCubes [39] and adapted for occupancy fields in ODC [16]. Neural Methods such as NMC [4] and NDC [5] replace analytical rules with neural networks that directly predict vertex positions, but suffer from either complex parameterizations or a lack of manifold guarantees.

3. Method

In this section, we present TopoMesh for high-fidelity mesh autoencoding, as demonstrated in Fig. 2. We first introduce the architecture of *Topo-VAE* (Sec. 3.1), including sparse voxel-point encoding and decoupled decoding, which outputs meshes under DMC format. We then describe the explicit mesh-level loss (Sec. 3.2) and training strategy (Sec. 3.3) that realize the full potential of our VAE. Finally, we introduce *Topo-Remesh* (Sec. 3.4), a robust algorithm that converts any mesh into a clean, DMC-compliant representation while preserving sharp features.

3.1. Topo-VAE

The central objective of Topo-VAE is efficient encoding and differentiable decoding of mesh attributes, *i.e.*, vertex placement and connectivity. This allows us to directly apply supervision on these fundamental attributes, which we argue is key to preserving fine geometric details. Specifically, we encode an input mesh, represented by its vertices V_i and its normals N_i, into compact, sparse voxels, and decode them back into an explicit mesh defined by output vertices V_o and faces F_o. This process can be succinctly formulated as:

$$z = \mathcal{E}(V_i, N_i), \quad (V_o, F_o) = \mathcal{D}(z) \tag{1}$$

where $\mathcal{E}$ and $\mathcal{D}$ represent encoder and decoder, respectively. We detail the core design of these components below.

Sparse Voxel-Point Encoder. We treat input mesh vertices as a point cloud and encode them into sparse voxel features. Naively encoding dense points into sparse voxels is computationally intractable. For example, attending from $20k$ sparse voxels at 64^3 resolution to 2M input points requires an attention map of 74GB. Our key observation is that each point lies exclusively within a single voxel, enabling us to replace full attention with sparse local attention where each point interacts only with its enclosing voxel. As illustrated in Fig. 3a, this local mechanism compresses the attention map: each column contains only one non-zero entry, reducing storage from $O(N \times P)$ to $O(P)$ (74GB to 3.8MB in our example), where N and P denote the number of voxels and points, respectively. We further reduce computation

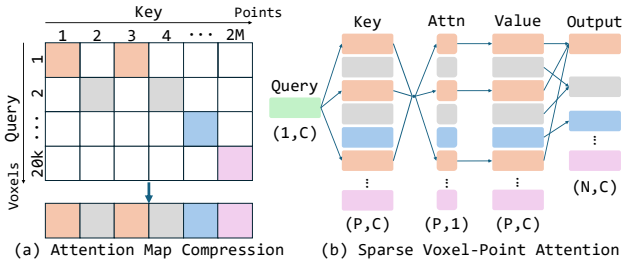

Figure 3. (a) Compression of the sparse attention map. (b) A single, shared query aggregates all keys from the same voxel to form the output feature of this voxel.

by normalizing point coordinates within each voxel to local coordinates, which allows all voxels to share a single learnable query token, as demonstrated in Fig. 3b. The sparse voxel-point cross-attention is formulated as follows:

$$O_i = \sum_{j=1}^{n_i} \text{Softmax}_i\left(\frac{QK_j^T}{\sqrt{d}}\right) \cdot V_j \tag{2}$$

where O_i is the output feature of voxel i, computed by aggregating features of n_i points within it. The query Q is the shared learnable token, and K_j and V_j are linear projections of the j-th point features within voxel i. The Softmax normalizes attention scores over all points within the voxel.

Decoupled Topology and Geometry Decoder. Our decoder builds upon FlexiCubes [39], which parameterizes mesh generation through: SDF values s and deformation vectors δ for voxel corners, and interpolation weights α, β, γ for voxels controlling the precise vertex placement. However, the coupled nature of SDF in standard FlexiCubes creates training instability: the sign of s determines topology, while its magnitude affects geometry, resulting in entangled supervision signals that compete during optimization.

We redesign this process to enable direct mesh-level supervision and stable training. Specifically, we decouple the SDF s into occupancy o (sign) and magnitude u, categorizing parameters into topology and geometry components:

$$\text{Topo} = \{o, \gamma\}, \quad \text{Geom} = \{u, \alpha, \beta, \delta\} \tag{3}$$

This separation allows mesh generation to proceed in two stages: output faces F_o are determined exclusively by topological parameters, while vertices V_o depend on all parameters:

$$F_o = \text{DMC}(o) \tag{4}$$

$$V_o = \text{FlexiCubes}(o \times u, \alpha, \beta, \delta, \gamma) \tag{5}$$

where DMC extracts mesh connectivity from the Dual Marching Cubes framework, and FlexiCubes interpolates vertex positions differentiably. This explicit separation is crucial for stable training: by independently supervising topology (via occupancy) and geometry (via vertex placements), we prevent the "tug-of-war" instability that arises when entangled topology and geometry losses compete during optimization (detailed in Sec. 3.3).

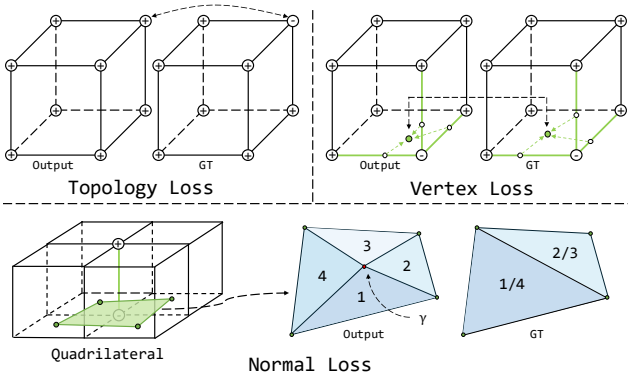

Figure 4. Illustration of explicit mesh-level loss.

3.2. Explicit Mesh-Level Loss

Our topological unification framework enables direct supervision on mesh attributes, *i.e.*, topology, vertex positions, and surface orientations, as illustrated in Fig. 4. Rather than supervising intermediate representations such as SDF values, our losses operate directly on the output mesh. Topology is supervised via occupancy o at grid corners, while vertex and normal losses drive the optimization of all geometric parameters (u, α, β, δ) to produce accurate geometry.

Topology Loss. We supervise occupancy o at grid corners with Binary Cross-Entropy (BCE) Loss against the ground-truth signs o_{gt}:

$$\mathcal{L}_{\text{topo}} = \text{BCE}(o, o_{\text{gt}}) \tag{6}$$

Vertex Loss. To supervise vertex placement, we apply L1 loss directly on vertex positions v against ground-truth positions v_{gt}.

$$\mathcal{L}_{\text{vert}} = \text{L1}(v, v_{\text{gt}}) \tag{7}$$

Normal Loss. To supervise surface orientation and quad triangulation, we apply L1 loss on face normals n. DMC produces quadrilaterals, which FlexiCubes triangulates using the γ parameters. However, FlexiCubes employs different triangulation strategies: splitting each quad into four triangles during training but only two during inference. To provide supervision for the finer triangulation, we duplicate each ground-truth triangle to supervise its corresponding pair of predicted triangles, as illustrated in Fig. 4:

$$\mathcal{L}_{\text{normal}} = \text{L1}(n, n_{\text{gt}}) \tag{8}$$

Total Loss. In addition to mesh-level losses, we include regularization terms: the FlexiCubes regularization loss L_{reg} [39], a consistent loss L_{con} that penalizes variance of attributes on grid corners [55], voxel pruning loss L_{occ} during upsampling [13], and the standard KL-divergence loss L_{KL} [19]. The final loss is a weighted combination:

$$\begin{aligned} L = &\lambda_{\text{topo}}\mathcal{L}_{\text{topo}} + \lambda_{\text{vert}}\mathcal{L}_{\text{vert}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} \\ &+ \lambda_{\text{occ}}\mathcal{L}_{\text{occ}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \end{aligned}$$

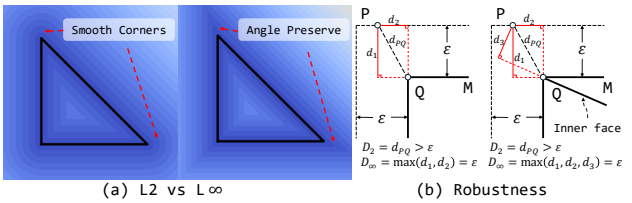

Figure 5. (a) Angle-preserving of L∞ metric. (b) 2D illustration of L∞ distance. P lies on the isosurface that preserves angles.

3.3. Training Scheme

Our SDF decoupling design (Sec. 3.1) provides the foundation for stable training by separating topology and geometry into independent parameter sets. However, we empirically find this insufficient to prevent training instability. We trace the remaining challenges to a "tug-of-war" dynamic between topology and geometry supervision. Initially, geometry losses provide no meaningful signal as the network struggles with topology. Once the network begins predicting correct topology for a region, geometry losses suddenly activate with large magnitudes, introducing disruptive gradients that destabilize topology learning. This creates an oscillation where neither topology nor geometry converges stably. To break this cycle, we introduce several training strategies that leverage our decoupled architecture to stabilize and accelerate convergence.

Teacher Forcing. To bypass the conditional dependency between topology and geometry, we employ Teacher Forcing during training. Instead of using the network's predicted occupancy in Eq. 5, we provide ground-truth topology o_{gt} to the decoder. This ensures that geometric parameters receive stable, meaningful gradients from the first iteration, as they operate under correct topological configurations. While this introduces a discrepancy between training and inference, we observe a negligible impact on reconstruction quality, which we attribute to the smoothness of our latent space and the decoder's ability to predict accurate topology after sufficient training.

GT-guided Voxel Pruning. Voxel pruning is essential for training efficiency at high resolutions. However, pruning based on early network predictions can be detrimental, often removing critical voxels and creating holes [13, 23] and training instability [38]. We adopt a GT-guided strategy that preserves sparse voxels within a narrow band around the ground-truth surface during training, ensuring both geometric integrity and computational efficiency.

Progressive Resolution Training. Our local-centric encoder design is adaptive to multiple resolutions, enabling progressive training. We begin at coarse resolution, allowing the model to capture global shape, then progress to finer resolutions where the model refines geometric details. This coarse-to-fine curriculum accelerates overall training convergence.

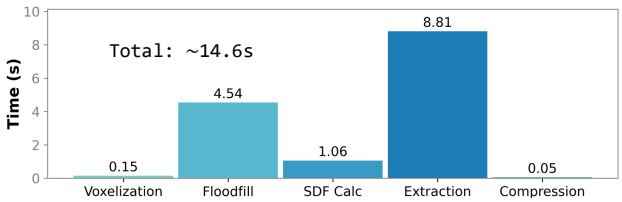

Figure 6. Remesh pipeline and execution time.

3.4. Topo-Remesh

Our approach demands that ground-truth meshes share the same DMC structure as our decoder to enable the direct mesh-level correspondences (Sec. 3.1, 3.2). To meet this requirement, we develop Topo-Remesh, a remeshing pipeline that converts arbitrary meshes into DMC-compliant representations while preserving geometric fidelity. A fundamental challenge in this process is surface dilation, a necessary step to ensure the output encloses a proper volume. Prior methods [2, 63] rely on L2-based dilation, which inherently smooths sharp corners, as illustrated in Fig. 5a. While post-processing methods [15, 23, 35] have been proposed to recover sharp features, they often introduce artifacts such as self-intersections and surface noise.

L∞ Metric. The limitation of L2 distance stems from its point-wise nature. As shown in Fig. 5b, for a spatial point P and its nearest point Q on mesh M, standard L_2 distance measures only the direct distance between P and Q, ignoring the local surface structure around Q and thus rounding corners during dilation. To preserve angles, we need a distance formulation that incorporates local geometry. Inspired by properties of the L∞ norm [8], we define the distance from point P by considering all triangles $T(Q)$ incident to Q. Our L∞ distance is the maximum L2 distance from P to the plane of each incident triangle:

$$D_\infty(P, Q) = \max_{T_i \in T(Q)} d(P, \Pi_i) \tag{9}$$

where Π_i is the plane containing triangle T_i, and $d(P, \Pi_i)$ denotes the Euclidean distance from point P to plane Π_i. The formulation is inherently angle-preserving: inflating the local surface at Q by offsetting each incident plane outward by ε creates a polyhedral envelope, and P lies on its boundary. This formulation is robust to topological defects and incorrect normals, as it requires no knowledge of topology or face directions, and is also robust to internal structure. Please refer to the supplementary for more details.

Fully GPU-accelerated Remeshing. As shown in Fig. 6, our pipeline consists of five sequential stages: Voxelization, Flood-fill, SDF Calculation, Isosurface Extraction, and Compression. We first voxelize the input mesh, then use flood-fill to locate voxels within a narrow band around the surface boundary. We compute L∞ distance at grid corners to further identify surface-intersection voxels. For surface extraction, we employ ODC [16], which extracts meshes

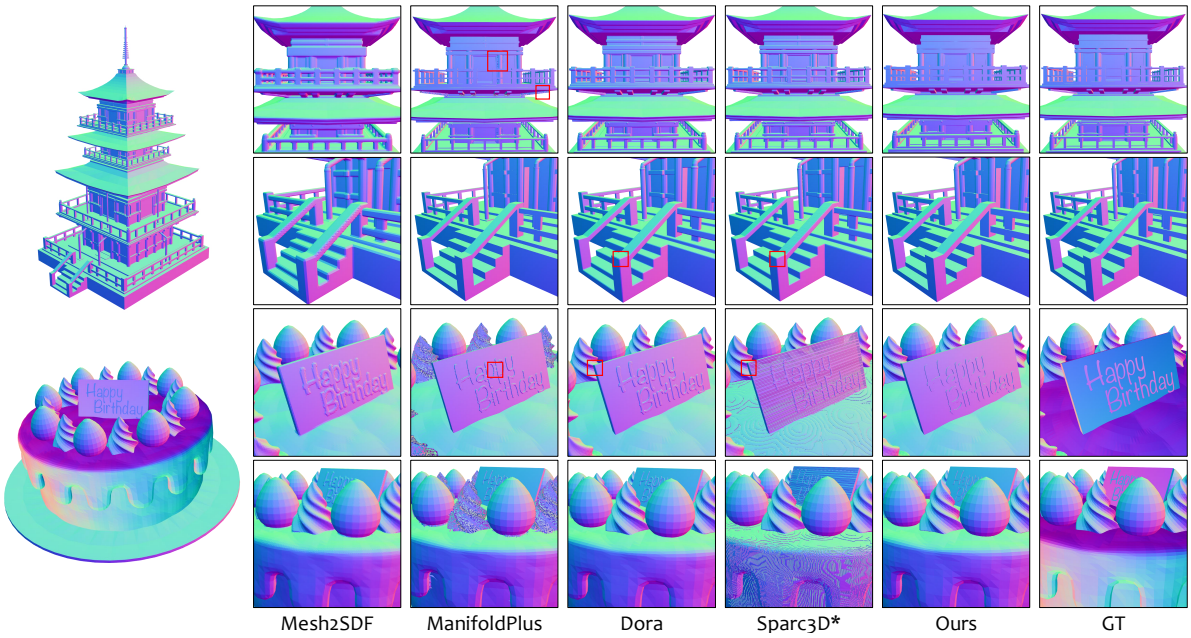

Figure 7. Visual comparison of remeshing. Our method produces clean meshes with sharp features. $*$ means a re-implementation version.

Table 1. Quantitative result of remeshing. Our method preserves fine details with the highest efficiency. CD is multiplied by 10^5.

Method	Device	Thingi10K [65]			Objaverse [9]			Efficiency	
		CD ↓	F1 ↑	ANC ↑	CD ↓	F1 ↑	ANC ↑	Time ↓	Size ↓
Mesh2SDF [48]	CPU	1.980	0.932	0.955	3.349	0.931	0.829	552.3	66.5
ManifoldPlus [15]	CPU	**1.347**	0.976	0.981	**0.620**	**0.991**	0.780	79.4	92.7
Dora [2]	GPU	1.492	0.972	0.970	1.057	0.987	0.961	116.3	112.6
Sparc3D [23]	GPU	1.436	**0.978**	0.975	2.864	0.970	0.929	175.9	121.6
Ours	GPU	1.479	0.978	**0.984**	0.964	0.988	**0.964**	**18.5**	**28.7**

that preserve sharp features through iteratively querying occupancy on grid edges and faces to compute the optimal vertex positions. All stages are implemented on GPU, enabling end-to-end remeshing at 1024^3 resolution in approximately 15 seconds.

DMC-based Mesh Compression. We design an efficient compression scheme for high-resolution DMC-based mesh structures. Specifically, for meshes up to 1024^3 resolution, we store a set of primitives, including integer coordinates of valid voxels (3×10 bits), occupancy of voxel corners (8 bits), intra-voxel vertex offsets (3×10 bits), and triangulation decisions (3 bits), which can be quickly reassembled to form a mesh. This representation achieves a 76% compression ratio, competitive with Draco [12]'s 84%, while being orders of magnitude faster at 0.05 seconds versus Draco's 7 seconds for encoding and decoding.

4. Experiment

We evaluate our method on both remeshing and autoencoding tasks. We first describe the experimental setup (Sec. 4.1), then present results for Topo-Remesh (Sec. 4.2) and Topo-VAE (Sec. 4.3), followed by ablation studies (Sec. 4.4). Finally, we present image-to-3D generation results using Topo-VAE (Sec. 4.5).

4.1. Setting

Dataset. The training dataset contains $320k$ high-quality meshes from Sketchfab [9]. For evaluating remeshing, we sample 500 objects each from Objaverse [9] and Thingi10K [65] (1k total). For autoencoding evaluation on complex geometries, we use the L3-L4 subset of Dora-Bench [2], which comprises 1.4k objects from GSO [11], ABO [7], Meta [10], and Objaverse [9]. We further introduce *Topo-Bench*, a supplementary benchmark of 1k objects from Objaverse [9], Thingi10K [65], and Toys4K [42], selected based on their number of sharp edges.

Training. Training is conducted on 64 A800 GPUs using AdamW [31] with a constant learning rate of 0.0001. We employ the progressive resolution schedule for both VAE and DiT. The VAE is trained with a batch size of 64 for 160k, 160k, and 380k steps at 32^3, 64^3, and 128^3 resolutions, respectively (700k steps total), and DiT is trained with a batch size of 512 for 200k, 300k, and 300k steps at 32^3, 64^3, 128^3 resolutions. (800k steps total).

4.2. Remesh

Baselines. We compare against Mesh2SDF [48], ManifoldPlus [15], Dora [2], and a re-implemented version of

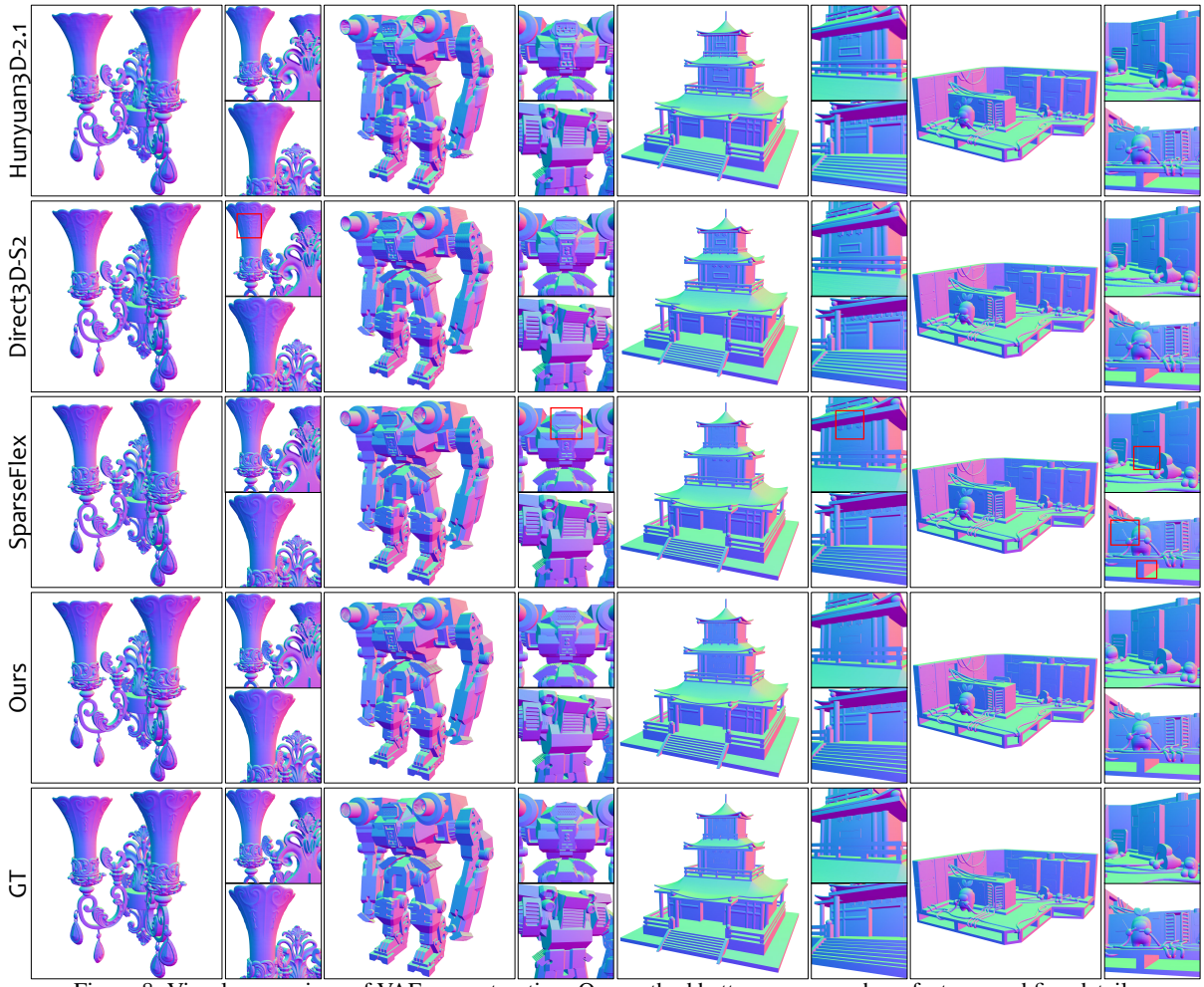

Figure 8. Visual comparison of VAE reconstruction. Our method better preserves sharp features and fine details.

Sparc3D [23]. We use Chamfer Distance (CD), F1 Score (F1), and Absolute Normal Consistency (ANC) as metrics, with all outputs reconstructed at 1024^3 resolution.

Quality. We present both qualitative and quantitative comparisons in Fig. 7 and Tab. 1. All L2-based methods, including Mesh2SDF, Dora, and Sparc3D, fail to preserve sharp edges and corners, as highlighted by red boxes. While projection-based (ManifoldPlus) and rendering-based (Sparc3D) refinement can slightly improve 3D metrics, they often introduce artifacts, such as self-intersections and noise. In contrast, our method produces clean meshes that faithfully preserve sharp features, such as the railings and handrails of the attic and birthday cards.

Efficiency. Tab. 1 shows that our method processes 1024^3-resolution meshes in 18.5 seconds, significantly faster than baselines, which require at least one minute. Our compression yields compact outputs, averaging 28.7MB per mesh.

4.3. Autoencoding

Baselines. We compare against VecSet-based VAEs including TripoSG [22], Dora [2], Hunyuan3D-2.1 [45]; and Sparse Voxel-based VAEs including Trellis [55], Direct3D-S2 [54], SparseFlex [13]. Beyond standard metrics (CD, F1, ANC), we introduce F1-Sharp to quantify sharp feature preservation: an F1 Score computed on points sampled near edges and corners with dihedral angles sharper than 30 degrees, using the sharp edge sampling method from Dora [2]. All methods generate meshes at 1024^3 resolution, except Trellis (256^3).

Reconstruction Quality Fig. 8 and Tab. 2 show that our method better reconstructs sharp corners (90-degree turns on sharp corners) and fine geometric details (holes on the robot), whereas all baselines fail to recover. While SparseFlex achieves comparable metrics on overall shapes, our method uses only a quarter of its tokens and demonstrates significantly better sharp feature preservation (5.9% and 7.1% improvement in F1-Sharp on the two benchmarks).

4.4. Ablation Study

L∞ Distance. We replace L∞ with L2 distance in our remeshing pipeline and visualize the dihedral angle distribution in Fig. 9. The histogram shows that our L∞ distance

Table 2. Quantitative result of VAE. All methods except Trellis produce meshes at 1024^3 resolution. Our method outperforms other methods in preserving sharp features. CD is multiplied by 10^5.

Method	#Latent	#Dim	Topo-Bench				Dora-Bench [2]			
			CD ↓	F1 ↑	F1-S ↑	ANC ↑	CD ↓	F1 ↑	F1-S ↑	ANC ↑
TripoSG [22]	4096	64	2.658	0.893	0.715	0.965	1.697	0.959	0.717	0.976
Dora [2]	4096	64	2.167	0.905	0.754	0.968	1.814	0.964	0.768	0.977
Hunyuan3D-2.1 [45]	4096	64	2.538	0.888	0.767	0.965	1.606	0.954	0.770	0.980
Trellis [55]	12832	8	18.616	0.583	0.308	0.893	14.716	0.715	0.279	0.915
Direct3D-S2 [54]	76386	16	2.713	0.813	0.694	0.920	2.313	0.881	0.819	0.968
SparseFlex [13]	244691	8	1.840	0.920	0.873	0.992	1.625	0.973	0.844	0.994
Ours	56006	8	1.882	0.917	0.932	0.993	1.126	0.973	0.915	0.995

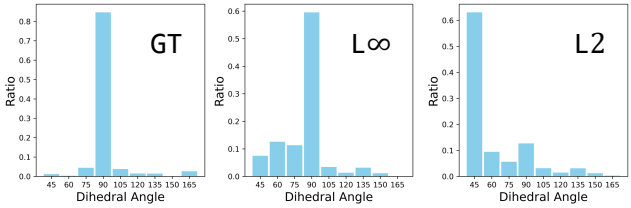

Figure 9. Dihedral angle distributions from different remeshing methods.

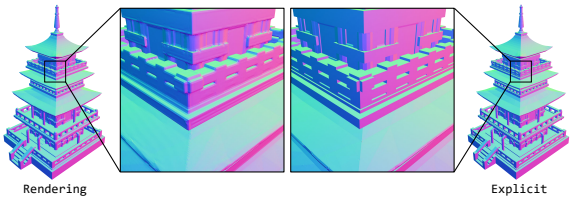

Figure 10. Comparison of direct mesh-level supervision versus rendering-based supervision.

Table 3. Quantitative results of ablations.

Method	CD ↓	F1 ↑	F1-S ↑	ANC ↑
render loss	1.731	0.776	0.711	0.932
mesh-level loss	0.150	0.975	0.991	0.999
32res	1.812	0.933	0.790	0.968
64res	1.693	0.940	0.869	0.983
128res	1.126	0.973	0.915	0.995

preserves the distribution of sharp angles, whereas the L2 variant smooths the geometry, collapsing sharp features into nearly planar surfaces.

Mesh-Level Supervision We conduct a single-shape overfitting experiment to compare direct mesh-level supervision against rendering-based supervision. As shown in Fig. 10 and Tab. 3 (first two rows), rendering-based supervision cannot reconstruct fine details and sharp edges accurately. In contrast, direct supervision enables nearly lossless reconstruction, demonstrating the advantage of our mesh-level correspondence framework.

Multi-Resolution Inference. Our VAE is adaptive to varying resolutions. We show quantitative results on Dora-bench in Tab. 3 (Last three rows), each compared with its corresponding ground truth.

Table 4. Quantitative generation results on Toys4K.

Method	Hunyuan3D-2.1 [45]	Trellis [55]	Direct3D-S2 [54]	Ours
FID ↓	59.43	59.61	45.33	**42.48**
KID ($\times 10^3$) ↓	5.97	6.03	5.47	**4.63**

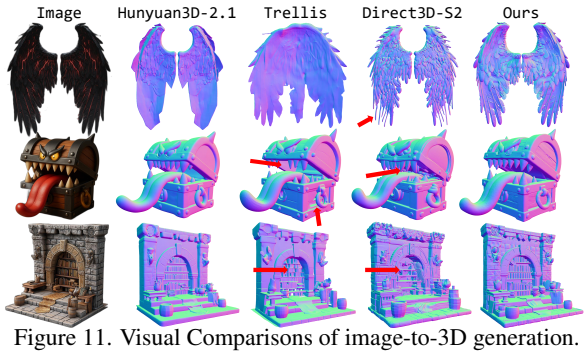

Figure 11. Visual Comparisons of image-to-3D generation.

4.5. 3D Generation

We validate the effectiveness of Topo-VAE as a powerful backbone for image-to-3D generation. We adopt the two-stage generation pipeline from Trellis [55], but replace the structure generation stage with a VecSet-based model for stability, similar to Ultra3D [3]. As shown in Tab. 4 and Fig. 11, our results exhibit sharper geometric details and better alignment with input images, whereas other methods suffer from noise or holes (highlighted by red arrows).

5. Limitation

Our sparse voxel-based VAEs generate millions of voxels when upsampling to high resolutions, which requires significant computational resources and time. Our remesh algorithm is limited by base resolution and therefore struggles to capture extremely fine details smaller than the voxel size.

6. Conclusion

In this paper, we present TopoMesh, a novel framework for high-fidelity mesh autoencoding. We identify the *representation mismatch* as the key bottleneck in existing VAEs and propose the principle of Topological Unification, which enables direct supervision of fundamental mesh attributes. Extensive experiments demonstrate that our VAE achieves superior performance in reconstruction fidelity, especially in sharp features and fine geometric details.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2023YFF0905104), the Natural Science Foundation of China (No. 62132012, 62361146854), and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[2] Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 5, 6, 7, 8

[3] Yiwen Chen, Zhihao Li, Yikai Wang, Hu Zhang, Qin Li, Chi Zhang, and Guosheng Lin. Ultra3d: Efficient and high-fidelity 3d generation with part attention. *arXiv preprint arXiv:2507.17745*, 2025. 3, 8

[4] Zhiqin Chen and Hao Zhang. Neural marching cubes. *ACM Trans. Graph.*, 2021. 3

[5] Zhiqin Chen, Andrea Tagliasacchi, Thomas A. Funkhouser, and Hao Zhang. Neural dual contouring. *ACM Trans. Graph.*, 2022. 3

[6] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[8] Wikipedia contributors. L-infinity. https://en.wikipedia.org/wiki/L-infinity. 5

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[10] Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, Sean Christofferson, James Fort, Xiaqing Pan, Mingfei Yan, Jiajun Wu, Carl Yuheng Ren, and Richard A. Newcombe. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6

[11] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *International Conference on Robotics and Automation (ICRA)*, 2022. 6

[12] Google. Draco 3d data compression. https://github.com/google/draco. 6

[13] Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 3, 4, 5, 7, 8

[14] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[15] Jingwei Huang, Yichao Zhou, and Leonidas J. Guibas. Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. *arXiv preprint arXiv:2005.11621*, 2020. 5, 6

[16] Jisung Hwang and Minhyuk Sung. Occupancy-based dual contouring. In *SIGGRAPH Asia*, 2024. 2, 3, 5

[17] Tao Ju, Frank Losasso, Scott Schaefer, and Joe D. Warren. Dual contouring of hermite data. *ACM Trans. Graph.*, 2002. 3

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 3

[19] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 2, 4

[20] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[21] Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native diffusion and interactive geometry refiner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[22] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 3, 7, 8

[23] Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025. 2, 3, 5, 6, 7

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

[25] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[26] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. Meshformer : High-quality mesh generation with 3d-guided reconstruction model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

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

[28] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[29] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[30] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, 1987. 2, 3

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[32] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[34] Gregory M. Nielson. Dual marching cubes. In *IEEE Visualization*, 2004. 2, 3

[35] Seonghun Oh, Xiaodi Yuan, Xinyue Wei, Ruoxi Shi, Fanbo Xiang, Minghua Liu, and Hao Su. Pamo: Parallel mesh optimization for intersection-free low-poly modeling on the GPU. In *Computer Graphics Forum*, 2025. 5

[36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[37] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[38] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 5

[39] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 2023. 2, 3, 4

[40] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[41] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[42] Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[43] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[44] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[45] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 3, 7, 8

[46] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[47] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[48] Peng-Shuai Wang, Yang Liu, and Xin Tong. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Trans. Graph.*, 2022. 6

[49] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[50] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. CRM: single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[51] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 3

[52] Guanjun Wu, Jiemin Fang, Chen Yang, Sikuang Li, Taoran Yi, Jia Lu, Zanwei Zhou, Jiazhong Cen, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Xinggang Wang, and Qi Tian. Unilat3d: Geometry-appearance unified latents for single-stage 3d generation. *arXiv preprint arXiv:2509.25079*, 2025. 3

[53] Shuang Wu, Youtian Lin, Yifei Zeng, Feihu Zhang, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[54] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2, 3, 7, 8

[55] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 4, 7, 8

[56] Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. Octfusion: Octree-based diffusion models for 3d shape generation. In *Computer Graphics Forum*, 2025. 3

[57] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[58] Yinghao Xu, Zifan Shi, Yifan Wang, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[59] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[60] Biao Zhang and Peter Wonka. Lagem: A large geometry model for 3d representation learning and diffusion. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[61] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 2023. 2, 3

[62] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[63] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Trans. Graph.*, 2024. 2, 3, 5

[64] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[65] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10, 000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 6

[66] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3