DiffGI: Differentiable Geometry Images for High-Fidelity Thin-Shell 3D Generation

Eungjune Shim, Hansol Lee, and Eunjung Ju

CLO Virtual Fashion Inc., South Korea
e.june.shim@gmail.com, {river, kate}@clo3d.com

Abstract. Existing 3D generative models predominantly rely on implicit volumetric representations, which inherently enforce watertight topology and struggle to faithfully represent thin-shell and non-manifold geometries such as garments. While geometry image-based approaches offer a surface-centric alternative, existing methods typically rely on discrete binary occupancy maps whose resolution-dependent boundary encoding causes staircase artifacts and information loss upon downsampling, while surface reconstruction remains a non-differentiable post-processing step disconnected from the learning pipeline.

To address this, we propose Differentiable Geometry Image (DiffGI)[1], an end-to-end 3D-to-2D mapping framework that seamlessly integrates surface representation and geometric optimization. DiffGI replaces conventional binary maps with a continuous 2D Truncated Signed Distance Function (TSDF), which encodes boundary position at subpixel precision within a fixed grid resolution, effectively eliminating resolution-dependent staircase artifacts even under aggressive downsampling. Building on this continuous field, we introduce a differentiable Marching Squares algorithm based on analytical linear interpolation, allowing gradients from 3D surface losses to propagate seamlessly back to the 2D latent space.

Leveraging this differentiable pipeline, we train a DiffGI-VAE augmented with a geometry-aware normal rendering loss to compress complex 3D surfaces into an ultra-compact 32×32 latent space. Finally, we instantiate a transformer-based latent diffusion model on top of this space for conditional 3D generation, showing that the proposed representation readily supports efficient generative modeling. Extensive experiments on garment and object datasets demonstrate that our method achieves superior reconstruction fidelity and boundary precision compared to prior geometry-image and voxel-based approaches, while requiring significantly fewer computational resources.

Keywords: 3D Mesh Generation · Differentiable Geometry Images · Truncated Signed Distance Function · Thin-Surface Modeling · Latent Diffusion Models

[1] Project page: https://ejshim.github.io/diffgi/

1 Introduction

Recent advances in combining large-scale data with deep learning have driven rapid progress in 3D generation. Neural implicit field representations such as SDFs, volumetric occupancy, and NeRFs learn continuous 3D spatial functions, enabling differentiable volume rendering and 3D distillation from 2D diffusion models.

However, these methods generally assume watertight surfaces, posing fundamental limitations in representing thin-shell and open-boundary structures critical in practical applications. When generating digital garments or furniture frames using implicit fields, networks often fail to maintain thin surfaces, resulting in artificial thickness or front-back blending. Moreover, meshes obtained via Marching Cubes lack UV coordinates, making them difficult to use in industrial pipelines requiring physics simulation and material editing.

Geometry image representations—which parameterize 3D surfaces onto 2D grids (specifically, multi-chart geometry images [32] that partition a surface into multiple UV charts)—offer an alternative, treating coordinates and attributes as image-like tensors and enabling direct use of 2D generative models. Omages [46], GIMDiffusion [11], and GarmageNet [19] have demonstrated this approach for UV-friendly 3D generation. However, existing methods rely on binary occupancy maps that are resolution-dependent and induce staircase artifacts, often requiring high-resolution grids (e.g., 768×768 [11]) to partially compensate. Moreover, post-processing steps such as boundary snapping are non-differentiable, creating a structural disconnect between learning and mesh reconstruction.

Meanwhile, Differentiable Marching Cubes, DMTet, and FlexiCubes have made iso-surface extraction differentiable for implicit representations, but operate on 3D grids with cubic memory scaling and without natural UV parameterization. Our goal is to bring the advantages of differentiable iso-surface extraction to the geometry image family on 2D UV grids, achieving lighter computation and UV-friendly meshes simultaneously.

We propose Differentiable Geometry Image (DiffGI), which replaces binary occupancy with a continuous 2D TSDF and introduces Differentiable Marching Squares (DMS) for fully differentiable mesh reconstruction. This allows gradients from 3D surface losses to propagate seamlessly to the 2D latent space. We build a VAE with geometry-aware normal rendering loss and a transformer-based latent diffusion model, demonstrating efficient generation of thin non-manifold meshes in an ultra-compact 32×32 latent space (Figure 2). As illustrated in Figure 1, these design choices yield noticeably sharper boundaries and better preservation of thin-shell structures compared to occupancy-based methods.

Our contributions are:

1. **Subpixel 2D Boundary Representation:** A geometry-image representation that replaces the binary occupancy maps of prior work with a continuous 2D TSDF. While the advantage of signed distance fields over occupancy grids is well established for volumetric 3D representations, we realize it in the 2D multi-chart geometry-image setting, where the continuous field encodes boundary position at subpixel precision within a fixed grid resolution

and removes the resolution dependency and staircase artifacts of occupancy maps.

2. **Fully Differentiable Mesh Reconstruction:** A tensor-based Differentiable Marching Squares algorithm that makes 3D reconstruction amenable to backpropagation, bridging the gap between 2D optimization and 3D mesh recovery.
3. **Ultra-Compact Latent Space for Efficient Generation:** The TSDF representation and geometry-aware normal rendering loss enable compression of complex non-manifold surfaces into a 32×32×4 latent space, supporting real-time 3D generation even on consumer-grade hardware.

Fig. 1: VAE reconstruction results with our TSDF-based DiffGI representation and normal rendering loss (left) versus an occupancy-based geometry image without normal loss (right), showing noticeably sharper boundaries and better preservation of thin-shell structures.

2 Related Work

2.1 Implicit and Explicit Representations for 3D Generation

Recent 3D generation has primarily built on implicit field representations—signed/unsigned distance fields, occupancy fields, and radiance fields—which provide continuous geometry and integrate naturally with diffusion- or rendering-based supervision [4,5,8,9,12,14,24,26,28,39,51,52], enabling a wave of large-scale 3D generation and reconstruction systems [2, 16, 20, 41, 43, 45, 48, 50]. UDF-based methods relax the closed-surface bias of SDFs [12, 24, 26, 52], but remain field-based and still require dedicated mesh extraction and post-processing [17,40,44], limiting their applicability to thin garment panels, open boundaries, and downstream tasks such as simulation and material authoring. Explicit surface representations preserve mesh structure and UV parameterization, making them

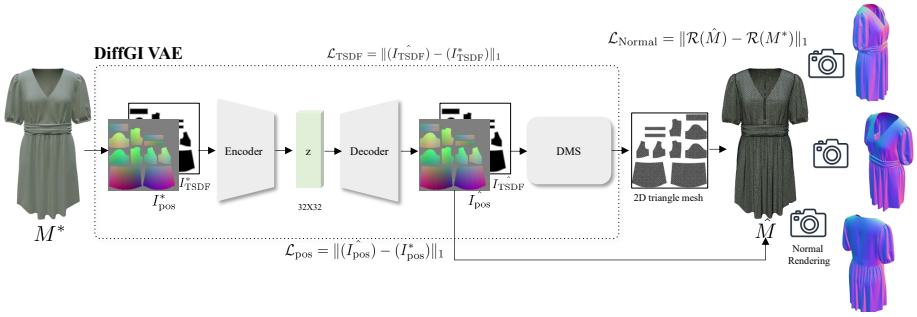

Fig. 2: Overview of the proposed DiffGI framework. The input 3D mesh is first mapped to a 2D TSDF-based DiffGI representation, which is then encoded by a DiffGI-VAE into a compact latent space. The decoder reconstructs the TSDF geometry image, from which a 3D triangle mesh is recovered via Differentiable Marching Squares. Pixel-space losses on the TSDF and position maps, together with a geometry-aware normal rendering loss, are jointly applied to enable end-to-end optimization of high-fidelity 3D surfaces.

more compatible with graphics pipelines [15, 47, 49], yet their irregular discrete topology makes end-to-end modeling challenging. Our work retains a structured 2D surface representation while enabling differentiable, boundary-aware surface recovery.

2.2 Geometry-Image-Based Surface Representations

Geometry images map irregular 3D surfaces onto regular UV-aligned 2D grids, enabling surface coordinates and attributes to be treated as image-like tensors [3, 15, 32, 35]. Recent methods such as GIMDiffusion [11], Omages [46], and GarmageNet [19] show that this representation pairs well with 2D generative backbones for UV-friendly 3D generation. However, these pipelines still rely on binary occupancy or mask channels [11, 19, 46], making boundary localization resolution-dependent and often requiring heuristics such as boundary snapping or chart masking. DiffGI retains the geometry-image prior but replaces binary support with a continuous 2D TSDF on a single 256×256 tensor, enabling continuous boundary localization instead of hard mask-based recovery.

2.3 Differentiable Iso-Surface Extraction

Differentiable iso-surface extraction makes field-to-surface conversion part of the optimization loop. Starting from Marching Cubes [25], numerous methods have pursued this goal, notably DMTet [33] and FlexiCubes [34], along with others [1,6,7,17,21,23,36–38]. These methods consistently demonstrate that surface-level supervision is more effective when extraction participates in optimization. However, most operate on 3D voxel or tetrahedral grids, where memory scales cubically and UV-aligned surfaces are not directly available [1, 6, 7, 17, 21, 33, 34].

DiffGI brings this idea to geometry images: we introduce Differentiable Marching Squares (DMS) over a 2D TSDF in UV space, combining end-to-end surface optimization with the efficiency and UV compatibility of structured 2D representations.

3 Method

3.1 Continuous TSDF Geometry Image Representation

Prior geometry-image-based methods combine a position map encoding 3D coordinates with a binary occupancy map indicating the valid region. However, binary occupancy maps encode boundaries as hard 0/1 transitions, making boundary localization inherently resolution-dependent: upon downsampling to practical training resolutions, fine boundary details are destroyed and staircase artifacts emerge.

To address this, the DiffGI representation replaces the binary occupancy map with a **Truncated Signed Distance Function (TSDF)** defined as a continuous function on the 2D plane. Because the TSDF encodes the distance to the nearest boundary rather than a binary label, it preserves subpixel boundary information even at low resolution, enabling high-fidelity reconstruction from aggressively compressed grids. The Mesh-to-DiffGI conversion is performed as an offline preprocessing step before training, and proceeds as follows:

1. **UV Packing & Global Scaling:** UV patches are arranged on the 2D plane using an AABB-based packing algorithm with a uniform global scale and inter-patch padding to prevent bleeding, iteratively optimized to maximize fill ratio.
2. **Spatial Sampling & Position Map:** The mesh surface is sampled on a 1024×1024 regular grid via barycentric interpolation, producing a 3-channel position map.
3. **Dilation:** Edge pixel values are propagated outward into the undefined background via dilation, preventing boundary corruption during downsampling.
4. **2D TSDF Transform:** For every pixel on the 1024×1024 UV grid, including those in the empty background outside the charts, we compute the signed Euclidean distance in pixels to the nearest chart contour (positive inside a chart, negative outside), clamped at a truncation distance of 15 pixels.
5. **Bilinear Downsampling:** The 1024×1024 tensor is downsampled to $256 \times 256 \times 4$. Unlike binary maps, the continuous TSDF ensures no information destruction or staircase artifacts during downsampling.

3.2 Differentiable Marching Squares (DMS)

When recovering a 3D mesh from the 2D DiffGI tensor generated within a deep learning architecture (DiffGI-to-Mesh), the reconstruction process must be differentiable with respect to the network weights in order to backpropagate geometric losses from 3D space to the 2D model. However, the traditional Marching

Squares algorithm determines topology through a discrete lookup table, blocking gradients at this stage and making learning within the deep learning computation graph infeasible.

To address this, we propose the Differentiable Marching Squares (DMS) module, which formulates vertex positions as continuous functions of the underlying TSDF field. On the 2D TSDF grid, when the corner values ϕ_A and ϕ_B of two adjacent pixels have opposite signs ($\phi_A \cdot \phi_B < 0$), the Intermediate Value Theorem guarantees that a surface boundary (zero-crossing) exists between them. The relative position $x \in [0, 1]$ at which the boundary point lies is derived through linear interpolation. To ensure numerical stability, we employ the following interpolation formula with a regularization constant ϵ:

$$x = \frac{\phi_A}{\phi_A - \phi_B + \epsilon \cdot \mathrm{sgn}(\phi_A - \phi_B)} \tag{1}$$

Here, ϵ (e.g., 10^{-5}) serves to prevent gradient explosion in regions where the two pixel values are nearly equal ($\phi_A \approx \phi_B$), which would otherwise cause the denominator to approach zero. Since $\mathrm{sgn}(\cdot)$ is locally constant and thus treated as a non-differentiable constant during backpropagation, gradients flow solely through the continuous ratio $\phi_A/(\phi_A - \phi_B)$, preserving a smooth optimization landscape. The subpixel vertex 2D UV coordinates extracted through this formula are then transformed into final 3D vertex coordinates $V = (X, Y, Z)$ via bilinear grid sampling on the 3-channel position map tensor.

The key advantage of this algorithm lies in its optimizability during the backward pass. The geometric error (e.g., normal loss) $\mathcal{L}$ computed at the reconstructed 3D vertices V is smoothly propagated back to the 2D TSDF tensor and position map tensor in the form of $\frac{\partial \mathcal{L}}{\partial V}$ via the chain rule. Among the 16 topological configurations of Marching Squares, Case 6 (BR+TL) and Case 9 (BL+TR)—where two pairs of diagonal corners have opposite signs—correspond to topologically ambiguous saddle point cases. In these cases, both interpretations of connecting or separating the two regions are mathematically valid. In this work, we deterministically adopt the convention of always treating the two patches as separate independent regions. While the topological configuration choice itself is discrete, the coordinates of each boundary vertex within the chosen topology remain continuous functions of the TSDF values according to Eq. (1). Therefore, the backpropagation path is fully preserved even in ambiguous cases. The module is implemented in a vectorized fashion using PyTorch tensor operations, guaranteeing significantly faster computation at $O(N^2)$ complexity compared to the $O(N^3)$ complexity of existing 3D volumetric extraction methods (DMC, DMTet, etc.).

3.3 Geometry-Aware Latent Compression (DiffGI-VAE)

Directly generating high-quality 3D surfaces in the high-dimensional tensor space of $256 \times 256 \times 4$ resolution is limited in terms of memory and training efficiency. We introduce the DiffGI-VAE framework to compress this data into a much

smaller and more compact space. To accelerate convergence, we initialize the DiffGI-VAE with pretrained VAE weights from Stable Diffusion 1.5 [31], which has learned general-purpose spatial compression priors from large-scale natural images. The channel count of the first convolutional layer is expanded to accommodate the 4-channel input scheme (Position 3 + TSDF 1), and zero-initialization is applied to the newly added weights to prevent catastrophic forgetting.

To faithfully preserve 3D geometric shapes while aggressively compressing the latent space dimensions to $32 \times 32 \times 4$, we design the total loss function $\mathcal{L}_{total}$ as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{Pos} + \lambda_{TSDF}\mathcal{L}_{TSDF} + \lambda_{Normal}\mathcal{L}_{Normal} + \lambda_{KL}\mathcal{L}_{KL} \tag{2}$$

Here, $\mathcal{L}_{Pos}$ and $\mathcal{L}_{TSDF}$ are L1 losses that enforce pixel-level reconstruction accuracy. However, pixel-level losses alone are insufficient to adequately constrain surface curvature and orientation information, leaving the preservation of high-frequency geometric features—such as fine fabric wrinkles and sharp edges—unguaranteed. To address this, we introduce a **geometry-aware normal rendering loss ($\mathcal{L}_{Normal}$)**. The mesh differentiably reconstructed through the DMS module is rendered into a normal map image using the differentiable rasterizer nvdiffrast [18], and the L1 error against the normal map identically rendered from the ground truth mesh is directly minimized as follows:

$$\mathcal{L}_{Normal} = \|\mathcal{R}(\hat{M}) - \mathcal{R}(M^*)\|_1 \tag{3}$$

where $\mathcal{R}(\cdot)$ denotes the nvdiffrast rendering function, $\hat{M}$ is the reconstructed mesh, and M^* is the ground truth mesh. This encourages high-frequency geometric features to be compressed into and reconstructed from the latent space without distortion. The advantages of combining the TSDF-based representation with the geometry-aware normal rendering loss over occupancy-based representations, in terms of boundary sharpness and preservation of thin non-manifold structures, are quantitatively examined in the ablation study in Section 4.

3.4 Latent Diffusion on DiffGI Latent Space

To verify that the proposed representation naturally supports high-quality conditional 3D generation on the 32×32 latent space constructed by DiffGI-VAE, we train a transformer-based latent diffusion model.

Unlike natural photographs, geometry image tensors require capturing global topological connectivity rather than relying on spatial inductive biases of pixel positions. We therefore depart from the U-Net architecture and adopt the Diffusion Transformer (DiT) [29] as the core backbone, which allows unrestricted global attention among input tokens. We also employ a flow-matching [22] formulation-based scheduler, which has been shown to achieve high-quality generation with fewer inference steps. To support diverse downstream applications, we construct three conditional generation models:

1. **Label-Conditioned Generation:** Generates global shapes conditioned on input class labels (e.g., furniture, lamp). A DiT-B/2 architecture is used to comprehensively learn shape characteristics.
2. **Image-Conditioned Generation:** Infers unseen back surfaces from a single front-view image to generate a complete 3D mesh. Semantic embedding vectors extracted by the DINOv2-Large [27] vision foundation model are injected into DiT-L/2 via a cross-attention mechanism to perform complex shape prediction.
3. **Occupancy-Conditioned Generation:** Generates physically plausible 3D garment draping and wrinkles conditioned on 2D sewing patterns or silhouette information. Since the input condition already strongly constrains the spatial structure in this task, preserving local spatial features is relatively more important than global attention. Accordingly, instead of a global attention-based DiT, a lightweight UNet-Tiny architecture with skip connections is selectively adopted to ensure computational efficiency.

To mitigate positional bias from fixed UV layouts, we apply geometric data augmentation via random 90°-multiple rotations and re-packing at the UV chart level during training. These transformations alter only the 2D layout while preserving the geometric signals within each chart, effectively expanding the training distribution and improving generalization (visual examples and quantitative ablation in the supplementary material).

4 Experiments

4.1 Experimental Setup and Evaluation Metrics

Datasets. We evaluate on three benchmarks: **ABO** [10] (7,900 commercial 3D furniture scans with complex topologies and open structures), **GarmageSet** [19] (14,801 physics-simulation-ready 3D garments with strong non-manifold characteristics), and **WARDROBE** [13] (∼300 real clothing images used for qualitative evaluation of zero-shot image-to-3D generalization).

Evaluation Metrics. For per-sample geometric evaluation, we randomly sample 100,000 points from both the generated and ground truth mesh surfaces. We report **Chamfer Distance (CD)** for macroscopic shape accuracy, **Normal Consistency (NC)**, computed as one minus the average L1 distance between normal maps rendered from four fixed viewpoints (so that higher is better), and **F1-score** at a distance threshold of 0.01. To assess boundary quality, we use **Boundary Chamfer Distance (BCD)**, which computes CD only on open-boundary points, and **Hausdorff Distance (HD,** d_H**)** for worst-case local distortion. For distribution-level evaluation, we measure **P-FID** and **P-KID** using pretrained PointNet++ [30] features, along with **EMD** and **JSD** between generated and real point cloud distributions. Table 1 summarizes the reconstruction and encoding fidelity of all methods.

Table 1: Quantitative comparison of 3D reconstruction and encoding fidelity on the ABO and GarmageSet datasets. We report CD, EMD, JSD, and NC for Omages, GarmageNet, and our DiffGI-VAE. All baselines use the same Omages-style tessellation (Tess.); for GarmageNet we additionally report its native official extraction (Official). For GarmageNet, N denotes the number of garment panels, each represented as a fixed 72-dimensional vector.

Method	Representation Size	ABO Dataset				GarmageSet			
		CD ($\times 10^{-3}$) ↓	EMD ↓	JSD ($\times 10^{-3}$) ↓	NC ↑	CD ($\times 10^{-3}$) ↓	EMD ↓	JSD ($\times 10^{-3}$) ↓	NC ↑
omages64	64×64×4	0.89	0.25	0.92	**0.89**	1.31	0.17	1.79	0.95
GarmageNet (Tess.)	$N \times 72$	-	-	-	-	2.19	0.21	32.61	0.88
GarmageNet (Official)	$N \times 72$	-	-	-	-	1.89	0.17	5.51	0.90
ours	32×32×4	**0.83**	**0.23**	**0.89**	0.83	**0.46**	**0.16**	**1.24**	**0.96**

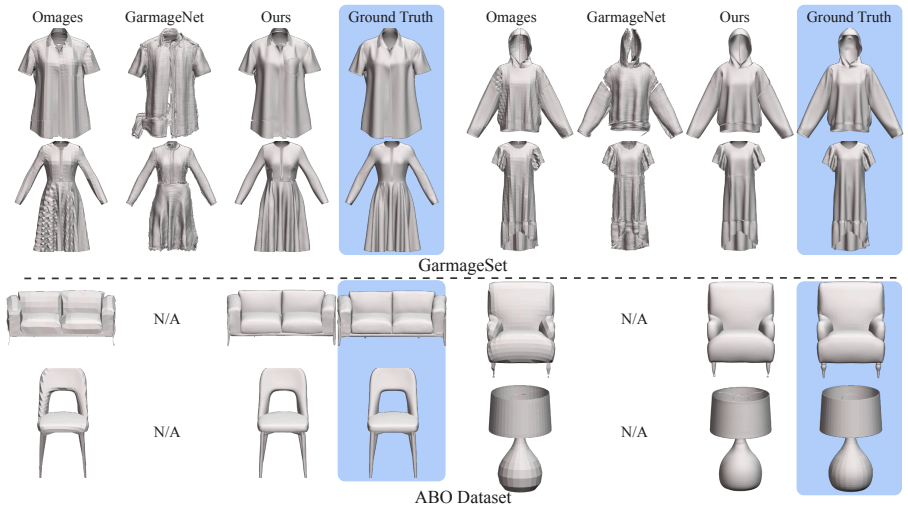

Fig. 3: Qualitative comparison of VAE reconstructions on the ABO and GarmageSet datasets, illustrating that our DiffGI-VAE yields sharper boundaries and better preservation of thin-shell details than Omages and GarmageNet.

4.2 Reconstruction Fidelity (DiffGI-VAE)

We compare DiffGI-VAE against Omages and GarmageNet on the ABO and GarmageSet datasets. Omages directly converts meshes to geometry images without VAE compression; GarmageNet uses per-panel geometry images with a fixed panel count, making it inapplicable to ABO (entries excluded). To evaluate each representation with its intended pipeline, each method uses its native mesh extractor: Differentiable Marching Squares for DiffGI and the Omages-style tessellation for Omages. Since GarmageNet's native extraction is a slow multi-stage non-learning procedure (∼5.1 s per mesh), we report it under the same Omages-style tessellation (Table 1, Tess.) for the main comparison and additionally under its native official extraction (Official). DiffGI remains superior under both GarmageNet settings.

As shown in Table 1 and Figure 3, DiffGI-VAE consistently outperforms baselines in CD, EMD, and JSD on both datasets. Notably, our method compresses a

256×256 geometry image into a 32×32×4 VAE latent, whereas Omages directly operates on a 64×64×4 geometry image without VAE compression; despite this aggressive compression, DiffGI-VAE achieves superior reconstruction fidelity. On ABO, NC is slightly lower than Omages, which we attribute to the information loss inherent in 8× spatial compression, particularly on flat furniture surfaces where Omages' uncompressed representation retains normal information more directly. On GarmageSet, however, our method achieves the highest NC, confirming state-of-the-art reconstruction quality for thin non-manifold garment structures. This demonstrates that the continuous TSDF representation with fully differentiable optimization effectively preserves curvature and fine details even under substantial latent compression.

Table 2: Ablation study of representation (Occ. vs. TSDF) and normal rendering loss (NL) for our DiffGI-VAE on GarmageSet. TSDF-based variants outperform occupancy-based ones across all metrics, and enabling NL yields the highest NC with the lowest CD, EMD, and JSD.

Rep.	$\mathcal{L}_{\text{Normal}}$	CD ($\times 10^{-3}$) ↓	EMD ↓	JSD ($\times 10^{-3}$) ↓	NC ↑
Occ	×	1.503	0.171	4.539	0.906
Occ	✓	1.313	0.166	4.257	0.947
TSDF	×	0.595	0.165	2.169	0.921
TSDF	✓	**0.461**	**0.160**	**1.244**	**0.961**

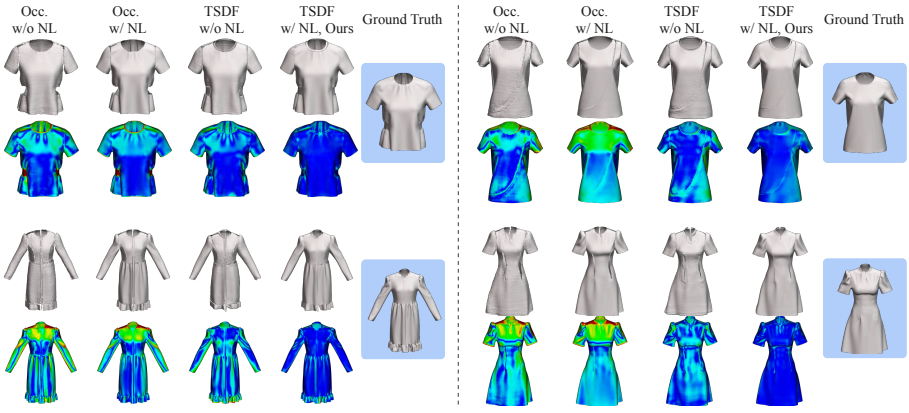

Fig. 4: Qualitative ablation study of our VAE on the GarmageSet dataset. The visual comparisons demonstrate how different representations (Occ. vs. TSDF) and the normal rendering loss affect the 3D reconstruction quality.

4.3 Ablation Study: Differentiable Mesh Reconstruction and Normal Loss

To verify that the performance gains of the DiffGI framework result from the cumulative effect of each design component, we conduct an ablation study that systematically decomposes the contributions of the representation (Occ. vs. TSDF) and the geometry-aware normal rendering loss ($\mathcal{L}_{Normal}$). Since occupancy maps do not admit differentiable mesh extraction by default, we reconstruct the occupancy variants with a differentiable formulation of the Omages-style tessellation, so that all four variants receive the same normal rendering supervision and the comparison isolates the representation rather than the extraction step.

The results in Table 2 demonstrate that each of the two design components independently yields significant performance gains.

Effect of representation (Occ. vs. TSDF). When only the representation is switched from Occ. to TSDF without $\mathcal{L}_{Normal}$, CD decreases by more than half from 1.503×10^{-3} to 0.595×10^{-3}, and JSD also improves substantially from 4.539×10^{-3} to 2.169×10^{-3}. This quantitatively confirms that the continuous TSDF encodes geometric information in boundary regions far more effectively than binary occupancy maps.

Effect of geometry-aware normal rendering loss ($\mathcal{L}_{Normal}$). Enabling $\mathcal{L}_{Normal}$ on the same representation consistently improves all metrics. In particular, under the TSDF-based setting, NC increases from 0.921 to 0.961, with CD and JSD further improving as well. Notably, the NC of the Occ. + $\mathcal{L}_{Normal}$ combination (0.947) exceeds the NC of TSDF without $\mathcal{L}_{Normal}$ (0.921), suggesting that $\mathcal{L}_{Normal}$ enforces normal orientation information strongly enough to partially compensate for the limitations of the representation. However, in terms of CD, EMD, and JSD, TSDF without $\mathcal{L}_{Normal}$ still outperforms Occ. + $\mathcal{L}_{Normal}$, indicating that the contribution of the representation itself is more fundamental for overall shape reconstruction.

As shown in Figure 4, the Occ. without $\mathcal{L}_{Normal}$ setting—trained with pixel-level losses only—exhibits prominent high-frequency noise and aliasing artifacts at mesh boundaries and surfaces. In the final configuration combining the TSDF representation with $\mathcal{L}_{Normal}$, these artifacts are substantially mitigated and boundaries are reconstructed with noticeably greater sharpness.

Effect of VAE initialization. We additionally train the same DiffGI-VAE architecture from random initialization and observe nearly identical final reconstruction fidelity to the SD1.5-initialized model on GarmageSet (CD 0.47 vs. 0.46, NC 0.96 for both), confirming that the reconstruction gains stem from the DiffGI representation and differentiable reconstruction pipeline rather than the SD1.5 VAE initialization; detailed results and convergence analysis are provided in Sec. 2 of the supplementary material.

4.4 Computational Efficiency and Inference Performance

Table 3 compares inference cost across methods using each method's official recommended settings. Volumetric models such as TRELLIS require up to 16.28 GB

Table 3: Computational Efficiency & Inference Performance. We compare peak memory usage and inference latency across different model scales, conditioning tasks, and hardware setups. Our DiffGI framework demonstrates exceptional efficiency, capable of running even on a consumer-grade CPU.

Method	Hardware	Peak VRAM (GB) ↓	Time (sec) ↓
TRELLIS-image	RTX A6000 Ada	16.28	4.52
TRELLIS.2 512^2	RTX A6000 Ada	2.65	12.35
TRELLIS.2 1024^2	RTX A6000 Ada	6.41	45.14
GarmageNet	RTX A6000 Ada	0.96	0.40
Omages	RTX A6000 Ada	2.49	52.0
Ours-Label	RTX A6000 Ada	1.18	0.50
Ours-Image	RTX A6000 Ada	3.22	0.80
Ours-Image	RTX 4070 (12GB)	3.22	1.21
Ours-Image	MacBook M4 (CPU)	–	8.52

Table 4: Label-conditioned 3D generation performance on the ABO dataset, reported as P-FID and P-KID per category and averaged (Mean).

	Chair		Lamp		Sofa		Table		Mean	
	Omages	Ours	Omages	Ours	Omages	Ours	Omages	Ours	Omages	Ours
P-FID	22.30	**11.15**	52.48	**26.55**	15.81	**10.49**	33.99	**31.45**	31.14	19.91
P-KID	0.08	**0.03**	0.26	**0.07**	0.06	**0.04**	**0.10**	0.11	0.12	0.06

VRAM, and Omages takes 52 seconds per sample. In contrast, DiffGI operates in a compact 2D latent space, completing image-conditioned generation in ∼1.2 s on a consumer GPU (RTX 4070, 3.22 GB VRAM) and ∼8.5 s on CPU only (MacBook M4), demonstrating scalability from servers to edge devices.

4.5 Downstream Generative Tasks: Open-surface 3D Generation

Label-Conditioned Generation We generate 512 3D samples per class on ABO and measure P-FID and P-KID. As shown in Table 4 and Figure 5, the proposed model consistently improves P-FID across all categories, with the largest reductions in Chair (50%) and Lamp (49%). For Lamp, the absolute P-FID remains elevated due to limited training samples, but P-KID (0.07 vs. Omages 0.26) confirms robust generation quality. Qualitatively, Omages produces broken thin frames due to binary masks, while DiffGI generates smooth, continuous structures via subpixel boundary estimation.

Image-Conditioned Generation: Comparison with Foundation Models For the *single-view image-to-3D* task, we compare DiffGI against foundation-based models (TRELLIS [43], TRELLIS.2 [42]) and GarmageNet [19] on GarmageSet. Our goal is not to claim universal superiority, but to examine the practical advantages of a surface-centric representation for thin-shell and open-boundary garment geometry.

As shown in Table 5, DiffGI achieves the best CD (1.35×10^{-2}), F1 (0.48), d_H (8.42×10^{-2}), and BCD (2.91×10^{-2}) with only 23K vertices on average. The low BCD in particular confirms that the TSDF-based representation and differentiable reconstruction pipeline stably preserve thin fabric boundaries. Qualitatively (Figure 6), TRELLIS tends to produce excessive thickness on thin-shell

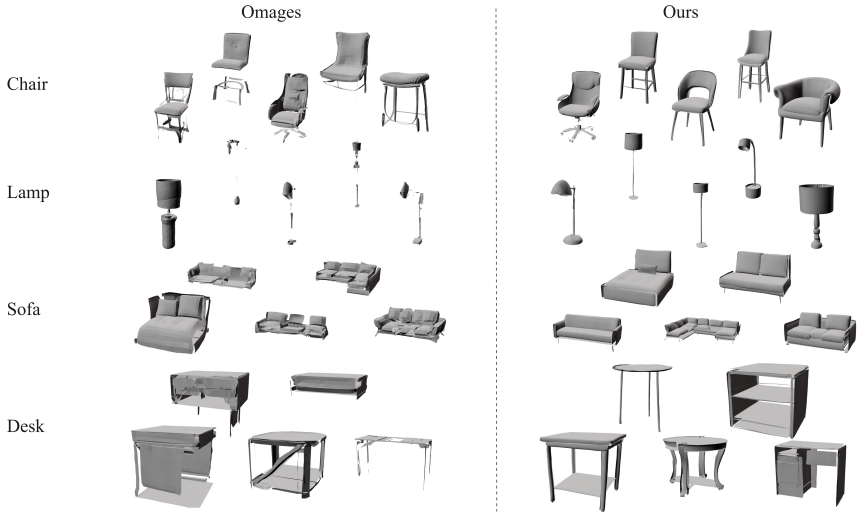

Fig. 5: Qualitative label-conditioned generation results on the ABO dataset. We compare Omages and our DiffGI-based diffusion model, showing improved reconstruction of thin frames and open-boundary structures, with fewer staircase artifacts.

Table 5: Quantitative comparison of image-to-3D generation on GarmageSet. TRELLIS and TRELLIS.2 are general-purpose foundation models trained on large-scale multi-category 3D data, whereas GarmageNet and our DiffGI are trained specifically on GarmageSet. We report this comparison to highlight the practical advantages of a surface-centric representation for thin-shell garment geometry, rather than to claim universal superiority.

Method	**#Vert.** ↓	**CD** ($\times 10^{-2}$) ↓	**F1** ↑	**HD** ($\times 10^{-2}$) ↓	**BCD** ($\times 10^{-2}$) ↓
TRELLIS	109K	3.44 ± 6.97	0.28 ± 0.14	15.38 ± 27.57	N/A
TRELLIS.2 512^2	380K	11.01 ± 7.64	0.27 ± 0.14	69.70 ± 50.00	12.44 ± 8.56
GarmageNet	526K	4.31 ± 3.03	0.20 ± 0.12	23.76 ± 11.12	5.64 ± 2.94
Ours (DiffGI)	**23K**	$\mathbf{1.35 \pm 0.47}$	$\mathbf{0.48 \pm 0.12}$	$\mathbf{8.42 \pm 3.25}$	$\mathbf{2.91 \pm 0.83}$

garments due to its watertight assumption, while GarmageNet exhibits residual boundary aliasing from binary occupancy maps. DiffGI reconstructs cleaner boundaries with more compact meshes.

Although our image-conditioned model is trained on front-view rendered normal maps, at inference it also generalizes to real RGB front-view photographs, as shown by the zero-shot results on the WARDROBE dataset in the supplementary material, demonstrating the extensibility of our representation to diverse conditioning inputs.

Occupancy-Conditioned Generation We generate 3D garments conditioned solely on 2D sewing pattern silhouettes. Despite using a lightweight UNet-Tiny decoder, visually natural draping and wrinkles are stably produced (Figure 7). Locally modifying only the sleeve region in the 2D occupancy map yields consis-

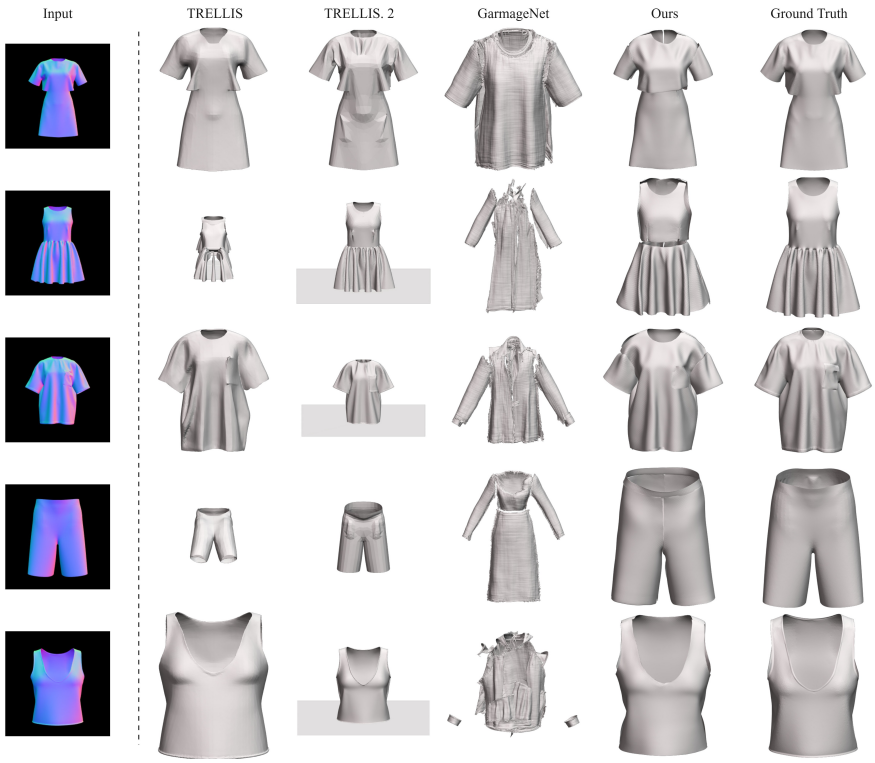

Fig. 6: Qualitative single-view image-to-3D comparison on GarmageSet. From left to right, each row shows the input front-view rendered normal map, the results of TRELLIS, TRELLIS.2, and GarmageNet, our DiffGI generation, and the ground-truth mesh. DiffGI preserves thin-shell details and produces cleaner open boundaries with far fewer vertices than the watertight foundation models (TRELLIS, TRELLIS.2) and the occupancy-based GarmageNet.

tent updates to the corresponding 3D geometry while preserving the remaining patterns, demonstrating that 2D pattern-level edits directly propagate to 3D mesh variations.

5 Conclusion, Limitations, and Future Work

We proposed Differentiable Geometry Image (DiffGI), a fully differentiable framework that replaces binary occupancy maps with a continuous 2D TSDF, enabling subpixel-precise boundary reconstruction on geometry images. Our Differentiable Marching Squares (DMS) module brings the mesh reconstruction step inside the learning graph, allowing 3D surface losses to backpropagate seamlessly to the 2D latent space. Built on this, DiffGI-VAE compresses complex non-manifold surfaces into a $32\times32\times4$ latent space with a geometry-aware nor-

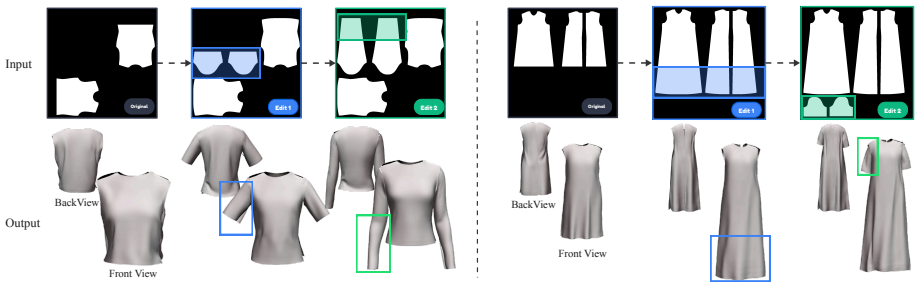

Fig. 7: Occupancy-conditioned garment generation and local edit propagation. Given a 2D occupancy map as input, modifying only the sleeve region in 2D leads to consistent updates of the corresponding 3D geometry, demonstrating that local pattern edits are faithfully reflected in the generated garment mesh.

mal rendering loss, consistently outperforming prior geometry-image methods on both garment and furniture benchmarks. A transformer-based latent diffusion model trained on this space achieves competitive or superior generation quality with significantly less compute, running in real time even on consumer-grade hardware.

Limitations. TSDF-based linear interpolation can introduce localized rounding on extreme sharp edges (e.g., mechanical parts), and the current framework does not generate RGB textures or PBR attributes alongside geometry. In addition, since each UV chart is reconstructed independently and the DMS module treats adjacent patches as separate regions, the recovered geometry can show visible seams where neighboring charts meet, such as between the upper and lower parts of the dress in Figure 6. These boundary discontinuities do not affect our reconstruction and generation metrics, but they pose additional challenges for downstream physical simulation. Enforcing cross-chart boundary consistency is left for future work.

Future work. We plan to (1) integrate feature-preserving iso-surfacing (e.g., dual contouring) into DMS for sharp-edge domains, (2) extend the latent space to jointly generate high-resolution textures and material maps, and (3) introduce a two-stage pipeline separating 2D pattern layout generation from 3D shape reconstruction for stronger controllability over complex garment designs.

Acknowledgements

We thank Hyun Kang, Seungoh Han, Sihun Cha, Dong-sig Kang, and Gyoo-Chul Kang for valuable discussions and feedback throughout this work. We are also grateful to CLO Virtual Fashion for providing the research environment and resources that made this work possible.

References

1. Binninger, A., Wiersma, R., Herholz, P., Sorkine-Hornung, O.: TetWeave: Isosurface extraction using on-the-fly delaunay tetrahedral grids for gradient-based mesh optimization. ACM Transactions on Graphics (2025)
2. Boss, M., Huang, Z., Vasishta, A., Jampani, V.: SF3D: Stable fast 3D mesh reconstruction with UV-unwrapping and illumination disentanglement. In: CVPR (2025)
3. Carr, N.A., Hoberock, J., Crane, K., Hart, J.C.: Rectangular multi-chart geometry images. In: Symposium on Geometry Processing (2006)
4. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion NeRF: A unified approach to 3D generation and reconstruction. In: ICCV (2023)
5. Chen, W., Lin, C., Li, W., Yang, B.: 3PSDF: Three-pole signed distance function for learning surfaces with arbitrary topologies. In: CVPR (2022)
6. Chen, Z., Tagliasacchi, A., Funkhouser, T., Zhang, H.: Neural dual contouring. ACM Transactions on Graphics (2022)
7. Chen, Z., Zhang, H.: Neural marching cubes. ACM Transactions on Graphics (2021)
8. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: SDFusion: Multimodal 3D shape completion, reconstruction, and generation. In: CVPR (2023)
9. Chou, G., Bahat, Y., Heide, F.: Diffusion-SDF: Conditional generative modeling of signed distance functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2262–2272 (2023)
10. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: ABO: Dataset and benchmarks for real-world 3D object understanding. In: CVPR (2022)
11. Elizarov, S., Rowles, C., Donné, S.: Geometry image diffusion: Fast and data-efficient text-to-3D with image-based surface representation. In: International Conference on Learning Representations (ICLR) (2025)
12. Fainstein, M., Siless, V., Iarussi, E.: DUDF: Differentiable unsigned distance fields with hyperbolic scaling. In: CVPR. pp. 4484–4493 (2024)
13. freshersstaff: Wardrobe vision dataset. `https://www.kaggle.com/datasets/freshersstaff/wardrobe-vision-dataset` (2025), accessed: 2025-12-01
14. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. In: ICLR (2022)
15. Gu, X., Gortler, S.J., Hoppe, H.: Geometry images. ACM Transactions on Graphics (2002)
16. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: LRM: Large reconstruction model for single image to 3D. In: ICLR (2024)
17. Hou, F., Chen, X., Wang, W., Qin, H., He, Y.: Robust zero level-set extraction from unsigned distance fields based on double covering. ACM Transactions on Graphics (2023)
18. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. arXiv:2011.03277 (2020)
19. Li, S., Liu, R., Liu, C., Wang, Z., He, G., Li, Y.L., Jin, X., Wang, H.: GarmageNet: A multimodal generative framework for sewing pattern design and generic garment modeling. ACM Transactions on Graphics (TOG) (2025)
20. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., Cao, Y.P.: TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models. CoRR (2025)

21. Liao, Y., Donné, S., Geiger, A.: Deep marching cubes: Learning explicit surface representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
23. Liu, Z., Feng, Y., Xiu, Y., Liu, W., Paull, L., Black, M.J., Schölkopf, B.: Ghost on the shell: An expressive representation of general 3D shapes. In: ICLR (2024)
24. Long, X., Lin, C., Liu, L., Liu, Y., Wang, P., Theobalt, C., Komura, T., Wang, W.: NeuralUDF: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In: CVPR. pp. 20834–20843 (2023)
25. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics (1987)
26. Meng, X., Chen, W., Yang, B.: NeAT: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In: CVPR. pp. 248–258 (2023)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
28. Park, J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: CVPR (2019)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
30. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems (2017)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
32. Sander, P.V., Wood, Z.J., Gortler, S.J., Snyder, J., Hoppe, H.: Multi-chart geometry images. In: Symposium on Geometry Processing (2003)
33. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. In: NeurIPS (2021)
34. Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N., Gao, J.: Flexible isosurface extraction for gradient-based mesh optimization. ACM Transactions on Graphics (2023)
35. Sinha, A., Bai, J., Ramani, K.: Deep learning 3D shape surfaces using geometry images. In: ECCV (2016)
36. Son, S., Gadelha, M., Zhou, Y., Fisher, M., Xu, Z., Qiao, Y.L., Lin, M.C., Zhou, Y.: DMesh++: An efficient differentiable mesh for complex shapes. In: ICCV (2025)
37. Son, S., Gadelha, M., Zhou, Y., Xu, Z., Lin, M.C., Zhou, Y.: DMesh: A differentiable mesh representation. In: Advances in Neural Information Processing Systems (2024)
38. Stippel, C., Mujkanovic, F., Leimkühler, T., Hermosilla, P.: Marching neurons: Accurate surface extraction for neural implicit shapes. ACM Transactions on Graphics (2025)
39. Wang, L., Chen, W., Meng, X., Yang, B., Li, J., Gao, L., et al.: HSDF: Hybrid sign and distance field for modeling surfaces with arbitrary topologies. In: NeurIPS (2022)
40. Wang, R., Li, J., Zeng, D., Ma, X., Xu, Z., Zhang, J., Zhao, Q.: GenUDC: High quality 3D mesh generation with unsigned dual contouring representation. In: Proceedings of the 32nd ACM International Conference on Multimedia (2024)

41. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: CRM: Single image to 3D textured mesh with convolutional reconstruction model. In: ECCV (2024)
42. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3D generation. arXiv preprint arXiv:2512.14692 (2025)
43. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3D latents for scalable and versatile 3D generation. In: CVPR (2025)
44. Xie, R., Huang, K., Luo, X., Chen, Y., Wang, L., Wang, Q., Ye, Q., Chen, W., Zheng, W., Huo, Y.: LDM: Large tensorial SDF model for textured mesh generation. Graphical Models **140**, 101271 (2025)
45. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. CoRR (2024)
46. Yan, X., Lee, H.H., Wan, Z., Chang, A.X.: An object is worth 64x64 pixels: Generating 3D object via image diffusion. In: International Conference on 3D Vision (3DV) (2025)
47. Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Wu, Z., Jiang, Y.G., Mei, T.: DreamMesh: Jointly manipulating and texturing triangle meshes for text-to-3D generation. In: ECCV (2024)
48. Yang, X., Shi, H., Zhang, B., Yang, F., Wang, J., Zhao, H., Liu, X., Wang, X., Lin, Q., Yu, J., et al.: Hunyuan3D 1.0: A unified framework for text-to-3D and image-to-3D generation. CoRR (2024)
49. Yu, X., Yuan, Z., Guo, Y.C., Liu, Y.T., Liu, J., Li, Y., Cao, Y.P., Liang, D., Qi, X.: TEXGen: a generative diffusion model for mesh textures. ACM Transactions on Graphics (2024)
50. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. CoRR (2025)
51. Zheng, X., Pan, H., Wang, P., Tong, X., Liu, Y., Shum, H.: Locally attentional SDF diffusion for controllable 3D shape generation. ACM Transactions on Graphics (SIGGRAPH) (2023)
52. Zhou, J., Zhang, W., Ma, B., Shi, K., Liu, Y.S., Han, Z.: UDiFF: Generating conditional unsigned distance fields with optimal wavelet diffusion. In: CVPR. pp. 21496–21506 (2024)