Monocular Avatar Reconstruction via Cascaded Diffusion Priors and UV-Space Differentiable Shading

Hong Li[1,2⋆], Minqi Meng[⋆1], Yanjun Liang[1], Chongjie Ye[3], Houyuan Chen[4], Weiqing Xiao[5], Xianda Guo[6†], Guojun Lei[7], Xuhui Liu[8], Chaojie Yang[1], Yanlun Peng[2], Hao Zhao[9], and Baochang Zhang[1‡]

[1] Beihang University, China [2] Great Wall Motors, China
[3] CUHK-Shenzhen, China [4] HKUST, China
[5] Nanjing University, China [6] Wuhan University, China
[7] Zhejiang University, China [8] KAUST, Saudi Arabia
[9] AIR, Tsinghua University, China
Project Page: marcus-avatar.github.io

Abstract. Reconstructing high-fidelity, relightable 3D avatars from a single in-the-wild image is a challenging ill-posed problem, primarily hindered by the scarcity of high-quality PBR data and the complexity of disentangling illumination from intrinsic materials. In this paper, we present a data-efficient framework that leverages the robust priors of a unified pre-trained diffusion backbone to sequentially address texture completion, delighting, and material decomposition. Unlike existing methods that rely on fragmented pipelines or extensive proprietary datasets, we utilize cascaded Low-Rank Adaptations (LoRAs) to adapt the strong generative prior of the diffusion model for each sub-task in UV space. Specifically, we first employ an Inpainting LoRA to complete missing UV textures caused by occlusion, leveraging the model's semantic understanding to generate semantically and photometrically coherent details. Subsequently, a Light-Homogenization LoRA and a novel Cross-Intrinsic Attention mechanism are introduced to remove baked-in lighting and collaboratively synthesize pixel-aligned PBR maps (Albedo, Normal, Roughness, Specular, and Displacement). To ensure physical plausibility, we impose a UV-space differentiable BRDF shading loss during the decomposition stage, forcing the generative process to adhere to the rendering equation without the artifacts typical of rasterization-based supervision. Extensive experiments demonstrate that our method, trained on fewer than 100 real 3D scans, generates comprehensive, 4K-resolution PBR assets with superior realism and generalization compared to state-of-the-art methods, and all training code and model weights will be released upon acceptance.

Keywords: 3D avatar · face reconstruction · Diffusion model · material estimation · differentiable rendering

⋆ Equal contribution. †Project leader. ‡Corresponding author.

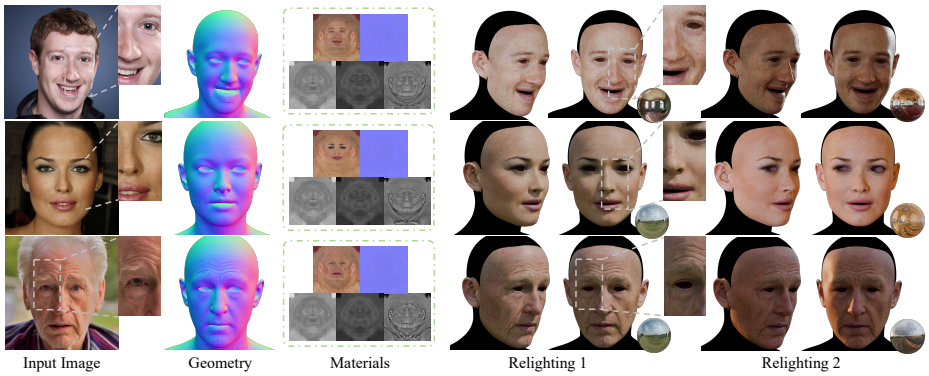


Fig. 1: High-Fidelity 3D Avatar Reconstruction. From a single input image, we reconstruct high-fidelity 3D geometry and PBR materials (albedo, normal, packed maps) to enable relightable avatar synthesis. As shown in the novel lighting results (right), our method accurately recovers fine details (e.g., wrinkles, moles) whilst maintaining identity consistency.

1 Introduction

Photorealistic avatars are pivotal to immersive experiences in virtual reality, multimedia, and video games [51,58]. These personalized digital doubles serve as a crucial interface, bridging physical identities with digital environments. While Light Stage systems [12] have long established the gold standard for acquiring high fidelity avatars, their prohibitive cost and reliance on complex, studio based capture setups severely limit their accessibility for mass adoption. Therefore, monocular 3D face reconstruction, recovering geometry and texture from a single in-the-wild image, remains a fundamental challenge in computer vision. In this paper, we focus on relightable face and head PBR asset reconstruction rather than full-body, dynamic, or accessory-complete avatars.

Traditional approaches rely mainly on 3D Morphable Models (3DMMs) [5, 39], which represent facial variations within a compact linear subspace. While robust, these linear models inherently struggle to capture high frequency details. To overcome this, recent research has shifted towards neural synthesis paradigms. Early works focused on texture completion [13, 19, 20, 32, 59]; however, they typically produce textures with baked-in illumination, rendering the assets unsuitable for relighting in physically based rendering (PBR) pipelines.

To achieve disentangled PBR reconstruction, recent methods typically train generative networks (GANs or Diffusion) to predict nonlinear texture bases [2, 9, 16, 18, 27–29, 38]. However, these approaches face a critical data efficiency and generalization dilemma. First, high quality PBR training data is scarce: synthetic data often lacks realism, while real Light Stage scans are difficult to scale. Second, methodological bottlenecks persist: GAN-based methods suffer from mode collapse and over smoothing, while recent attempts to finetune diffusion models on small real datasets (e.g., UltraAvatar [62]) often degrade the model's prior

knowledge, leading to poor generalization on in-the-wild images. Furthermore, most existing frameworks rely on differentiable rasterization with 2D screen-space supervision. This optimization paradigm is sensitive to occlusions (e.g., hair, glasses) present in input images, often inadvertently "baking" these artifacts into the reconstructed texture.

In this work, we present a novel framework that addresses these challenges by synergizing the robust priors of large-scale generative models with a modern rendering workflow (Fig. 1). Our core insight is that we can circumvent the need for massive proprietary datasets by effectively augmenting a compact set of high-quality scans via modern rendering engines, incorporating pose and occlusion priors derived from public in-the-wild face datasets. By combining these augmented physical priors with a pre-trained modern diffusion backbone, our method achieves high-fidelity material estimation without requiring extensive real-world data. Crucially, we enforce physically based rendering constraints directly on the estimated materials in UV space. This strategy ensures the physical consistency and high quality of the textures while mitigating the occlusion artifacts commonly observed in previous rasterization-based approaches.

Our pipeline begins with a precise geometric alignment step. To support high-fidelity texture generation, we employ a robust geometry estimator enhanced by DINOv3 features, ensuring accurate UV mapping. With the geometry fixed, our core contribution lies in a unified diffusion-based texture framework adapted via cascaded LoRA modules. This backbone sequentially performs texture inpainting, light homogenization, and intrinsic material estimation. To ensure physical plausibility, we introduce a Cross-Intrinsic Attention mechanism that coordinates distinct material branches (Albedo, Normal, Roughness, Specular, Displacement), ensuring coherent high-frequency details. Trained on fewer than 100 real 3D scans, our model generates comprehensive, 4K-resolution PBR assets that generalize seamlessly to diverse in-the-wild images.

In summary, our key contributions are:

- We propose a data-efficient framework for high-fidelity 3D avatar reconstruction that achieves state-of-the-art quality using fewer than 100 real 3D scans, effectively solving the data scarcity bottleneck via a Blender-augmented synthetic data pipeline.
- We design a unified diffusion-based texture pipeline featuring a novel Cross-Intrinsic Attention mechanism, which enforces structural coherence and physical plausibility across generated material modalities.
- Leveraging generative priors and UV-space physical constraints, our method effectively avoids the pitfalls of direct 2D supervision, producing 4K-resolution PBR assets robust to occlusions and generalizing well to in-the-wild inputs.

2 Related Works

2.1 3D Geometry Reconstruction

Monocular 3D face geometry reconstruction remains a fundamental challenge [52]. The dominant paradigm relies on parametric representations, utilizing 3DMMs [4,

5, 34, 53] or blendshapes [35] to regress coefficients via deep learning [3, 10, 14, 17, 21, 43, 55, 63]. While semantically controllable, these methods are inherently constrained by the linear subspace, failing to capture high-frequency details. To address this, HRN [31] proposes a hierarchical iterative refinement strategy but suffers from optimization instability (e.g., overfitting artifacts). Similarly, HiFace [8] attempts to enhance expressiveness by expanding the high-fidelity 3DMM basis, yet it still struggles to resolve intricate micro-structures like wrinkles and pores. The second category, non-linear reconstruction models, endeavors to surmount the inherent limitations of linear models in representing high-frequency geometric details. Instead of relying on fixed linear bases, these approaches employ DNNs, GNNs [1, 16, 41, 49, 50], or Implicit Neural Representations (INRs) [11, 60, 61] to learn complex mappings directly from data. This enables the capture of subtle facial variations and intricate details like wrinkles, leading to higher fidelity reconstructions. However, they typically demand extensive, high-quality 3D training data, and the resulting geometry often lacks the explicit semantic disentanglement offered by parametric models. Targeting high-fidelity texture synthesis, we instead enhance a linear parametric model [4] with DINOv3 [44] semantic priors. This yields a geometric scaffold sufficiently accurate for extracting visible texture cues, effectively supporting our generative pipeline without the complexity of non-linear models.

2.2 Facial texture reconstruction

Monocular facial texture reconstruction has evolved from 3DMM-based parameter regression [14, 45] to high-fidelity generative approaches. The latter typically formulates texture generation as a UV-space completion problem. Early approaches utilized the editing and generative capabilities of Generative Adversarial Networks (GANs) [23–25] to hallucinate invisible regions from multi-view inputs [13, 19, 20]. More recently, latent diffusion models [42] have been introduced for this task; for instance, UV-IDM [32] and FreeUV [59] employ diffusion processes in latent space to generate high-quality texture details. But these completion-based methods inextricably bake the illumination into the texture, rendering the reconstructed assets unsuitable for relighting or downstream applications.

To achieve disentangled, physically-based reconstruction, recent research seeks to replace linear texture bases with high-quality non-linear bases generated by GANs or Diffusion models. These methods typically rely on large-scale synthetic datasets [2,9,40] or high-quality scans from Light Stage devices [16,18,27–29,38], employing differentiable rendering for optimization. While MoSAR [16] achieved complete texture reconstruction, its GAN-based backbone often suffers from mode collapse and over-smoothed results. Moreover, approaches relying on standard 2D differentiable rasterization often fail to explicitly handle occlusions, inadvertently baking artifacts from hair or glasses into the facial texture. NextFace [15] utilizes differentiable ray tracing to explicitly model self-occlusion for high-quality material disentanglement; however, compared to rasterization-based methods, it incurs prohibitive computational overhead and lacks generaliz-

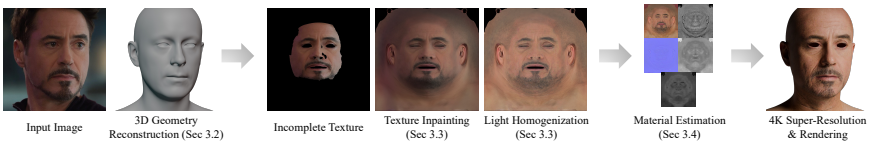


Fig. 2: Overview of our high-fidelity 3D avatar reconstruction pipeline. From a single image, we recover 3D geometry and an incomplete texture map, then repair and homogenize it for illumination. We subsequently predict PBR maps (albedo, roughness, specular, normal) and displacement, enabling photorealistic 4K rendering.

ability. UltraAvatar [62] attempts to fine-tune early versions of diffusion models on small datasets, yet the generation quality remains limited, and the model loses the capability to generalize to in-the-wild images.

Fundamentally, these methods are constrained by a data-efficiency dilemma: synthetic data often lacks realism, while real Light Stage data is expensive and difficult to scale. In contrast, we utilize a modern rendering engine [1] to augment limited real texture data. By leveraging the powerful generative priors of large-scale diffusion models [26, 46, 48, 57], we achieve high-fidelity generation of comprehensive material components with fewer than 100 real samples, seamlessly generalizing to in-the-wild images.

3 Methods

An overview of our method is illustrated in Fig. 2. Given a single in-the-wild input image, we first estimate the 3D face geometry and camera parameters, which are utilized to project the image into UV space to obtain an incomplete texture map (Sec. 3.2). To reconstruct a complete and canonicalized texture representation, we introduce a shared pre-trained diffusion backbone as the core component. Adapted via cascaded LoRA [22] modules, this backbone is flexibly applied to a sequence of sub-tasks: first, a texture inpainting stage completes missing regions caused by occlusion; this is followed by a light homogenization step that removes baked-in illumination to establish a unified lighting baseline (Sec. 3.3). Building on this, a multi-branch material estimation module utilizes the normalized texture to predict key intrinsic material parameters, including albedo, normal, roughness, specular, and displacement maps (Sec. 3.4). Finally, all predicted maps are upscaled to 4K resolution by a super-resolution network, ensuring high-fidelity rendering output (Sec. 3.5).

3.1 Dataset Preparation

Our training data are constructed by combining large-scale in-the-wild images with a small set of high-quality scanned PBR textures, following the pipeline illustrated in Fig. 3. We use FFHQ [24] and CelebAMask-HQ [30] as sources

[1] https://www.blender.org/

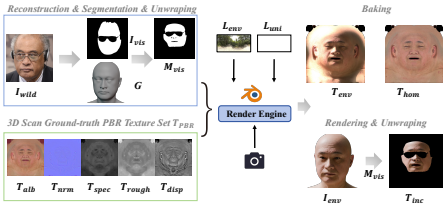


Fig. 3: Synthetic Data Generation Pipeline. Starting from an image I_{wild}, we reconstruct geometry $\mathbf{G}$ and compute visibility mask M_{vis}. We assign high-quality PBR textures T_{PBR} to $\mathbf{G}$.

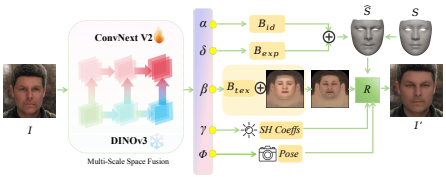


Fig. 4: Geometry estimation model. Combining ConvNeXt V2 and DINOv3, our encoder integrates local details with semantic priors for geometry estimation.

of in-the-wild facial images I_{wild}, which provide diverse geometry, pose, expression, illumination, and occlusion. These images are used to train the geometry reconstruction module and to derive realistic visibility masks M_{vis}. For material learning, we employ a compact scan dataset D_{scan} [2] containing fewer than 100 professional 3D face scans. Each scan provides a complete set of physically-based material maps, including albedo (T_{alb}), normal (T_{nrm}), roughness (T_{rough}), specular (T_{spec}), and displacement (T_{disp}), which together form the PBR texture set T_{PBR}. Rather than relying on appearance diversity, we treat these scans as physically accurate anchors. As shown in Fig. 3, we first reconstruct facial geometry $\mathbf{G}$ from I_{wild} and compute the visibility mask M_{vis} in UV space. The scanned PBR textures T_{PBR} are then assigned to $\mathbf{G}$ and rendered under diverse lighting L_{env} to produce shaded textures T_{env}, as well as under uniform lighting L_{uni} to obtain light-homogenized targets T_{hom}. To generate realistic incomplete inputs, we render images I_{env}, re-unwrap them into UV space, and multiply by M_{vis} to obtain T_{inc}. These paired data supervise texture inpainting, light homogenization, and intrinsic material estimation.

3.2 3D Geometry Reconstruction

Accurate 3D geometry forms the foundation of our high-fidelity avatar pipeline. Following standard paradigms [14], we regress the parameters of a 3D Morphable Model (3DMM) from a single image using deep convolutional networks. To enhance geometric fidelity beyond conventional linear models, we introduce a multi-scale encoder architecture together with a robust hybrid landmark supervision strategy.

We adopt the Hifi3D++ basis as our parametric face model. The parameter vector $\chi = \{\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$ represents identity $\boldsymbol{\alpha} \in \mathbb{R}^{532}$, expression $\boldsymbol{\delta} \in \mathbb{R}^{45}$, texture $\boldsymbol{\beta} \in \mathbb{R}^{439}$, Spherical Harmonic lighting $\boldsymbol{\gamma} \in \mathbb{R}^{9}$, and perspective camera

[2] https://www.3dscanstore.com.

pose $\boldsymbol{\phi}$. The 3D shape S and texture T are modeled as:

$$\begin{aligned} S(\boldsymbol{\alpha}, \boldsymbol{\delta}) &= \bar{S} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\delta}, \\ T(\boldsymbol{\beta}) &= \bar{T} + \mathbf{B}_{tex}\boldsymbol{\beta}, \end{aligned} \tag{1}$$

where $\bar{S}$ and $\bar{T}$ denote the mean shape and texture, and $\mathbf{B}$ are the corresponding PCA bases. The reconstructed image is obtained through a differentiable renderer $\mathcal{R}$ as $I' = \mathcal{R}(\chi)$.

To capture both fine-scale geometric details and high-level semantic structure, we propose a **Multi-Scale Space Fusion (MSSF) encoder** (Fig. 4). MSSF combines a trainable ConvNeXt V2 backbone [56] for local feature extraction with a frozen DINOv3 model [44] that provides strong self-supervised semantic priors. Multi-scale features from both branches are fused before parameter regression, significantly improving reconstruction accuracy. Despite using a linear decoder, this design enables our method to outperform nonlinear approaches such as MoSAR [16] in recovering subtle facial structures.

The network is trained using a self-supervised objective:

$$\mathcal{L} = \lambda_{pho}\mathcal{L}_{pho} + \lambda_{lan}\mathcal{L}_{lan} + \lambda_{per}\mathcal{L}_{per} + \lambda_{reg}\mathcal{L}_{reg},$$

where $\mathcal{L}_{pho}$ is a photometric loss weighted by a skin-attention mask to handle occlusions, $\mathcal{L}_{per}$ is a perceptual loss to improve visual fidelity, and $\mathcal{L}_{reg}$ regularizes the 3DMM coefficients. To address instability of standard 68-point landmark detectors around the mouth, we introduce an **Enhanced Landmark Loss** $\mathcal{L}_{lan}$ based on an 88-point hybrid configuration, which combines stable facial contour landmarks [6] with robust mouth keypoints from MediaPipe [37].

3.3 Texture Inpainting and Light Homogenization

Leveraging the paired data constructed in Sec. 3.1, we perform texture completion and light homogenization using a unified diffusion backbone operating in UV space, adapted via task-specific LoRA modules. Both tasks are formulated as conditional diffusion processes that share the same pre-trained transformer and differ only in their conditioning inputs and supervision targets (Fig. 5, right). During training, the backbone is kept frozen and only the LoRA parameters are optimized.

Texture Inpainting. Texture inpainting aims to recover a complete UV texture from an incomplete input T_{inc}. We formulate this task as conditional diffusion in latent space using the flow-matching objective. Given a noise level $\sigma \in (0, 1)$, the noisy latent is constructed as $z_\sigma = (1 - \sigma)\, z_0 + \sigma\, \epsilon$, where z_0 denotes the latent encoding of the ground-truth texture and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The model predicts the target velocity field $(\epsilon - z_0)$ by minimizing the flow-matching loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\sigma,\,\epsilon}\left[\left\| f_{\text{inp}}(z_\sigma, T_{\text{inc}}, \sigma) - (\epsilon - z_0)\right\|_2^2\right]. \tag{2}$$

The incomplete texture T_{inc} is provided as conditioning, enabling the model to infer missing regions while preserving visible content.

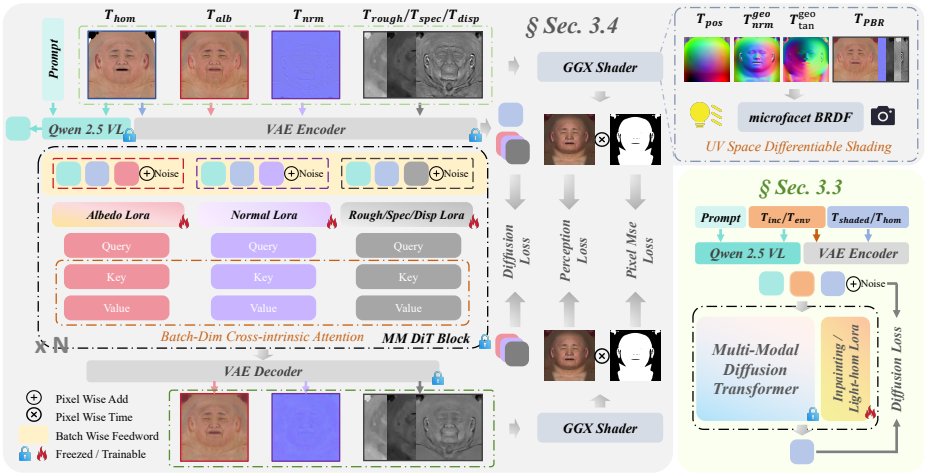


Fig. 5: Training pipeline for texture generation. A unified diffusion transformer uses task-specific LoRA modules for UV-space texture inpainting, light homogenization, and PBR material estimation. Inpainting and homogenization (Sec. 3.3) use paired supervision. For material estimation (Sec. 3.4), the homogenized texture T_{hom} feeds three parallel LoRA branches (albedo, normal, packed) interacting via cross-intrinsic attention to generate PBR maps, supervised by a differentiable GGX BRDF shader.

Light Homogenization. Although texture completion yields visually complete UV maps, illumination remains scene-dependent. To enable intrinsic material estimation, we introduce a light homogenization stage that maps shaded textures into a canonical lighting domain. Using the same diffusion backbone and flow-matching formulation, the conditioning input and supervision target differ: the model is conditioned on the shaded texture T_{env} and supervised by its uniformly illuminated counterpart T_{hom}. Normalizing textures into a consistent lighting space simplifies and stabilizes subsequent material estimation.

3.4 Physically-based Material Estimation

This stage estimates physically-based material properties from the light-homogenized texture T_{hom}. Our goal is to recover a complete PBR representation in UV space, including albedo T_{alb}, normal T_{nrm}, and a compact reflectance map T_{rsd} that packs roughness T_{rough}, specular T_{spec}, and displacement T_{disp}. A naïve approach is to train independent diffusion adapters for each material attribute. While the semantic priors from the pre-trained backbone are effectively leveraged, independently generated material maps often lack physical consistency. Specifically, independently trained branches tend to overfit local high-frequency pixel variations, misinterpreting appearance cues as geometric structures, which leads to fragmented artifacts and spatially incoherent micro-structures.

Joint Diffusion with Cross-Intrinsic Attention. To enforce physical coherence across material modalities, we adopt a joint diffusion strategy based on *cross-intrinsic attention*. We introduce three task-specific LoRA adapters, $\Delta\theta_{\text{alb}}$, $\Delta\theta_{\text{nrm}}$, and $\Delta\theta_{\text{rsd}}$, corresponding to albedo, normal, and reflectance-related properties, while sharing a common diffusion backbone. During training, the three material modalities are processed jointly by stacking them along the batch dimension. Within each transformer block, standard self-attention is replaced by a cross-intrinsic attention mechanism that enables explicit information exchange across modalities. Queries are computed independently within each modality branch, while keys and values are formed by concatenating features from all material branches. Let $\mathbf{K} = [\mathbf{k}_{\text{alb}}, \mathbf{k}_{\text{nrm}}, \mathbf{k}_{\text{rsd}}], \quad \mathbf{V} = [\mathbf{v}_{\text{alb}}, \mathbf{v}_{\text{nrm}}, \mathbf{v}_{\text{rsd}}]$ denote the concatenated keys and values across modalities. Given query features $\mathbf{q}_i$ from modality i, cross-intrinsic attention outputs the updated token features $\mathbf{h}_i$ as

$$\mathbf{h}_i = \text{Attn}(\mathbf{q}_i, \mathbf{K}, \mathbf{V}), \text{Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d}}\right)\mathbf{V}. \tag{3}$$

This design allows each material branch to directly attend to complementary geometric and reflectance cues from the others, encouraging spatially aligned fine-scale structures while preserving domain-specific adaptation through separate LoRA parameters.

UV-Space Differentiable Shader. Joint diffusion enforces spatial alignment but does not guarantee physically meaningful decomposition. We therefore introduce a UV-space differentiable BRDF shading loss that explicitly couples material prediction with image formation. During training, the clean latent estimate $\hat{\mathbf{z}}_0$ is reconstructed from the noisy state $\mathbf{z}_\sigma$ and the predicted velocity field $\hat{\mathbf{v}}_\theta$ (Eq. 2) as:

$$\hat{\mathbf{z}}_0 = \mathbf{z}_\sigma - \sigma\, \hat{\mathbf{v}}_\theta(\mathbf{z}_\sigma, \sigma, \mathbf{c}). \tag{4}$$

Decoding $\hat{\mathbf{z}}_0$ through the VAE yields the predicted material maps $\hat{T}_{\text{alb}}$, $\hat{T}_{\text{nrm}}$, and $\hat{T}_{\text{rsd}}$. To ensure stable and physically grounded supervision, we decouple texture decoding from subject-specific geometry. A fixed template face is used, for which we precompute its world-space position $\tilde{T}_{\text{pos}}$, geometric normal $\tilde{T}^{\text{geo}}_{\text{nrm}}$, and geometric tangent $\tilde{T}^{\text{geo}}_{\text{tan}}$. Given these geometric priors, shading is evaluated using a GGX microfacet shader:

$$\hat{T}_{\text{shaded}} = \mathcal{S}_{\text{GGX}}\Big(\hat{T}_{\text{alb}}, \hat{T}_{\text{nrm}}, \hat{T}_{\text{rough}}, \hat{T}_{\text{spec}}, \hat{T}_{\text{disp}}; \tilde{T}_{\text{pos}}, \tilde{T}^{\text{geo}}_{\text{nrm}}, \tilde{T}^{\text{geo}}_{\text{tan}}, L, V\Big), \tag{5}$$

where L and V denote sampled directional light and view directions. Implementation details of $\mathcal{S}_{\text{GGX}}$ are provided in the Appendix A.1. During training, camera viewpoints are sampled using a hybrid strategy: with probability 0.5, views are sampled uniformly along a 360° azimuth at a fixed radius with random vertical perturbation; otherwise, a near-frontal view with small spatial jitter is used. Directional lighting is initialized from the camera direction and randomly perturbed, and a low-intensity uniform ambient term sampled from $[0.15, 0.3]$ is added to prevent degenerate shadows.

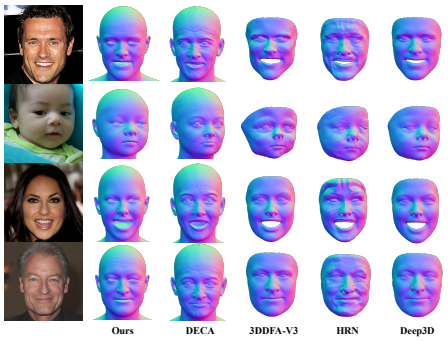


Fig. 6: Comparison of 3D face reconstruction methods.

Table 1: Reconstruction errors (NMSE) on the REALY benchmark. Lower is better.

Method	Error (mm)				All
	Nose	Mouth	Forehead	Cheeks	
DECA	1.697 ± 0.355	2.516 ± 0.839	2.394 ± 0.576	1.479 ± 0.535	2.010
Deep3D	1.719 ± 0.354	1.368 ± 0.439	2.015 ± 0.449	1.528 ± 0.501	1.657
HRN	1.722 ± 0.330	1.357 ± 0.523	1.995 ± 0.476	1.072 ± 0.333	1.537
HiFace (w/o syn)	1.227 ± 0.407	1.787 ± 0.439	1.454 ± 0.382	1.762 ± 0.436	1.558
MoSAR	1.499 ± 0.366	1.424 ± 0.462	1.950 ± 0.559	1.128 ± 0.303	1.500
3DDFA-V3	1.584 ± 0.308	1.237 ± 0.375	1.809 ± 0.394	1.110 ± 0.328	1.435
HiFace	1.036 ± 0.280	1.450 ± 0.413	1.324 ± 0.334	1.291 ± 0.362	1.275
Ours	1.619 ± 0.381	1.376 ± 0.477	1.784 ± 0.464	1.181 ± 0.402	1.490
w/o Enhanced $\mathcal{L}_{lan}$	1.457 ± 0.341	1.670 ± 0.529	1.803 ± 0.453	1.140 ± 0.405	1.518
w/o MSSF	1.728 ± 0.441	1.785 ± 0.514	2.179 ± 0.610	1.530 ± 0.548	1.806
Baseline	2.592 ± 0.510	2.664 ± 0.528	2.273 ± 0.563	2.524 ± 0.474	2.514

Training Objective. The material estimation network is optimized using a hybrid objective that combines diffusion supervision with rendering-based constraints:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{img}} \|\hat{T}_{\text{shaded}} - T_{\text{shaded}}\|_2^2 + \lambda_{\text{lpips}} \,\text{LPIPS}(\hat{T}_{\text{shaded}}, T_{\text{shaded}}). \quad (6)$$

By coupling flow-matching diffusion with physically grounded re-rendering losses, the model is guided to produce material maps that exhibit consistent physical behavior under novel illumination.

3.5 Super-Resolution

As a final step, we apply super-resolution to obtain 4K UV texture maps. We treat super-resolution as a post-processing module and fine-tune a Real-ESRGAN [54] model to upscale all predicted texture maps from 1K to 4K.

4 Experiments

Implementation Details. For geometry reconstruction, we train a ConvNeXt V2 Base backbone augmented with a frozen DINOv3 encoder for 50 epochs using a batch size of 128. For texture generation, we adopt LongCat-Image-Edit [47] as the unified diffusion transformer (DiT) backbone. LoRA adapters (rank 32) are injected into the attention and feedforward projection layers of the DiT. Each adapter contains approximately 91M parameters, accounting for only 0.7% of the backbone. We use JoyCaption[3] to generate captions for ground-truth data at each training stage, while inference combines captions from the input image with fixed templates for intermediate tasks. All models are trained on 8 NVIDIA H100 80GB GPUs using BF16 precision and the AdamW optimizer [36] with a learning rate of 1×10^{-4}. The inpainting and light-homogenization LoRAs are trained for 20k steps with a batch size of 64. The intrinsic material LoRAs follow

[3] `https://github.com/fpgaminer/joycaption`

a two-stage schedule, consisting of 10k steps of independent training followed by 10k steps of joint training, using a batch size of 8 and loss weights $\lambda_{img} = 0.5$ and $\lambda_{lpips} = 0.1$.

4.1 Geometry Reconstruction

Quantitative and qualitative analysis. We evaluate our geometry reconstruction on the REALY benchmark [7], which provides high-quality 3D scans and images of 100 subjects. Following the standard protocol, we report the Normalized Mean Squared Error (NMSE) between reconstructed and ground-truth meshes over four facial regions: nose, mouth, forehead, and cheeks. Quantitative results are shown in Tab. 1. Our method achieves the third-best overall performance with an NMSE of 1.490. Notably, despite relying on a linear 3DMM representation, our approach outperforms MoSAR [16], which adopts a more expressive nonlinear geometry model.

While our method trails slightly behind 3DDFA-V3 [55] and HiFace [8] numerically, it is important to note that HiFace heavily relies on large-scale synthetic images and real 3D meshes for training. Under comparable settings that restrict training to in-the-wild images, HiFace exhibits inferior performance compared to our method. Similarly, although 3DDFA-V3 achieves marginally lower errors using segmentation-based landmark clustering, qualitative inspection reveals a tendency to overfit the input image. As shown in Fig. 6, 3DDFA-V3 produces unstable geometric bumps on smooth regions and fails to recover fine wrinkle details. We further compare our method with DECA [17], HRN [31], and Deep3D [14]. Benefiting from the estimated displacement and normal maps, our method faithfully reconstructs high frequency facial details, such as forehead wrinkles, eye region creases, and smile lines. In contrast, HRN suffers from severe overfitting artifacts, Deep3D lacks fine-scale details, and DECA often generates wrinkles that are misaligned with the input image.

Ablation Study. As shown in Tab. 1 bottom, adding the enhanced $\mathcal{L}_{lan}$ to a ConvNeXt baseline improves mouth-region accuracy. Furthermore, our MSSF encoder achieves comprehensive improvements across all metrics, underscoring the critical role of DINOv3's semantic priors.

4.2 Texture Reconstruction

Texture Inpainting and Light Homogenization Fig. 7 and Tab. 2 demonstrate the effectiveness of our pipeline in handling occlusions and illumination. For texture inpainting (middle block), we compare against UV-IDM [32] and HRN [31]. Due to its reliance on an overfitting projection loss without explicit occlusion modeling, HRN tends to bake external objects into the UV texture. As shown in the second and fourth rows, hair strands and hats are erroneously preserved, and large pose-induced self-occlusions are poorly handled. UV-IDM avoids projection artifacts but struggles to preserve high-frequency identity details, exhibiting noticeable spotting artifacts (e.g., at the temples in the first row). In contrast, our

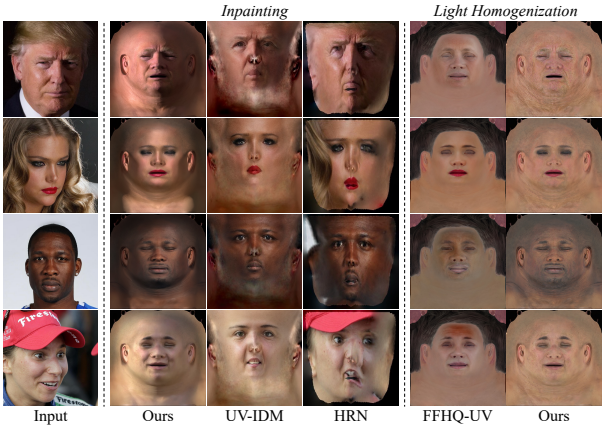


Fig. 7: Texture inpainting and light homogenization. Left: inputs. Middle: HRN shows baked-in occlusions while UV-IDM lacks fine detail; our method removes occlusions and restores high-frequency detail. Right: unlike FFHQ-UV, our homogenization yields uniformly lit, detail-rich textures for intrinsic decomposition.

Table 2: Quantitative evaluation of texture inpainting and light homogenization.

(a) Texture Inpainting

Method	PSNR↑	SSIM↑	LPIPS↓	CSIM↑
UV-IDM	19.39	0.6149	0.1674	0.269
HRN	15.70	0.6193	0.2982	0.441
Ours	**22.44**	**0.8003**	**0.0621**	**0.540**

(b) Light Homogenization

Method	CSIM↑	BS↓
FFHQ-UV	0.1340	5.738
Ours	**0.4667**	**3.963**

inpainting module produces clean and realistic skin textures in occluded regions while preserving identity-specific details. Moreover, the resulting illumination appears more volumetric, benefiting from our integration of physically based rendering into the data generation process.

For light homogenization (right block), we compare against FFHQ-UV [2]. As shown in the third row, FFHQ-UV retains scene-dependent illumination, including specular highlights on the forehead and shadows around facial features, and is further affected by occlusions (e.g., the red artifact in the fourth row). Our method effectively suppresses such illumination and occlusion artifacts, producing flat, uniformly lit textures that preserve fine appearance cues, including eyebrows, makeup, facial redness (rows 1, 2, and 4), as well as skin tone and beard structure (row 3). This provides a robust and consistent input for subsequent PBR material estimation.

Material estimation. We compare our method with state-of-the-art 3D avatar reconstruction methods: MoSAR [16], FitMe [27], and Relightify [38]. Since their official implementations are unavailable, we utilize test cases from [16] for a fair visual comparison.

As shown in Figures 8 and 9, our pipeline demonstrates superiority in three key aspects. First, regarding fidelity and skin tone (Fig. 8), MoSAR suffers from GAN-induced waxy over-smoothing and a systematic bias towards pale skin, fail-

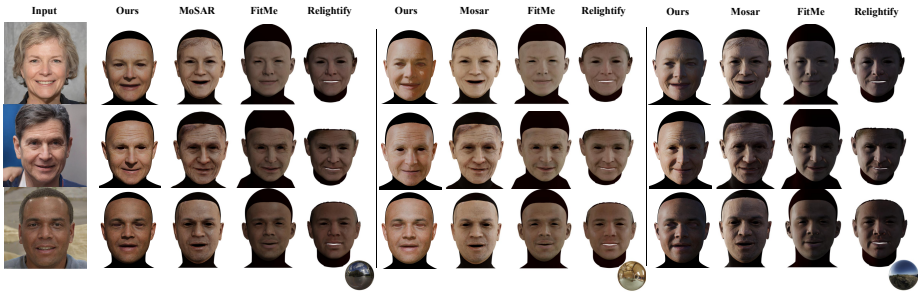


Fig. 8: Qualitative comparison of avatar relighting under novel environments, against MoSAR, FitMe and Relightify. MoSAR yields waxy, desaturated skin and visible artifacts; FitMe and Relightify show poor albedo estimates and often bake hair into textures. Our method better disentangles lighting and occlusion, producing photorealistic, detailed avatars with accurate skin tones.

ing to recover the dark pigmentation in Row 3. Similarly, FitMe and Relightify exhibit unnatural albedo estimation. In contrast, our method faithfully preserves authentic skin texture and intrinsic coloration. Second, concerning occlusion robustness, rasterization-based methods like MoSAR struggle to separate skin from occluders, often baking hair into the texture (forehead in Rows 1-2) or generating artifacts (left temple in Row 3). Our UV-space physical constraints and generative priors effectively synthesize clean, artifact-free skin in occluded areas. Finally, our approach even outperforms the closed-source commercial ChatAvatar, providing superior geometric details and material completeness—as seen in the beard in the first row, as well as the facial spots and wrinkle details in the second and third rows.

Ablation Study. We first examine the necessity of the light homogenization stage in Fig. 10a. The top row shows results from the full pipeline, while the bottom row removes light homogenization and performs material estimation directly on the shaded inpainted texture. From a learning perspective, light homogenization acts as a crucial normalization step that significantly constrains the solution space. Without it, the network faces a much harder one-to-many inverse rendering problem, requiring simultaneous material hallucination and implicit decomposition of unknown, complex lighting from a single texture map. As shown in the bottom row, this expanded solution space leads to severe optimization instability: the model fails to recover valid material properties and instead converges to degenerate solutions, where high-contrast shading is baked into the albedo and misinterpreted as high-frequency geometric noise in the normal and displacement maps. These results confirm that explicit lighting normalization is essential for stabilizing training and achieving physically disentangled reconstruction [33].

We further analyze the qualitative impact of our multi-branch LoRA-based joint material estimation strategy in Fig. 10b. A naive approach using separate LoRAs leads to modality-specific overfitting: as shown in the *Separate* column,

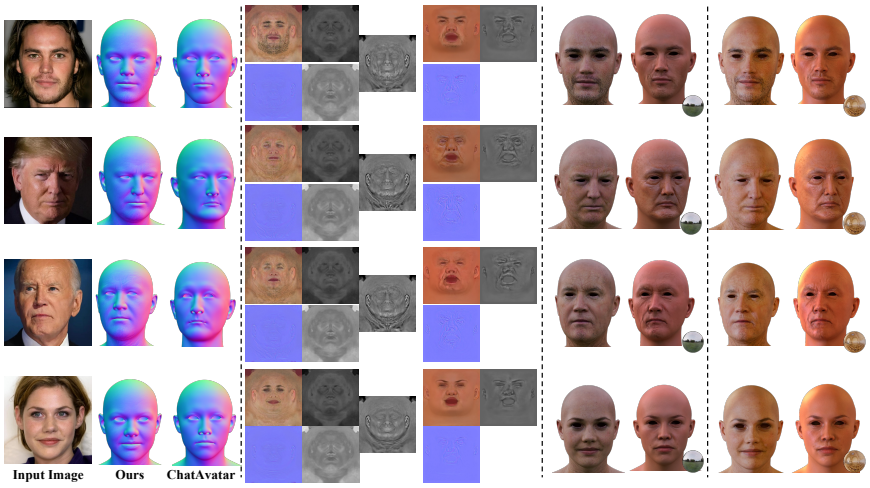
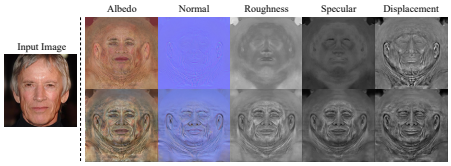


Fig. 9: Comparison with ChatAvatar. Geometry (Left): our method recovers high-frequency details (wrinkles, nasolabial folds) while the baseline over-smooths. **Materials (Middle):** the baseline provides Albedo/Normal/Specular only; we predict full PBR including Roughness and Displacement. **Relighting (Right):** under novel HDRIs, our relighting better preserves identity and photorealism.

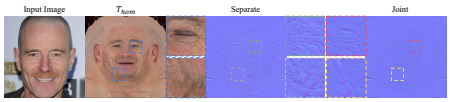
the network over-interprets high-frequency signals as geometric deformation, producing severely noisy normal maps. In particular, stubble and skin pores are reconstructed as sharp geometric spikes (gray box), while wrinkles appear as fragmented scratches rather than continuous folds (green box). In contrast, our joint strategy combines Cross-Intrinsic Attention with a UV-space differentiable BRDF shading loss, enabling explicit inter-modal coupling and enforcing physical consistency. This constraint suppresses geometry hallucination from surface pigmentation and restricts geometric variation to shading-consistent structures. As shown in the *Joint* column, this yields structurally coherent geometry, with continuous crow's feet (red box) and smooth skin regions (yellow box).

5 Conclusions

In this work, we present a novel framework for high-fidelity 3D avatar reconstruction from a single in-the-wild image. By synergizing modern rendering workflows with the robust priors of large-scale diffusion models, our approach effectively overcomes the data scarcity and generalization bottlenecks inherent in this task. Building upon a robust geometric alignment enhanced by DINOv3 priors, we introduce a unified diffusion-based pipeline for texture generation, facilitated by cascaded LoRA modules. This pipeline, through texture restoration, illumination homogenization, and intrinsic material estimation guided by a Cross-Intrinsic Attention mechanism combined with UV-space differentiable BRDF shading loss, successfully disentangles intrinsic materials from complex illumination without



(a) Ablation on light homogenization. Top (Ours): Homogenization ensures fast convergence to clean, physically disentangled materials. **Bottom (w/o):** Without it, the solution space expands and the texture→material mapping becomes ill-posed; even after **40k steps**, the model fails to reach a physically plausible result. Shadows are mistaken for dark albedo (burnt artifacts) and geometric noise (noisy normals), reflecting sensitivity to lighting variance.



(b) Visual ablation on material estimation. Normals from independent LoRAs (Separate) vs our multi-branch (Joint). **Separate:** Lacking physical and cross-modal constraints, it overfits high-frequency detail—mistaking skin texture and stubble for geometry, causing fragmented crow's feet and grainy cheeks. **Joint:** With Cross-Intrinsic Attention and a differentiable BRDF loss, we separate pigmentation from depth, producing coherent wrinkles and smooth, physically plausible skin while suppressing non-geometric noise.

Fig. 10: Texture Reconstruction ablation study.

relying on screen-space supervision. Trained on fewer than 100 real 3D scans, our framework achieves state-of-the-art performance in recovering photorealistic, relightable, and geometrically consistent avatars.

Limitations. Our method targets relightable face and head PBR assets; it does not reconstruct full bodies, dynamic hair, clothing, or accessories. Treating hair, glasses, and hands as occlusions improves face-centric material recovery, but may lose identity-specific details behind occluders and fail under extreme expressions. Future work will explore complete heads, stronger nonlinear or 3DGS-based representations, larger open PBR datasets, and bias-aware evaluation.

Social impact. Our method lowers the barrier to generating high-quality 3D facial avatars from a single image. Intended for authorized asset creation in film, games, virtual production, and telepresence, it could also enable non-consensual digital doubles, deepfake-ready identities, or impersonation assets from unauthorized images, harming privacy and identity security. It should therefore be used only with authorized images and valid usage rights. For released code and models, usage terms will prohibit impersonation, non-consensual identity replication, and other identity misuse; watermarking, metadata disclosure, and forgery-detection compatibility are feasible downstream mitigation options.

6 Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2023YFC3306401), National Natural Science Foundation of China (No. 62576018), Zhejiang Provincial Natural Science Foundation of China (No. LD24F020007), Beijing Natural Science Foundation (No. L244043), and the Ministry of Economic Development of the Russian Federation (agreement 000000C313925P3X0002, grant No. 139-15-2025-004 dated 17.04.2025). We also thank Great Wall Motors for their support.

References

1. Aliari, M.A., Beauchamp, A., Popa, T., Paquette, E.: Face editing using part-based optimization of the latent space. In: CGF (2023)
2. Bai, H., Kang, D., Zhang, H., Pan, J., Bao, L.: Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In: CVPR (2023)
3. Bai, Z., Cui, Z., Liu, X., Tan, P.: Riggable 3d face reconstruction via in-network optimization. In: CVPR (2021)
4. Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J., Yu, D., Zhang, Z.: High-fidelity 3d digital human head creation from rgb-d selfies. TOG (2021)
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
6. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: ICCV (2017)
7. Chai, Z., Zhang, H., Ren, J., Kang, D., Xu, Z., Zhe, X., Yuan, C., Bao, L.: Realy: Rethinking the evaluation of 3d face reconstruction. In: ECCV (2022)
8. Chai, Z., Zhang, T., He, T., Tan, X., Baltrusaitis, T., Wu, H., Li, R., Zhao, S., Yuan, C., Bian, J.: Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In: ICCV (2023)
9. Dai, J., Wang, A., Ni, B., Cao, T.: High-quality facial albedo generation for 3d face reconstruction from a single image using a coarse-to-fine approach. arXiv:2506.13233 (2025)
10. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: CVPR (2022)
11. De Luigi, L., Cardace, A., Spezialetti, R., Zama Ramirez, P., Salti, S., Di Stefano, L.: Deep learning on implicit neural representations of shapes. In: ICLR (2023)
12. Debevec, P.: The light stages and their applications to photoreal digital actors. In: ACM SIGGRAPH 2012 Courses. pp. 1–10 (2012)
13. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S.: Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In: CVPR (2018)
14. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPRW (2019)
15. Dib, A., Bharaj, G., Ahn, J., Thébault, C., Gosselin, P.H., Romeo, M., Chevallier, L.: Practical face reconstruction via differentiable ray tracing. EG (2021)
16. Dib, A., Hafemann, L.G., Got, E., Anderson, T., Fadaeinejad, A., Cruz, R.M., Carbonneau, M.A.: Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In: CVPR (2024)
17. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. In: SIGGRAPH (2021)
18. Galanakis, S., Lattas, A., Moschoglou, S., Zafeiriou, S.: Fitdiff: Robust monocular 3d facial shape and reflectance estimation using diffusion models. In: WACV (2025)
19. Gecer, B., Deng, J., Zafeiriou, S.: Ostec: One-shot texture completion. In: CVPR (2021)
20. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: CVPR (2019)
21. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: ECCV (2020)

22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
26. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv:2506.15742 (2025)
27. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Deng, J., Zafeiriou, S.: Fitme: Deep photorealistic 3d morphable model avatars. In: CVPR (2023)
28. Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In: CVPR (2020)
29. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., Zafeiriou, S.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. PAMI (2022)
30. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
31. Lei, B., Ren, J., Feng, M., Cui, M., Xie, X.: A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In: CVPR (2023)
32. Li, H., Feng, Y., Xue, S., Liu, X., Zeng, B., Li, S., Liu, B., Liu, J., Han, S., Zhang, B.: Uv-idm: identity-conditioned latent diffusion model for face uv-texture generation. In: CVPR (2024)
33. Li, H., Ye, C., Chen, H., Xiao, W., Yan, Z., Xiao, L., Chen, Z., Xiang, J., Xu, S., Liu, X., Wang, Y., Zhang, B., Han, X., Yang, J., Zhao, H.: Near: Coupled neural asset-renderer stack. arXiv:2511.18600 (2025)
34. Li, R., Bladin, K., Zhao, Y., Chinara, C., Ingraham, O., Xiang, P., Ren, X., Prasad, P., Kishore, B., Xing, J., Li, H.: Learning formation of physically-based face attributes. In: CVPR (2020)
35. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. SIGGRAPH Asia (2017)
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: NeurIPS (2017)
37. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv:1906.08172 (2019)
38. Papantoniou, F., Lattas, A., Moschoglou, S., Zafeiriou, S.: Relightify: Relightable 3d faces from a single image via diffusion models. In: CVPR (2023)
39. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: AVSS (2009)
40. Rai, A., Gupta, H., Pandey, A., Carrasco, F.V., Takagi, S.J., Aubel, A., Kim, D., Prakash, A., De la Torre, F.: Towards realistic generative 3d face models. In: WACV (2024)
41. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: ECCV (2018)

42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
43. Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: CVPR (2019)
44. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv:2508.10104 (2025)
45. Smith, W.A.P., Seck, A., Dee, H., Tiddeman, B., Tenenbaum, J., Egger, B.: A morphable face albedo model. In: CVPR (2020)
46. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., Cai, X., Guan, Y., Hu, J.: Longcat-image technical report. arXiv:2512.07584 (2025)
47. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., et al.: Longcat-image technical report. arXiv:2512.07584 (2025)
48. Team, Z.I.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv:2511.22699 (2025)
49. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: CVPR (2019)
50. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR (2018)
51. Wang, C., Kang, D., Sun, H., Qian, S., Wang, Z., Bao, L., Zhang, S.H.: Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In: CVPR (2025)
52. Wang, H.: A review of 3d face reconstruction from a single image. arXiv:2110.09299 (2021)
53. Wang, L., Chen, Z., Yu, T., Ma, C., Li, L., Liu, Y.: Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In: CVPR (2022)
54. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW (2021)
55. Wang, Z., Zhu, X., Zhang, T., Wang, B., Lei, Z.: 3d face reconstruction with the geometric guidance of facial part segmentation. In: CVPR (2024)
56. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: CVPR (2023)
57. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv:2508.02324 (2025)
58. Wu, Y., Wu, Y., Li, W., Lu, Y., Feng, K., Chen, X.: Fastavatar: Towards unified fast high-fidelity 3d avatar reconstruction with large gaussian reconstruction transformers. arXiv:2508.19754 (2025)
59. Yang, X., Taketomi, T., Endo, Y., Kanamori, Y.: Freeuv: Ground-truth-free realistic facial uv texture recovery via cross-assembly inference strategy. In: CVPR (2025)
60. Zheng, M., Yang, H., Huang, D., Chen, L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20343–20352 (2022)
61. Zheng, M., Zhang, H., Yang, H., Chen, L., Huang, D.: Imface++: A sophisticated nonlinear 3d morphable face model with implicit neural representations. PAMI (2025)

62. Zhou, M., Hyder, R., Xuan, Z., Qi, G.: Ultravatar: A realistic animatable 3d avatar diffusion model with authenticity guided textures. In: CVPR (2024)
63. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European Conference on Computer Vision (ECCV) (2022)