TOPOS: HIGH-FIDELITY AND EFFICIENT INDUSTRY-GRADE 3D HEAD GENERATION

Anonymous authors
Paper under double-blind review

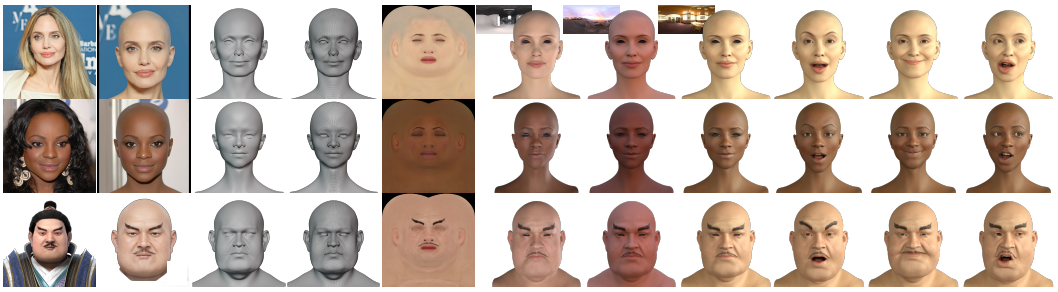


Figure 1: Our proposed TOPOS framework is capable of generating high-fidelity 3D head mesh and texture map given a single image. From left to right are the input images, edited images, generated geometry, mesh topology, generated texture maps, rendering images under different lighting conditions and animation results across different expressions, respectively. The upper-left insets illustrate the environment maps. Please zoom in for better inspection.

ABSTRACT

Industrial 3D head generation requires fixed reference topology for production-level rigging, skinning and animation, yet existing 3D generative models typically produce inconsistent mesh topologies which are incompatible with production pipelines. In this paper, we present TOPOS, a single image conditioned framework that jointly generates high-fidelity geometry and appearance under a unified industry-standard topology. To model heads under this unified topology, we proposed a novel variational autoencoder structure, termed TOPOS-VAE. Inspired by multi-modal large language models (MLLMs), TOPOS-VAE leverages the Perceiver Resampler to convert input pointclouds sampled from head meshes of diverse topologies into the target reference topology. Building upon TOPOS-VAE's structured latent space, we train a rectified flow transformer, TOPOS-DiT, to efficiently generate high-fidelity head meshes from a single image. We further present TOPOS-Texture, an end-to-end module that produces relightable and geometry-aligned UV texture maps from the same portrait image via fine-tuning a multi-modal image generative model. Extensive experiments demonstrate that TOPOS achieves state-of-the-art performance on 3D head generation, surpassing both classical face reconstruction methods and general 3D object generative models, highlighting its effectiveness for digital human creation. We will release our code and trained models to facilitate future research.

1 INTRODUCTION

High-fidelity 3D head generation is of vital importance in the film, computer animation and video game industries. However, it is still quite time-consuming for a skilled artist to create a realistic and industry-grade head asset, which is a labor-intensive process taking several hours or days per head. Moreover, industrial production pipelines typically require a fixed reference topology across all assets to enable rigging, skinning and downstream character animation. Therefore, developing an automatic algorithm for high-fidelity and efficient 3D head generation with a uniform topology and relightable texture map would be highly meaningful for digital human creation.

Traditional 3D face parametric models like 3DMMs Blanz & Vetter (1999); Paysan et al. (2009); Li et al. (2017) offer compact and controllable representations but lack the capacity to capture high-frequency geometric and textural details. On the other hand, recent deep generative approaches using point clouds Luo & Hu (2021); Vahdat et al. (2022), SDFs Zheng et al. (2023); Xiong et al. (2025b), or Flexicubes Xiang et al. (2025b); He et al. (2025b) to enable fine-grained modeling but rely on additional surface extraction steps such as Marching Cubes Lorensen & Cline (1998), yielding head

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

meshes with excessive vertices and unstructured connectivity. Auto-regressive mesh generation methods Chen et al. (2024a; 2025); Siddiqui et al. (2024); Lin et al. (2026); Xu et al. (2026) avoid this extraction but are constrained by the prohibitive cost of Transformer self-attention Vaswani et al. (2017). More recently, flow-based mesh generation methods He et al. (2025a); Li et al. (2026); Sun et al. (2026); Wang et al. (2026a); Zhao et al. (2026); Long et al. (2026); Wang et al. (2026b) directly generate artist meshes from images or text with parallel diffusion or flow-matching. However, none of the above methods enforce a consistent topology across generated instances, restricting their applicability to downstream tasks.

For facial appearance generation, inherent self-occlusions make hole-free texture reconstruction challenging Lattas et al. (2021); Han et al. (2024; 2025). Previous UV inpainting Zeng et al. (2022); Lei et al. (2023); Yang et al. (2025); Qiu et al. (2025b) or template completion Bai et al. (2023) methods introduce seams and identity drift, while inaccuracies in the unprojection process further degrade texture quality. Recent end-to-end approaches such as Uni-1 Luma AI (2026) lack explicit semantic alignment, leading to inconsistent texture representations.

In this paper, we propose TOPOS, a generative framework tailored for industry-grade 3D head generation, comprising TOPOS-VAE, TOPOS-DiT, and TOPOS-Texture. To address the scarcity of standardized topology data, TOPOS-VAE learns a continuous latent space and converts point clouds from diverse-topology meshes into a fixed reference topology. It combines Perceiver Resampler Alayrac et al. (2022) as the point cloud encoder with a GNN decoder, leveraging cross-modal alignment to translate unstructured pointclouds features into structured fixed-topology meshes. Building on this latent space, TOPOS-DiT is a rectified flow transformer conditioned on rendered portraits that directly generates meshes in about one second. TOPOS-Texture is a dedicated end-to-end pipeline fine-tuned from a pretrained multi-modal image generative model, Qwen-Image-Edit Wu et al. (2025), that produces relightable and geometry-consistent UV maps from a single portrait image. TOPOS-Texture is executed in parallel with TOPOS-DiT and is bottlenecked only by its backbone (around one minute in our implementation). As a result, the contributions of our paper are threefold:

- Inspired by multi-modal large language models, we propose TOPOS-VAE, which combines the Perceiver Resampler with Graph Neural Network to faithfully reconstruct high-quality 3D face meshes from input unstructured pointclouds, followed by TOPOS-DiT that directly generates meshes in this unified latent space.
- We introduce TOPOS-Texture, a feed-forward texture generation pipeline fine-tuned from a pretrained multimodal image generative model, which produces relightable and geometry-consistent UV maps without unprojection and inpainting artifacts.
- We present TOPOS, to our knowledge, the first end-to-end diffusion framework that directly outputs high-fidelity 3D head meshes that simultaneously preserve facial identity and conform to an industry-standard topology. Qualitative and quantitative experiments have been conducted to verify the superiority of our method over other existing face reconstruction and 3D generation methods.

2 Related Work

2.1 3D Face Reconstruction

Parametric models like 3DMMs Blanz & Vetter (1999); Paysan et al. (2009); Li et al. (2017) and their extensions Booth et al. (2018); Chai et al. (2022); Xu et al. (2020); Yang et al. (2020) provide fixed topology but lack high-frequency expressiveness due to the low-dimensional linear subspaces. Rendering-based methods Zielonka et al. (2022); Retsinas et al. (2024); Zeng et al. (2023); Giebenhain et al. (2025) improve identity consistency yet remain constrained by predefined parameter spaces. Non-parametric approaches using vertex regression Richardson et al. (2017); Zeng et al. (2019), NeRF Mildenhall et al. (2020); Wang et al. (2021), Tri-plane Chan et al. (2022) or 3DGS Kerbl et al. (2023); Hu et al. (2024); Chu & Harada (2024); Qiu et al. (2025a) achieve superior rendering realism but output volumetric or point-based representations, which is incompatible with production pipelines. Therefore, existing 3D face reconstruction methods struggle to simultaneously achieve high-fidelity geometry as well as an industry-standard and uniform topology.

2.2 3D Shape and Mesh Generation

Deep generative models Ho et al. (2020); Lipman et al. (2022) have advanced 3D shape generation across implicit fields Mescheder et al. (2019); Park et al. (2019); Zheng et al. (2023), point

clouds Nichol et al. (2022); Luo & Hu (2021), and sparse voxels Xiong et al. (2025b); He et al. (2025b), with both unstructured Jaegle et al. (2021); Zhang et al. (2023a); Zhao et al. (2025b) and structured Xiang et al. (2025b;a) latent formulations. However, extracted 3D meshes typically exhibit excessive vertices and irregular connectivity. Direct mesh generation methods, including autoregressive approaches Siddiqui et al. (2024); Chen et al. (2024a); Weng et al. (2024a); Tang et al. (2024); Chen et al. (2025); Lionar et al. (2025); Hao et al. (2024); Wang et al. (2025) and recent flow-based artist mesh generation methods He et al. (2025a); Li et al. (2026); Sun et al. (2026); Wang et al. (2026a); Zhao et al. (2026); Long et al. (2026); Wang et al. (2026b), produce coherent polygonal meshes but fail to enforce consistent topology across generated instances, which is incompatible with the uniform, reusable topology required by industrial production pipelines.

2.3 3D Face Texture Generation

Early face reconstruction methods represent appearance using 3DMM texture coefficients or per-vertex colors, which cannot capture high-frequency facial details. Optimization-based UV recovery methods Lattas et al. (2021); Han et al. (2024; 2025) produces relightable assets but requires multi-view or controlled capture. Single-view texture reconstruction methods address this problem via UV inpainting Bai et al. (2023); Zeng et al. (2022); Lei et al. (2023); Yang et al. (2025); Qiu et al. (2025b); Dai et al. (2025) or direct UV regression Wang et al. (2019a); Dib et al. (2024), yet struggle with self-occluded regions (e.g., inner mouth) and introduce seams or misalignment. Recent end-to-end texture synthesis methods like Uni-1 Luma AI (2026) leverage large generative models but lack explicit semantic alignment with mesh layout, yielding inconsistent textures. In contrast, our method explicitly enforces semantic alignment, producing UV textures that are more consistent and suitable for downstream avatar applications.

3 Method

In this section, we provide a detailed explanation of our industry-grade 3D head generation framework, TOPOS. Figure 2 shows an overview of our TOPOS framework, with details described below.

3.1 TOPOS-VAE

3.1.1 Encoder

A widely adopted design for 3D shape VAEs Zhao et al. (2025b; 2023) is to employ the encoder from 3DShape2VecSet Zhang et al. (2023a), which stacks cross-attention layers to encode input point clouds $\mathcal{P}$ using either learnable queries or farthest-point-sampled subsets as cross-attention queries. However, we argue that neither choice is suitable for our task. The latent space of TOPOS-VAE should form a set of **semantic anchors**, where each latent token corresponds to a specific semantic region on the fixed-topology head mesh. Learnable queries in 3DShape2VecSet struggle to establish such anchors due to the lack of iterative refinement with the input, leading to unstable training. On the other hand, farthest-point-sampled (FPS) queries, cannot guarantee consistent semantic correspondence across instances. As a result, 3DShape2VecSet-style encoders perform poorly on head mesh reconstruction under our fixed-topology setting in our pilot experiments.

Instead, inspired by the multi-modal large language models (MLLMs), we adopt the Perceiver Resampler Alayrac et al. (2022) as our encoder $\mathcal{E}$. The Perceiver Resampler maintains a set of learned queries $\mathcal{L} \in \mathbb{R}^{N\times d}$ shared across all instances. Unlike 3DShape2VecSet, where queries interact with the input only once via a single cross-attention layer, Perceiver Resampler iteratively refines its queries by simultaneously attending to both the input pointcloud and one another across multiple layers. This iterative refinement enables each query to specialize in a consistent semantic region across all training instances, naturally forming the desired semantic anchors. Since the queries are input-agnostic at initialization, their semantic roles emerge entirely through self-supervised learning, ensuring cross-instance consistency. The encoding process is formulated as $\mathcal{L} = \mathcal{E}(\mathcal{P})$.

3.1.2 Decoder

We employ a Graph Neural Network (GNN) as our decoder $\mathcal{D}$, which takes latent tokens $\mathcal{L}$ from the Perceiver Resampler and outputs head meshes with consistent topology. Given the shared face connectivity $\mathcal{F}$ across all training meshes, the mesh structure naturally defines a graph $\mathcal{G}$.

Graph Convolution For vertex v_i with feature F_i and neighbors $\mathcal{N}(i)$, our graph convolution is defined as:

$$F_i' = W_0 F_i + \textstyle\sum_{j\in\mathcal{N}(i)} W_1 F_j, \tag{1}$$

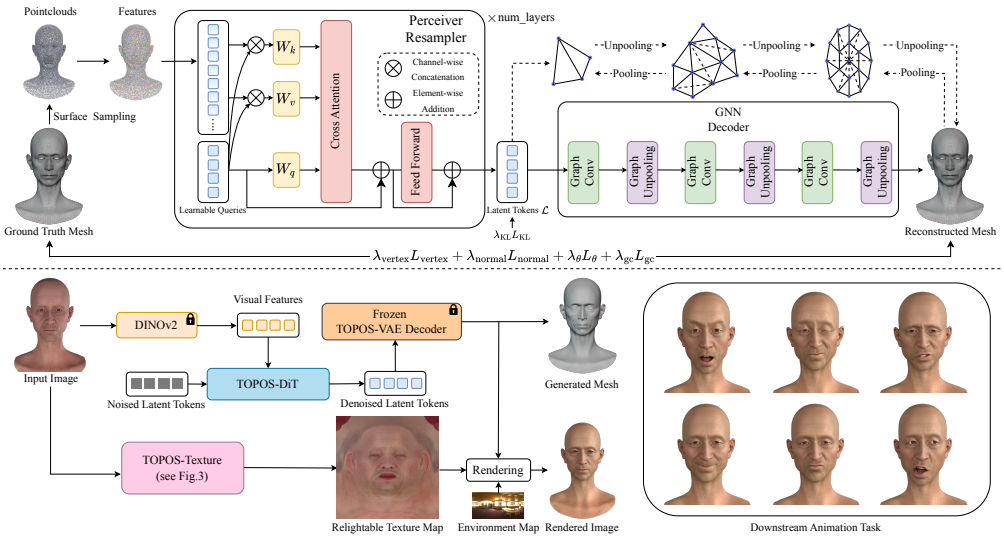


Figure 2: The top part of this figure shows the structure of our TOPOS-VAE, which utilizes Perceiver Resampler to encode the input pointclouds and GNN decoder to decode it into 3D head mesh with consistent topology. The bottom part shows the training and generation process of our TOPOS-DiT and TOPOS-Texture. Thanks to the unified topology, the generated 3D head mesh can be readily driven across different facial expressions for downstream applications.

where W_0 and W_1 are trainable weights. This formulation is similar to GraphSAGE Hamilton et al. (2017) and is simpler yet more effective than previous graph convolutions in 3D deep learning Wang et al. (2019b) that define update functions as MLPs.

Graph Pooling & Unpooling In our TOPOS-VAE, since GNNs are only employed in the decoder, the graph pooling operation is simply introduced to establish the mapping relation between vertices before and after pooling. Then we keep track of this mapping during graph unpooling in the decoder.

We adopt grid-based graph pooling Simonovsky & Komodakis (2017); Pang et al. (2023) for efficiency. Grid-based pooling employs regular voxel grids to cluster vertices within the same voxel into a single vertex, constructing a hierarchy of graph structures at multiple levels:

$$\mathcal{G}^{(0)} \sim \{\mathcal{V}_t, \mathcal{F}\}, \quad \mathcal{G}^{(l+1)} = \text{GridPool}^{(l)}\big(\mathcal{G}^{(l)}\big), \quad l = 0, \ldots, L-1, \tag{2}$$

where the initial graph structure $\mathcal{G}^{(0)}$ is fully determined by the template mesh $\mathcal{M}_t = \{\mathcal{V}_t, \mathcal{F}\}$ under our industry-standard topology setting, and we perform $L = 4$ successive downsampling operations. Notably, we set the number of learned queries of Perceiver Resampler N to match the number of vertices in the coarsest, i.e., the bottom-level graph $\mathcal{G}^L$, such that each query serves as a semantic anchor aligned with a vertex in $\mathcal{G}^L$. Since all meshes share a consistent topology, the pooling mapping is constructed once on the template mesh $\mathcal{M}_t$ and reused across all instances. Graph unpooling reverses this cached mapping in the decoder, and a final MLP head translates vertex features on $\mathcal{G}^{(0)}$ to 3D coordinates:

$$\hat{\mathcal{V}} = \mathcal{D}(\mathcal{L}), \quad \hat{\mathcal{M}} = (\hat{\mathcal{V}}, \mathcal{F}). \tag{3}$$

where $\hat{\mathcal{V}}$ and $\hat{\mathcal{M}}$ are output vertices and head meshes, respectively.

3.1.3 TOPOS-VAE Training

Our TOPOS-VAE is trained on the head mesh dataset under the consistent topology. Inspired by recent work Wang et al. (2026c), the objective combines a basic **vertex loss** L_{vertex}, which is defined as the L_2 distance between the decoded vertices $\hat{\mathcal{V}}$ and ground-truth $\mathcal{V}$, with three additional geometric losses that enforce a smooth surface and regular face connectivity.

Face normal consistency loss L_{normal} penalizes angular deviation between decoded and ground-truth face normals:

$$L_{\text{normal}} = \frac{1}{N_f} \sum_{f \in \mathcal{F}} \left(1 - \frac{\hat{\mathbf{n}}_f \cdot \mathbf{n}_f}{|\hat{\mathbf{n}}_f||\mathbf{n}_f|}\right). \tag{4}$$

For each triangle $f \in \mathcal{F}$, we compute the face normal $\mathbf{n}_f$ from its edge vectors via the cross product

and l_2 normalization.

Face angle consistency loss L_θ constrains local triangle shape to prevent degeneration by supervising interior cosines. For triangle f with vertices v_0, v_1, v_2, c_f^k denotes the cosine of the interior angle at v_k, which is computed from the edge vectors originating from v_k. L_θ is then defined as the L_1 distance between the decoded and ground-truth cosine values:

$$L_\theta = \frac{1}{N_f} \sum_{f \in \mathcal{F}} \left(|\hat{c}_f^0 - c_f^0| + |\hat{c}_f^1 - c_f^1| \right). \tag{5}$$

Discrete Gaussian curvature loss $L_{\mathbf{gc}}$ preserves fine surface curvature via the signed dihedral angle on each interior edge e shared by two faces with normals $\mathbf{n}_a, \mathbf{n}_b$:

$$\kappa_e = \frac{1}{\pi} \cdot \operatorname{sgn}\left((\mathbf{n}_a \times \mathbf{n}_b) \cdot \hat{\mathbf{e}}\right) \cdot \arctan \frac{|\mathbf{n}_a \times \mathbf{n}_b|}{\mathbf{n}_a \cdot \mathbf{n}_b}, \tag{6}$$

where $\hat{\mathbf{e}}$ is the unit edge direction and $\operatorname{sgn}(\cdot)$ is the sign function encoding local convexity or concavity. L_{gc} is defined as the L_1 distance between the decoded and ground-truth signed dihedral angle:

$$L_{\text{gc}} = \frac{1}{N_e} \sum_e |\hat{\kappa}_e - \kappa_e|. \tag{7}$$

Finally, a **KL-divergence loss** $L_{\mathbf{KL}}$ Kingma & Welling (2013) is adopted to regularize the distribution of latent token $\mathcal{L}$ toward a standard Gaussian. In summary, the total loss function of our TOPOS-VAE is defined as the weighted sum of all above losses:

$$L_{\text{VAE}} = \lambda_{\text{vertex}} L_{\text{vertex}} + \lambda_{\text{normal}} L_{\text{normal}} + \lambda_\theta L_\theta + \lambda_{\text{gc}} L_{\text{gc}} + \lambda_{\text{KL}} L_{\text{KL}}. \tag{8}$$

3.2 TOPOS-DiT

We employ rectified flow models Lipman et al. (2022) to generate the latent tokens $\mathcal{L}$ of TOPOS-VAE conditioned on the portrait images to realize our head mesh generation task. Rectified flow models use a linear interpolation forward process, $x(t) = (1-t)x_0 + t\epsilon$, which interpolates between data samples x_0 and noises ϵ with a timestep t. The backward process is represented as a time-dependent vector field, $v(x,t) = \nabla_t x$, moving noisy samples toward the data distribution. In our setting, we utilize a simple transformer backbone TOPOS-DiT to generate latent tokens $\mathcal{L} \in \mathbb{R}^{N \times d}$. The input noisy latent tokens, combined with positional embeddings, are fed into the TOPOS-DiT for denoising. Timestep information is incorporated using adaptive layer normalization (AdaLN) and a gating mechanism Peebles & Xie (2023). For input images, we adopt visual features from DINOv2 Oquab et al. (2023) and visual features are injected through cross attention layers as keys and values. The denoised latent tokens are further decoded into head meshes via TOPOS-VAE.

3.3 TOPOS-TEXTURE

Inspired by Zeng et al. (2025), we treat UV texture generation as a learnable unwrapping task. To overcome scarce high-quality facial texture data Bai et al. (2023), we fine-tune Qwen-Image-Edit Wu et al. (2025) via LoRA adapter Hu et al. (2021) on attention layers, keeping base weights, VAE and text encoder frozen (Figure 3). Crucially, we discard image tokens from the vision-language (VL) branch and retain only text hidden states as condition, avoiding conflicts between generic VL semantics and required UV spatial alignment.

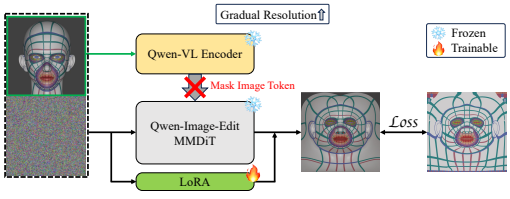
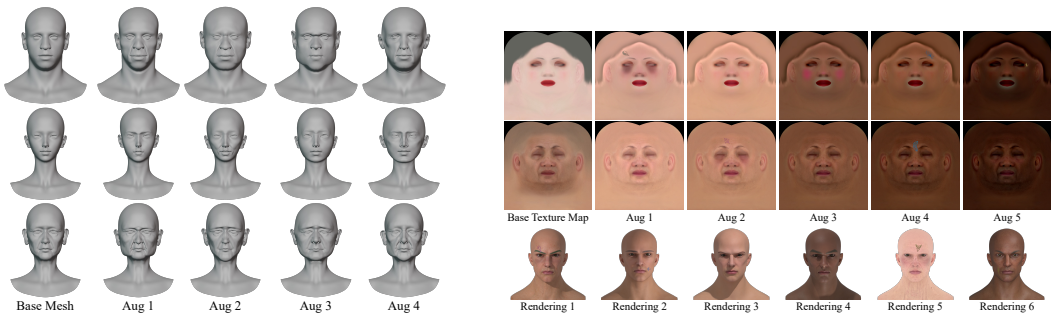


Figure 3: Pipeline of TOPOS-Texture. We fine-tune Qwen-Image-Edit via LoRA, discarding VL image tokens to enforce UV alignment.

Training uses the original flow matching objective with a fixed prompt. To prevent high-frequency details from dominating loss and causing global UV misalignment, we adopt a gradual resolution schedule: training begins at low resolution to establish UV alignment, and progressively introduces higher resolutions to recover sharpness. We further apply dynamic time-shift μ to the logit-normal noise sampling distribution, adapting μ to the image sequence length at each resolution to keep the noise schedule calibrated across stages. This decouples the learning of UV alignment from the recovery of texture sharpness.

Furthermore, to enable relightability, we render training images under randomized lighting while supervising against the same canonical texture map, following Chen et al. (2024b); Liang et al. (2025). This encourages implicit albedo-lighting disentanglement, producing illumination-free texture maps suitable for downstream rendering.



(a) Geometry augmentation results. Aug 1–4 are independent variants from the same base mesh. Please zoom in for details.

(b) Texture map augmentation results. Top two rows: Aug 1–5 from the same base texture. Bottom row: six renderings of the same mesh with different textures.

Figure 4: Visualization of our geometry and texture augmentation algorithms.

4 Experiments

4.1 Data Preparation

We adopt the MetaHuman Epic Games (2021) head mesh, which contains 24,049 vertices, 48,050 edges and 24,002 faces as our uniform reference topology. Our base dataset D_0 consists of 388 in-house head meshes, with 226 sharing the MetaHuman topology from internal 3D animation and video game productions, plus 162 in non-MetaHuman topologies from our internal asset library.

Building upon this, we design a geometric augmentation algorithm to increase data diversity. Specifically, we derive 500 augmented variants from each of the 388 base head meshes in D_0, resulting in the final dataset D, which contains $388 \times 500 = 194{,}000$ head meshes in total. Among these, we denote the subset of head meshes with MetaHuman topology as D_M, comprising $226 \times 500 = 113{,}000$ head meshes in total. Figure 4a visualizes the results of our proposed geometric augmentation, with algorithm details provided in the supplementary material.

For appearance and rendering, we first collect 356 different base texture maps from D. For each base texture map, we randomize 14 skin tones using the official MetaHuman implementation. Then, we apply procedural makeup augmentation over predefined semantic regions (e.g., eyelashes, blush, tattoos). We further perturb PBR strengths and randomize lighting via combinations of area, point and ambient lights, with camera poses slightly perturbed around the frontal view. All meshes are rendered without hair to avoid occlusion. For each mesh, we render eight images using eight random texture maps, and select one per training step of TOPOS-DiT and TOPOS-Texture. Figure 4b shows the texture augmentation results and different render images of the same head mesh.

Our test set D_{test} contains 269 high-quality textured 3D head meshes in MetaHuman topology derived from licensed digital human assets (including 66 official MetaHuman assets), covering diverse identities and appearances. The frontal-view rendering of each mesh serves as the input condition. The underlying textured meshes in D_{test} are used solely for evaluation and are not redistributed.

4.2 Training Details

TOPOS-VAE requires consistent topology and is trained on D_M only. Once trained, it can encode head meshes with diverse topologies. Therefore, TOPOS-DiT and TOPOS-Texture are both trained on the full D. All training runs on 16 NVIDIA H20 GPUs in PyTorch Paszke et al. (2019), taking 3, 7 and 2 days for TOPOS-VAE, TOPOS-DiT and TOPOS-Texture, respectively. More details of experimental settings and hyperparameters are provided in the supplementary material.

4.3 3D Head Mesh Reconstruction

We evaluate TOPOS-VAE reconstruction on D_{test} against two families of baselines: (i) open-source direct mesh generation methods that can generate artist-like meshes from pointclouds, including DeepMesh Zhao et al. (2025a), TreeMeshGPT Lionar et al. (2025), MeshAnythingV2 Chen et al. (2025), BPT Weng et al. (2024b), MeshSilksong Song et al. (2026), MeshRipple Lin et al. (2026), MeshFlow Li et al. (2026) and (ii) neural-field VAEs that recover dense meshes via post-processing algorithm, including Hunyuan3D 2.1 Team (2025), SparseFlex He et al. (2025b), TRELLIS.2 Xiang et al. (2025a), UltraShape 1.0 Jia et al. (2025). Following ConvONet Peng et al. (2020), we report L_1 Chamfer distance (CD), normal consistency (NC) and F-Score against ground truth head meshes. The details of these metrics are provided in the supplementary material.

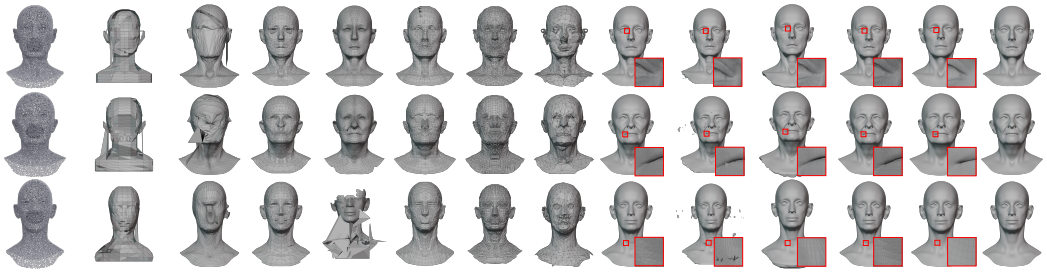


Figure 5: Qualitative results reconstructed by different methods on our test dataset D_{test}. The insets in the red boxes illustrate the topology and connectivity of the reconstructed head meshes.

Qualitative Results Figure 5 compares reconstructions on D_{test}. Direct mesh generation methods emit low-poly head meshes but exhibit noticeable artifacts and distorted structures. Neural-field models produce smooth surfaces yet generate excessive vertices and non-manifold head meshes. In contrast, TOPOS-VAE faithfully reconstructs input pointclouds with a clean and consistent topology.

Quantitative Results TOPOS-VAE outperforms all baselines on all metrics at the top of Tab. 1, while preserving exactly the input vertex and face counts. Neural-field reconstruction models produce millions of vertices, unsuitable for downstream tasks. Direct mesh generation methods yield meshes of poor geometric quality. TOPOS-VAE is non-autoregressive and post-processing free, enabling reconstruction within one second, markedly faster than all baselines.

Table 1: Quantitative results of 3D head meshes reconstruction on our test dataset D_{test}. The top shows comparison with other methods; the bottom presents ablation studies. V-num and F-num denote average vertex and face counts.

Method	CD ↓	NC ↑	F-Score ↑	V-num	F-num
MeshAnythingV2 Chen et al. (2025)	0.119	0.6546	0.0513	1236.78	2097.64
TreeMeshGPT Lionar et al. (2025)	0.0864	0.818	0.0751	4341.90	8073.21
BPT Weng et al. (2024b)	0.0417	0.900	0.113	1518.92	3001.90
DeepMesh Zhao et al. (2025a)	0.131	0.624	0.0513	8564.66	16225.93
MeshSilksong Song et al. (2026)	0.0641	0.867	0.0678	2364.97	4659.99
MeshRipple Lin et al. (2026)	0.0835	0.813	0.0651	4529.84	8903.90
MeshFlow Li et al. (2026)	0.0494	0.859	0.102	2347.17	7676.18
TRELLIS.2 (512) Xiang et al. (2025a)	0.124	0.787	0.0309	508577.93	1018609.32
TRELLIS.2 (1024) Xiang et al. (2025a)	0.124	0.787	0.0309	2032858.15	4070818.52
UltraShape 1.0 Jia et al. (2025)	0.119	0.605	0.0598	1346200.94	2689415.33
Hunyuan3D 2.1 Team (2025)	0.151	0.664	0.0330	634685.12	1269357.18
SparseFlex (512) He et al. (2025b)	0.124	0.786	0.0309	514463.00	1027306.41
SparseFlex (1024) He et al. (2025b)	0.124	0.787	0.0310	2047892.75	4092383.40
VecSet-Learn Zhang et al. (2023a)	0.0132	0.967	0.502	24049	24002
VecSet-FPS Zhang et al. (2023a)	0.0439	0.833	0.156	24049	24002
w/o Geo Loss	0.0094	0.93	0.636	24049	24002
TOPOS-VAE	**0.0055**	**0.987**	**0.915**	24049	24002

Ablation Studies We ablate two design choices in Figure 6 and the bottom part of Tab. 1. First, we replace the Perceiver Resampler with VecSet Zhang et al. (2023a) encoders using learnable queries (VecSet-Learn) or FPS queries (VecSet-FPS). The former converges slowly and loses fine facial details, while the latter fails to reconstruct complete head meshes. Second, training only with L_{vertex} ("w/o Geo Loss") yields significant structural artifacts and irregular topology, confirming the necessity of geometric supervision.

4.4 3D Head Mesh Generation

We evaluate TOPOS-DiT on our test set D_{test}, using the frontal view rendered images as input. We compare it with several 3DMM-based face reconstruction methods, including FFHQ-UV Bai et al. (2023), FLAME 2020 Li et al. (2017), which is implemented in FreeUV Yang et al. (2025), UV-IDM Li et al. (2024), HRN Lei et al. (2023) and Pixel3DMM Giebenhain et al. (2025), as well as representative 3D shape generation models, including DreamFace Zhang et al. (2023b), Hunyuan3D 2.1 Team (2025), TRELLIS.2 Xiang et al. (2025a), UltraShape 1.0 Jia et al. (2025) and commercial models Model A and Model B. All these methods accept single portrait image as input and output 3D head mesh. For closed-source DreamFace and two commercial models, we report only qualitative comparisons through their official web interfaces. Since different methods produce meshes in heterogeneous coordinate frames, scales and spatial extents, we align each generated mesh to its paired ground truth and report CD, NC and F-Score. Full implementation details of 3D head mesh alignment algorithm are provided in the supplementary material.

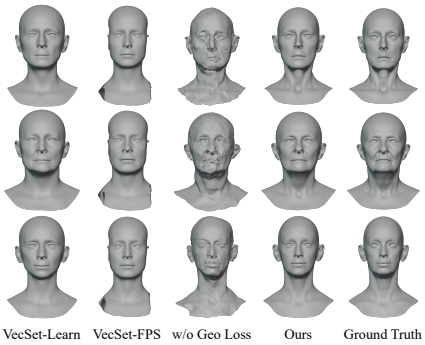


Figure 6: Qualitative results of our ablation studies on TOPOS-VAE using the same three head meshes in Figure 5.

Qualitative Results Figure 7 compares our method with the competing approaches. 3DMM-based methods recover plausible global shapes but miss fine geometric detail due to the limited expressiveness of parameters. General 3D generative models produce meshes with excessive vertices and irregular topology, as highlighted in red boxes. In contrast, TOPOS-DiT produces head

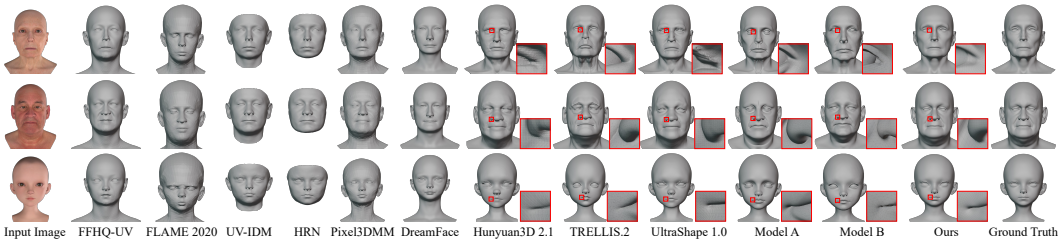


Figure 7: Qualitative results generated by different methods on our test dataset D_{test}. The insets in the red boxes illustrate the topology and connectivity of the generated head meshes.

meshes with clean topology, smooth connectivity and faithful identity fidelity, striking a balance between geometric expressiveness and mesh regularity.

Quantitative Results Tab. 2 demonstrates that TOPOS-DiT outperforms all baselines across CD, NC and F-Score. The substantially higher F-Score over 3DMM-based methods reflects our finer geometric detail beyond the parametric subspace. Compared with general 3D generative models, which produce millions of polygons, our generative results remain well-structured and directly usable in downstream pipelines.

Table 2: Quantitative results of 3D head meshes generation on D_{test}. V-num and F-num denote average vertex and face counts of generated meshes, respectively.

Method	CD ↓	NC ↑	F-Score ↑	V-num	F-num
FFHQ-UV Bai et al. (2023)	<u>0.0150</u>	<u>0.940</u>	0.564	20792	40832
FLAME 2020 Li et al. (2017)	0.0217	0.906	0.458	5118	9976
UV-IDM Li et al. (2024)	0.0461	0.812	0.438	53215	105954
HRN Lei et al. (2023)	0.0634	0.777	0.342	35709	70789
Pixel3DMM Giebenhain et al. (2025)	0.0185	0.920	0.515	5023	9976
Hunyuan3D 2.1 Team (2025)	0.0218	0.870	<u>0.667</u>	746179.62	1492399.77
TRELLIS.2 (1024) Xiang et al. (2025a)	0.0349	0.808	0.520	329620.02	485600.25
UltraShape 1.0 Jia et al. (2025)	0.0225	0.866	0.659	5147079.16	10294336.87
Ours	**0.00613**	**0.978**	**0.857**	24049	24002

4.5 3D Head Texture Generation

We compare TOPOS-Texture with methods capable of generating texture maps introduced in the previous subsection, the same two commercial models, as well as the recent multimodal reasoning model Uni-1 Luma AI (2026), which produces a single unwrapped UV map from three multi-view photographs of the same subject.

Beyond D_{test}, we evaluate on 159 uncurated in-the-wild portrait images (82 males and 77 females) from web sources, academic datasets Karras et al. (2019); Liu et al. (2015) and AI generation tools. Following recent practice Wang et al. (2026c), we frontalize faces and remove occluders using a fine-tuned FLUX.1 Kontext Labs et al. (2025) and the resulting edited images serve as the actual inputs fed into all compared methods. For Uni-1 Luma AI (2026), which uses FFHQ-UV Bai et al. (2023) topology, we only perform qualitative comparison on D_{test} using three multi-view GT renderings with official prompts.



Figure 8: Qualitative results generated by different methods on our test dataset D_{test}. The upper-left insets illustrate the environment maps. Please zoom in for better inspection.

On D_{test}, we render generated and GT meshes from seven viewpoints (one frontal + six random frontal-hemisphere) under six lighting environments at four illumination angles each. For each identity, we apply a facial mask derived from GT mesh to exclude eyes, neck, and shoulders. We report LPIPS Zhang et al. (2018), FID Parmar et al. (2022), KID Bińkowski et al. (2018), and CSIM Deng et al. (2019). For in-the-wild evaluation without GT meshes, we align all competing meshes to ours, render frontally under identical lighting and compute FID, KID as well as CSIM against the edited input image. To measure the geometric similarity, we additionally render an untextured version of each mesh and report CSIM* against the input edited image. The details of these metrics are provided in the supplementary material.

Qualitative Results. Figure 8 shows qualitative comparisons on D_{test}, where each three-row group renders one identity under three environment maps. Our method consistently produces high-fidelity skin textures and accurate color reproduction across all lighting conditions, closely matching the ground truth. 3DMM-based methods and DreamFace fail to recover out-of-distribution geometry such as elongated ears. FFHQ-UV, TRELLIS.2 and two commercial models produce overly smooth textures lacking fine skin detail. Uni-1 exhibits catastrophic texture-geometry misalignment and Hunyuan3D 2.1 shows noisy color artifacts with baked-in shading.

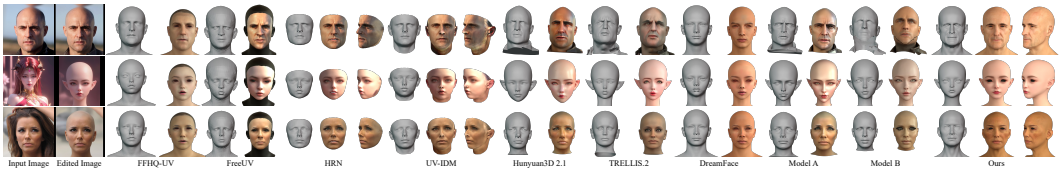


Figure 9: Qualitative results generated by different methods on in-the-wild images. For each method, we show both the geometry and texture renderings under the same environment map side by side. For HRN, UV-IDM and our method, we additionally include a profile view rendering.

In-the-wild comparisons are shown in Figure 9. In terms of geometry, 3DMM-prior methods and DreamFace are limited by their shape bases (e.g., failing on the elongated ears in row 2), while our method recovers identity-specific details such as nasolabial folds and chin structure. For texture, baked-in illumination is a common failure mode in HRN, UV-IDM, Hunyuan3D 2.1 and Model A, whereas our method recovers clean albedo that decouples appearance from lighting and remains faithful to each subject's intrinsic skin tone across novel views.

Table 3: Quantitative results of 3D head texture generation on D_{test} and in-the-wild images. Top: comparison with other methods. Bottom: ablation studies.

Method	D_{test}				In-the-wild			
	LPIPS ↓	FID ↓	KID ↓	CSIM ↑	FID ↓	KID ↓	CSIM ↑	CSIM* ↑
FFHQ-UV Bai et al. (2023)	0.0257	28.883	0.0334	0.738	58.093	0.0413	0.584	0.359
FreeUV Yang et al. (2025)	0.0533	26.671	0.0253	0.526	98.967	0.0877	0.512	0.367
UV-IDM Li et al. (2024)	0.0236	9.624	0.00664	0.752	63.276	0.0429	0.694	0.416
HRN Lei et al. (2023)	0.0321	8.289	0.00554	0.773	91.572	0.0836	**0.723**	0.398
Hunyuan3D 2.1 Team (2025)	0.0319	20.009	0.0149	0.655	81.011	0.0352	0.551	0.394
TRELLIS.2 Xiang et al. (2025a)	0.0875	57.233	0.0385	0.545	84.826	0.0346	0.521	0.422
w/o LoRA	0.0138	2.678	0.00210	0.826	45.348	0.0221	0.611	-
w/ VL Img Tokens	0.0142	2.992	0.00203	0.789	43.683	0.0206	0.606	-
w/o Res. Sched.	0.0139	2.933	0.00251	0.852	42.681	0.0195	0.605	-
Ours	**0.0119**	**2.456**	**0.00199**	**0.855**	**41.066**	**0.0187**	0.613	**0.438**

Quantitative Results. The top part of Tab. 3 reports the comparison on both D_{test} and in-the-wild images. On D_{test}, TOPOS-Texture is the best across all four metrics, with particularly large margins on FID and KID. On in-the-wild images, our method again attains the lowest FID and KID. The CSIM gap to HRN and UV-IDM stems from those methods baking the original illumination into their textures, which inflates identity similarity under the input viewpoint but degrades under novel views, as demonstrated in the profile view renderings in Figure 9. CSIM* further reflects this effect by evaluating untextured renderings, where our method achieves the highest score.

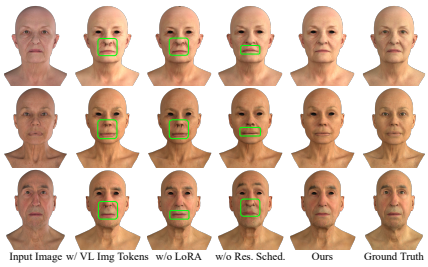


Figure 10: Qualitative results of our ablation studies on TOPOS-Texture. The green boxes highlight the texture-geometry misalignment.

Ablation Studies We validate our texture module's designs through three ablations: (1) retaining VL image tokens, (2) training directly at full resolution without the resolution schedule or adaptive time-shift μ and (3) unfreezing the backbone network ("w/o LoRA"). As reported in Tab. 3, all variants yield inferior metrics compared to our full method. We also show the visual effects of ablations in Figure 10, where these three variants exhibit notable texture-geometry misalignment, particularly in semantically sensitive regions such as nose and mouth. This consistent degradation confirms that discarding VL image tokens, the adaptive resolution schedule and frozen backbone are essential for high-quality texture unwrapping.

4.6 Application

Finally, we show that the industry-standard topology of our outputs seamlessly supports downstream production tasks. As shown in Figure 1, given a single portrait image, our TOPOS produces 3D head mesh that faithfully preserves geometry and appearance details. After being bound to the pre-defined facial rig, the generated meshes support rig-driven animation across diverse facial expressions, including mouth opening, eye blinking, brow furrowing and smiling, while maintaining geometric consistency and texture fidelity.

5 Conclusion

In this paper, we present TOPOS framework, which combines the Perceiver Resampler with GNN decoder for fixed-topology mesh modeling, a rectified flow transformer for 3D head geometry generation and a fine-tuned Qwen-Image-Edit model for relightable UV textures. Extensive experiments demonstrate state-of-the-art performance across 3D head mesh reconstruction, generation and texture generation. Since TOPOS conforms to the studio-defined reference topology, its generated heads are directly compatible with production pipelines for downstream rigging and character animation. We hope TOPOS offers a scalable solution for digital human creation and inspires future work on generalizable 3D human digitization as well as uniform topology 3D mesh generation.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks with required disclosure (including generating synthetic datasets, formulating mathematical claims or proofs, designing research methodologies, implementing code, or interpreting results). Additionally, we used generative AI tools solely for the task with recommended disclosure, specifically to edit the research paper to improve readability, grammar, and phrasing. All AI-assisted text edits were carefully reviewed and verified by the authors to ensure that the original technical meaning was strictly preserved without introducing any misrepresentations or factual errors. We take full responsibility for the final content of this work, including all text and claims.

ETHICS STATEMENT

This work advances single image conditioned 3D head generation. Our method is not intended for malicious use, and all synthesized content should clearly indicate its artificial nature. We acknowledge potential misuse, such as deepfakes, and are developing tools to help detect synthetic images and videos. At the same time, our technology can support education, communication assistance, and therapeutic applications, reflecting our commitment to responsible and ethical AI development.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 362–371, 2023.

Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pp. 586–606. Spie, 1992.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pp. 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. ISBN 0201485605. doi: 10.1145/311535.311556. URL https://doi.org/10.1145/311535.311556.

James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018.

Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm siggraph*, volume 2012, pp. 1–7. vol. 2012, 2012.

Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *European conference on computer vision*, pp. 74–92. Springer, 2022.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Zhibin Wang, Jingyi Yu, Gang Yu, et al. Meshxl: Neural coordinate field for generative 3d foundation models. *Advances in Neural Information Processing Systems*, 37:97141–97166, 2024a.

Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination, 2024b. URL https://arxiv.org/abs/2404.11593.

Yiwen Chen, Yikai Wang, Yihao Luo, Zhengyi Wang, Zilong Chen, Jun Zhu, Chi Zhang, and Guosheng Lin. Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13922–13931, 2025.

Dmitry Chetverikov, Dmitry Stepanov, and Pavel Krsek. Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm. *Image and vision computing*, 23(3):299–309, 2005.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gVM2AZ5xA6.

Jiashu Dai, Along Wang, Binfan Ni, and Tao Cao. High-quality facial albedo generation for 3d face reconstruction from a single image using a coarse-to-fine approach. *arXiv preprint arXiv:2506.13233*, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Abdallah Dib, Luiz Gustavo Hafemann, Emeline Got, Trevor Anderson, Amin Fadaeinejad, Rafael MO Cruz, and Marc-André Carbonneau. Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1770–1780, 2024.

Epic Games. Metahuman creator. https://www.unrealengine.com/en-US/metahuman-creator, 2021.

Xiang Feng, Chang Yu, Zoubin Bi, Yintong Shang, Feng Gao, Hongzhi Wu, Kun Zhou, Chenfanfu Jiang, and Yin Yang. Arm: Appearance reconstruction model for relightable 3d generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21425–21437. IEEE, 2025.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Yuxuan Han, Junfeng Lyu, and Feng Xu. High-quality facial geometry and appearance capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 697–707, 2024.

Yuxuan Han, Junfeng Lyu, Kuan Sheng, Minghao Que, Qixuan Zhang, Lan Xu, and Feng Xu. Facial appearance capture at home with patch-level reflectance prior. *ACM Transactions on Graphics (TOG)*, 44(4):1–16, 2025.

Zekun Hao, David W Romero, Tsung-Yi Lin, and Ming-Yu Liu. Meshtron: High-fidelity, artist-like 3d mesh generation at scale. *arXiv preprint arXiv:2412.09548*, 2024.

Xianglong He, Junyi Chen, Di Huang, Zexiang Liu, Xiaoshui Huang, Wanli Ouyang, Chun Yuan, and Yangguang Li. Meshcraft: Exploring efficient and controllable mesh generation with flow-based dits. *arXiv preprint arXiv:2503.23022*, 2025a.

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*, pp. 14822–14833, 2025b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Tanghui Jia, Dongyu Yan, Dehao Hao, Yang Li, Kaiyi Zhang, Xianyi He, Lanjiong Li, Jinnan Chen, Lutao Jiang, Qishen Yin, Long Quan, Ying-Cong Chen, and Li Yuan. Ultrashape 1.0: High-fidelity 3d shape generation via scalable geometric refinement. *arxiv preprint arXiv:2512.21185*, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9269–9284, 2021.

Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images, 2023.

Hong Li, Yutang Feng, Song Xue, Xuhui Liu, Bohan Zeng, Shanglin Li, Boyu Liu, Jianzhuang Liu, Shumin Han, and Baochang Zhang. Uv-idm: identity-conditioned latent diffusion model for face uv-texture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10585–10595, 2024.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813.

Weiyu Li, Antoine Toisoul, Tom Monnier, Roman Shapovalov, Rakesh Ranjan, Ping Tan, and Andrea Vedaldi. Meshflow: Efficient artistic mesh generation via meshvae and flow-based diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. Highlight.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models, 2025. URL https://arxiv.org/abs/2501.18590.

Junkai Lin, Hang Long, Huipeng Guo, Jielei Zhang, Jiayi Yang, Tianle Guo, Yang Yang, Jianwen Li, Wenxiao ZHANG, Matthias Nießner, and Wei Yang. Meshripple: Structured autoregressive generation of artist-meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12706–12718, June 2026.

Stefan Lionar, Jiabin Liang, and Gim Hee Lee. Treemeshgpt: Artistic mesh generation with autoregressive tree sequencing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26608–26617, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Hang Long, Tianhao Zhao, Junkai Lin, Youjia Zhang, Huipeng Guo, Rendong Liang, Jiale Xu, Jozef Hladký, Matthias Nießner, Yuanming Hu, and Wei Yang. Lato.2: Factorized 3d mesh generation with vertex and topology flow, 2026. URL https://arxiv.org/abs/2607.10623.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pp. 347–353. 1998.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luma AI. Uni-1. https://lumalabs.ai/uni-1, 2026.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

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 *ECCV*, 2020.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bo Pang, Zhongtian Zheng, Guoping Wang, and Peng-Shuai Wang. Learning the geodesic embedding with graph neural networks. *ACM Transactions on Graphics (TOG)*, 42(6):1–12, 2023.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11410–11420, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pp. 296–301. Ieee, 2009.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *european conference on computer vision*, pp. 523–540. Springer, 2020.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14184–14194, 2025a.

Yuda Qiu, Zitong Xiao, Yiwei Zuo, Zisheng Ye, Weikai Chen, and Xiaoguang Han. Avatartex: High-fidelity facial texture reconstruction from single-image stylized avatars. *arXiv preprint arXiv:2511.06721*, 2025b.

George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. Smirk: 3d facial expressions through analysis-by-neural-synthesis. *arXiv preprint arXiv:2404.04104*, 2024.

Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1259–1268, 2017.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19615–19625, 2024.

Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3693–3702, 2017.

Gaochao Song, Zibo Zhao, Haohan Weng, Jingbo Zeng, Rongfei Jia, and Shenghua Gao. Topology-preserved auto-regressive mesh generation in the manner of weaving silk. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qi Sun, Kiyohiro Nakayama, Jing Nathan Yan, Qixing Huang, Alexander Rush, Leonidas Guibas, Gordon Wetzstein, Jing Liao, and Guandao Yang. Meshflow: Mesh generation with equivariant flow matching. In *ACM SIGGRAPH Conference Papers*, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Jiaxiang Tang, Zhaoshuo Li, Zekun Hao, Xian Liu, Gang Zeng, Ming-Yu Liu, and Qinsheng Zhang. Edgerunner: Auto-regressive auto-encoder for artistic mesh generation. *arXiv preprint arXiv:2409.18114*, 2024.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in neural information processing systems*, 35:10021–10039, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chunshi Wang, Haohan Weng, Junliang Ye, Biwen Lei, Yang Li, Zibo Zhao, Zeqiang Lai, Kaiyi Zhang, Yunhan Yang, Zhuo Chen, et al. Polyflow: Continuous topology embedding flow matching for artist-style mesh generation. *arXiv preprint arXiv:2606.30673*, 2026a.

Hanxiao Wang, Biao Zhang, Weize Quan, Dong-Ming Yan, and Peter Wonka. iflame: Interleaving full and linear attention for efficient mesh generation. *arXiv preprint arXiv:2503.16653*, 2025.

Hanxiao Wang, Ying-Tian Liu, Yuan-Chen Guo, Qi-Yuan Feng, Zi-Xin Zou, Ding Liang, Biao Zhang, and Yan-Pei Cao. Nexus: Native mesh generation with diffusion. *ACM Transactions on Graphics (TOG)*, 45(4):1–14, 2026b.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

Ruizhe Wang, Chih-Fan Chen, Hao Peng, Xudong Liu, Oliver Liu, and Xin Li. Digital twin: Acquiring high-fidelity 3d avatar from a single image. *arXiv preprint arXiv:1912.03455*, 2019a.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019b.

Yunmu Wang, Zoubin Bi, Bowen Cai, Chenchu Rong, Jinlong Wang, Junchen Deng, Aocheng Huang, Jidong Jia, and Huan Fu. High-fidelity single-image head modeling with industry-grade topology. *arXiv preprint arXiv:2605.04524*, 2026c.

Haohan Weng, Yikai Wang, Tong Zhang, CL Chen, and Jun Zhu. Pivotmesh: Generic 3d mesh generation via pivot vertices guidance. *arXiv preprint arXiv:2405.16890*, 2024a.

Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, Tong Zhang, Shenghua Gao, and C. L. Philip Chen. Scaling mesh generation via compressive tokenization. *arXiv preprint arXiv:2411.07025*, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3d generation. *Tech report*, 2025a.

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*, pp. 21469–21480, 2025b.

Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 551–561. IEEE, 2025a.

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*, volume 44, pp. e70198. Wiley Online Library, 2025b.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6184–6193, 2020.

Rui Xu, Dafei Qin, Kaichun Qiao, Qiujie Dong, Huaijin Pi, Qixuan Zhang, Longwen Zhang, Lan Xu, Jingyi Yu, Wenping Wang, et al. Strips as tokens: Artist mesh generation with native uv segmentation. *arXiv preprint arXiv:2604.09132*, 2026.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 601–610, 2020.

Xingchao Yang, Takafumi Taketomi, Yuki Endo, and Yoshihiro Kanamori. Freeuv: Ground-truth-free realistic facial uv texture recovery via cross-assembly inference strategy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 326–337, 2025.

Chong Zeng, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. Renderformer: Transformer-based neural rendering of triangle meshes with global illumination. In *ACM SIGGRAPH 2025 Conference Papers*, 2025.

Hao Zeng, Wei Zhang, Changjie Fan, Tangjie Lv, Suzhen Wang, Zhimeng Zhang, Bowen Ma, Lincheng Li, Yu Ding, and Xin Yu. Flowface: Semantic flow-guided shape-aware face swapping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3367–3375, 2023.

Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2315–2324, 2019.

Xiaoxing Zeng, Zhelun Wu, Xiaojiang Peng, and Yu Qiao. Joint 3d facial shape reconstruction and texture completion from a single image. *Computational Visual Media*, 8(2):239–256, 2022.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023a.

Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Ruowen Zhao, Junliang Ye, Zhengyi Wang, Guangce Liu, Yiwen Chen, Yikai Wang, and Jun Zhu. Deepmesh: Auto-regressive artist-mesh creation with reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10612–10623, 2025a.

Tianhao Zhao, Youjia Zhang, Hang Long, Jinshen Zhang, Wenbing Li, Yang Yang, Gongbo Zhang, Jozef Hladkỳ, Matthias Nießner, and Wei Yang. Lato: 3d mesh flow matching with structured topology preserving latents. *arXiv preprint arXiv:2603.06357*, 2026.

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. *Advances in neural information processing systems*, 36:73969–73982, 2023.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025b.

Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pp. 250–269. Springer, 2022.

Timo Zinßer, Jochen Schmidt, and Heinrich Niemann. Point set registration with integrated scale estimation. In *International conference on pattern recognition and image processing*, pp. 116–119, 2005.

APPENDIX

TABLE OF CONTENTS

We first provide a brief overview of our supplementary material. This supplementary material consists of the following sections and contents:

- **section A**: Experimental settings and hyperparameters details.
- **section B**: Geometric augmentation algorithm details.
- **section C**: Full implementation details of 3D head mesh alignment algorithm.
- **section D**: Evaluation metrics details.
- **section E**: Physically Based Rendering.
- **section F**: Additional experimental results and visualizations.
- **section G**: Analysis of our current limitations.

A EXPERIMENTAL SETTINGS AND HYPERPARAMETERS

The encoder of our TOPOS-VAE uses 8 Perceiver Resampler blocks with hidden width 128, 8 attention heads of dimension 64 and an MLP expansion ratio of 4. The GNN decoder has $L = 4$ graph upsampling stages. The shape of latent tokens $\mathcal{L}$ is 1513×32. We optimize the network with AdamW optimizer Loshchilov & Hutter (2017) ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay = 0.1). The initial learning rate is 1×10^{-3}, decayed with a polynomial schedule of power 0.9. The loss weights are set to $\lambda_{\text{vertex}} = 2$, $\lambda_{\text{normal}} = 1$, $\lambda_\theta = 5$, $\lambda_{\text{gc}} = 5$, and $\lambda_{\text{KL}} = 1 \times 10^{-3}$, respectively.

With trained TOPOS-VAE frozen, TOPOS-DiT learns a conditional flow matching model in its latent space. We use a DiT-style transformer backbone with model channels 1024, 24 blocks, 16 attention heads with head dimension 64 and an MLP expansion ratio of 4. The input and output channels match the TOPOS-VAE's latent dimension of 32, while cross-attention to the image condition uses 1024 channels. The image features are extracted by a frozen DINOv2 ViT-L/14-with-registers Oquab et al. (2023) backbone, producing 1374 patch tokens of dimension 1024, layer-normalized before being fed into the DiT. The timesteps are drawn from a logit-normal distribution with mean 0 and standard deviation 1, (i.e. $t = \sigma(\mathcal{N}(0, 1))$). To enable classifier-free guidance, the image condition is dropped with probability 0.1 during training. The model is trained with AdamW Loshchilov & Hutter (2017) at a learning rate of 1×10^{-4} under the same polynomial schedule of TOPOS-VAE. At inference stage of TOPOS-DiT, we use a Flow Euler sampler with 50 steps and a CFG guidance scale of 3.0.

TOPOS-Texture fine-tunes Qwen-Image-Edit Wu et al. (2025) via LoRA with rank 32. The learning rate follows a cosine decay schedule from 5×10^{-5} to 3.5×10^{-5}. We adopt a progressive resolution training strategy, starting from 256^2 and gradually increasing the proportion of higher-resolution samples (384^2, 512^2, 768^2), until training exclusively on 1024^2 images. The input condition is dropped with probability 0.2 for classifier-free guidance. All other hyperparameters follow the default settings of Qwen-Image-Edit Wu et al. (2025). At inference stage of TOPOS-Texture, we apply a CFG guidance scale of 1.2.

B GEOMETRIC AUGMENTATION ALGORITHM

Our geometric augmentation algorithm is mainly implemented via region-wise blendshape sampling. We group the face into nine semantic regions: {*Brow, Eye, Nose, Mouth, Ear, Forehead, Cheekbone, FaceShape, Chin*}. For each base mesh M and each facial region r, we randomly select three blendshapes $\{B_{r,1}, B_{r,2}, B_{r,3}\}$ from the corresponding blendshape set, where $B_{r,i} \in \mathbb{R}^{V \times 3}$ denotes the i-th blendshape of region r, and V is the total number of head mesh vertices. Their coefficients are independently sampled within artist-defined bounds,

$$w_{r,i} \sim \mathcal{U}(l_{r,i}, u_{r,i}), \quad i = 1, 2, 3, \tag{9}$$

where $l_{r,i}$ and $u_{r,i}$ denote the lower and upper bounds specified by artists. To avoid excessive deformation within a local region, we normalize the coefficients only when their sum exceeds 1,

$$\tilde{w}_{r,i} = \begin{cases} \dfrac{w_{r,i}}{\sum_{j=1}^{3} w_{r,j}}, & \text{if } \sum_{j=1}^{3} w_{r,j} > 1, \\ w_{r,i}, & \text{otherwise.} \end{cases} \tag{10}$$

The augmented head mesh M' is then written as

$$M' = M + \sum_{r} \sum_{i=1}^{3} \tilde{w}_{r,i} B_{r,i}. \tag{11}$$

All parameter ranges are specified by artists to suppress implausible facial shapes, and no additional manual inspection or post-filtering is applied, which enables efficient large-scale geometric augmentation.

C 3D Head Mesh Alignment for Evaluation.

Since different methods output face meshes in heterogeneous coordinate frames, with varying scales, and with different spatial extents (some emit only the frontal face shell, while the ground-truth meshes cover the full head down to the neck and shoulders), a direct vertex-wise comparison is meaningless. Before computing any geometric metric, we therefore align each generated mesh $\mathcal{M}_P$ to its paired ground-truth mesh $\mathcal{M}_G$ by a 7-DoF similarity transform $(s, \mathbf{R}, \mathbf{t}) \in \mathbb{R}^+ \times SO(3) \times \mathbb{R}^3$ so that $\mathbf{p} \mapsto s\mathbf{R}\mathbf{p} + \mathbf{t}$ maps the facial region of $\mathcal{M}_P$ onto that of $\mathcal{M}_G$. The ground-truth 3D head mesh is kept fixed throughout.

We solve for $(s, \mathbf{R}, \mathbf{t})$ by trimmed similarity ICP Besl & McKay (1992); Zinßer et al. (2005); Chetverikov et al. (2005). At each iteration, every source vertex is matched to its nearest neighbor on $\mathcal{M}_G$ via a k-d tree, and only the top τ=50% of correspondences with the smallest residuals are retained. A closed-form Umeyama estimate Umeyama (2002) is computed on this trimmed set and composed with the running transform. The trimming is essential: it automatically rejects correspondences that belong to regions present in one mesh but not the other (e.g., neck and shoulders in the ground truth, or a frontal-only shell in the prediction), so that the fit is dominated by the mutually-overlapping facial region rather than being dragged by extraneous geometry.

Because ICP is non-convex, we initialize it carefully. A coarse centroid translation and a bbox-diagonal-based scale s_0 provide a starting point. On top of this, we perform a small grid search over $0°, 90°, 180°, 270°$ yaw rotations and over scale multipliers $\{0.5,\ 0.7,\ 1.0,\ 1.4\} \cdot s_0$; the scale sweep is necessary because the bbox-diagonal heuristic is biased when the source covers only a subset of the target. For each candidate, a short 30-iteration trimmed ICP is run, and the one with the lowest residual seeds a final full-length refinement (up to 300 iterations, τ=50%).

The recovered transform is then applied to every vertex of $\mathcal{M}_P$, while connectivity, UVs, materials, and vertex normals are preserved verbatim (normals are rotated by $\mathbf{R}$ only and renormalized, since uniform scaling preserves their direction). For methods that ship their output 3D head meshes as GLB files, the same similarity is applied to the scene-graph root transform so that the full material/texture/node hierarchy remains intact. This yields a set of aligned meshes on which we compute all reported metrics.

To prevent extraneous geometry retained by some methods (e.g., neck and shoulders) from dominating the metrics, we clip $\mathcal{M}_G$ to the axis-aligned bounding box of $\mathcal{M}_R$, inflated by 5%, via exact plane-based triangle slicing. $\mathcal{M}_R$ itself is left intact, so that predictions failing to cover the face are still penalized through incompleteness. All Euclidean distances are normalized by the bounding-box diagonal of $\mathcal{M}_R$, making the F-score thresholds scale-free and comparable across different 3D head meshes.

D Evaluation Metrics Details

We denote the generated mesh and ground-truth mesh by $\mathcal{M}_G$ and $\mathcal{M}_R$, on which we randomly sample a set of N = 10k points $G = \{x_i\}_1^N$ and $R = \{y_i\}_1^N$, respectively. Define $\mathcal{P}_A(x) =$

$\text{argmin}_{y \in A}||x - y||$, which finds the closest point of x from a point set A. The L_1 Chamfer distance is defined as:

$$CD = \frac{1}{N} \sum_i \|x_i - \mathcal{P}_R(x_i)\| + \frac{1}{N} \sum_i \|y_i - \mathcal{P}_G(y_i)\| . \tag{12}$$

We define $\mathcal{N}(x)$ as an operator that returns the corresponding normal of an input point, then the normal consistency is define as:

$$NC = \frac{1}{N} \sum_i \left|\mathcal{N}(x_i) \cdot \mathcal{N}\big(\mathcal{P}_R(x_i)\big)\right| + \frac{1}{N} \sum_i \left|\mathcal{N}(y_i) \cdot \mathcal{N}\big(\mathcal{P}_G(y_i)\big)\right|. \tag{13}$$

The F-Score is defined as the harmonic mean between the precision and the recall of points that lie within a certain distance between $\mathcal{M}_G$ and $\mathcal{M}_R$.

For 3D head texture evaluation, we calculate the perceptual similarity metric LPIPS Zhang et al. (2018) based on AlexNet Krizhevsky et al. (2012) between the rendered images of generated head meshes and ground truth head meshes.

We also utilize FID Parmar et al. (2022) and KID Bińkowski et al. (2018) to compare the distribution of these two images sets. FID is defined as:

$$\text{FID} = \|\mu_G - \mu_R\|^2 + \text{Tr}\big(\Sigma_G + \Sigma_R - 2(\Sigma_R \Sigma_G)^{1/2}\big), \tag{14}$$

and KID is defined as the squared Maximum Mean Discrepancy (MMD) between the two feature sets with a polynomial kernel $k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3$, where d is the feature dimension. Given features $\{g_i\}_{i=1}^m$ from G and $\{r_j\}_{j=1}^n$ from R, KID is computed using the unbiased estimator:

$$\begin{aligned} \text{KID} = &\frac{1}{m(m-1)} \sum_{i \neq i'}^{m} k(g_i, g_{i'}) + \frac{1}{n(n-1)} \sum_{j \neq j'}^{n} k(r_j, r_{j'}) \\ &- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(g_i, r_j), \end{aligned} \tag{15}$$

where G and R denote the features of the generated images set and ground truth images set, which is extracted by Inception-v3 model Szegedy et al. (2016). μ and Σ denote the mean and covariance matrices of each image set.

Finally, we utilize Cosine SIMilarity of identity features (CSIM) to measure the identity preservation between two portrait images, which can be either rendered images or in-the-wild portrait images. We calculate CSIM through the cosine similarity of two embeddings from the representative pretrained face recognition network ArcFace Deng et al. (2019).

E Physically Based Rendering

Our current model predicts canonical albedo rather than a complete set of physically based material attributes. Nevertheless, the predicted albedo can be readily incorporated into a standard PBR pipeline by complementing it with heuristic material parameters, inspired by Wang et al. (2026c), and rendering with a GGX-based microfacet BRDF under the Disney principled shading framework Burley & Studios (2012). This provides a practical way to enable PBR-compatible rendering with our current outputs, without requiring the model to explicitly infer the full set of intrinsic material properties. In future work, inspired by recent advances in relightable appearance reconstruction and PBR material generation Feng et al. (2025); Xiong et al. (2025a), we plan to collect richer PBR supervision, including roughness, normal maps, and other spatially varying material attributes, and extend our model to explicitly predict these properties for more realistic and controllable rendering under novel illumination.

F More Generative Results

We present more 3D head results generated by our TOPOS framework in figure 11. What's more, thanks to MetaHuman's facial capture capability Epic Games (2021), we drive the 3D heads generated by our TOPOS shown in the main manuscript and supplementary material using performances

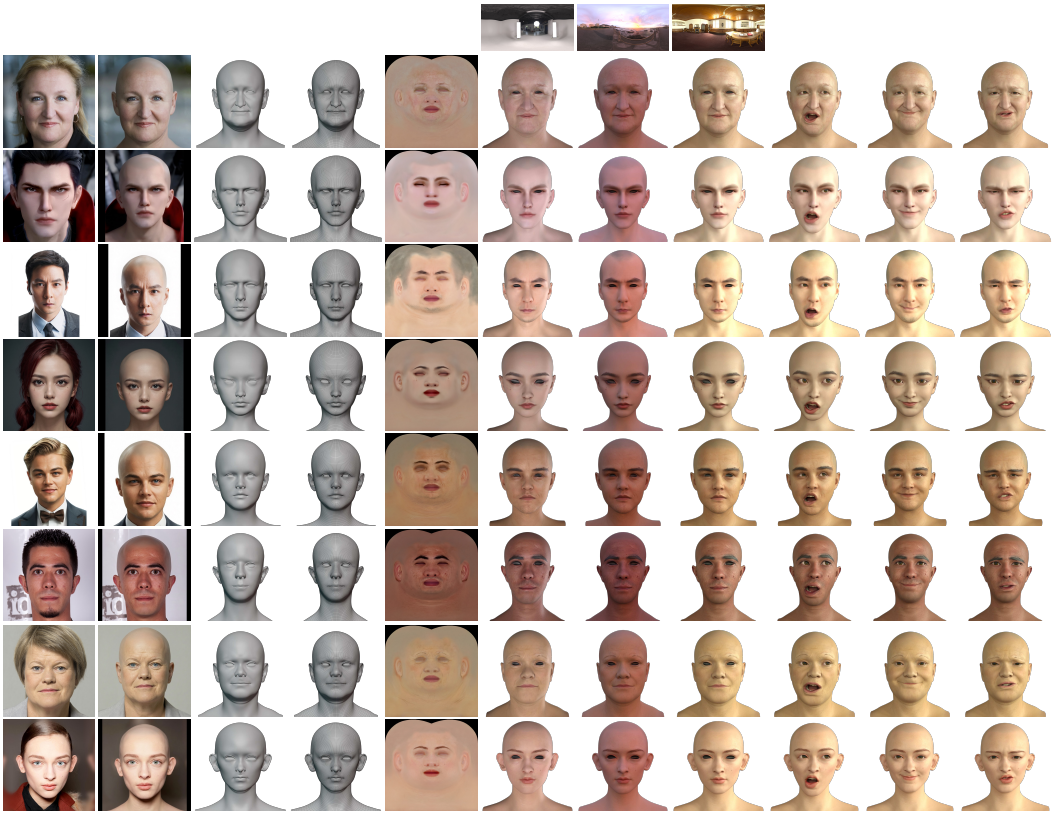


Figure 11: More generative results of our method on different in-the-wild images. From left to right are the input images, edited images, generated geometry, mesh topology, generated texture maps, rendering images under different lighting conditions and animation results across different expressions, respectively. The first row illustrates the environment maps. Please zoom in for better inspection.

captured from the professional facial motion actors, with results demonstrated in the supplementary video.

G LIMITATIONS

While our approach demonstrates strong overall performance, it may exhibit degradation under challenging scenarios such as severe occlusions, extreme low-resolution inputs, or extreme poses that exceed the correction capability of the current preprocessing pipeline. Integrating more robust and state-of-the-art image editing models (e.g., ChatGPT Images 2.5 or Nano Banana 2) would enhance the fidelity and robustness of face frontalization and occluder removal before passing to our TOPOS framework.