High-Quality Full-Head 3D Avatar Generation from Any Single Portrait Image

Yujie Gao[1], Chencheng Wang[1], Xianbing Sun[1], Jiahui Zhan[1], Wentao Wang[2], Yiyi Zhang[1], Haohua Zhao[1], Liqing Zhang[1*], Jianfu Zhang[1*]

[1]Shanghai Jiao Tong University,
[2]Shanghai AI Laboratory
{GYjgyj123, nyte_plus, fufengsjtu, jiahuizhan, yi95yi, haoh.zhao, zhang-lq, c.sis}@sjtu.edu.cn;
wangwentao@pjlab.org.cn;

Abstract

In this work, we introduce a novel high-fidelity full-head 3D avatar generation method from a single image, regardless of perspective, style, expression, or accessories. Prior works often fail to preserve consistent head geometry and facial details, primarily due to their limited capacity in modeling fine-grained facial textures and maintaining identity information. To address these challenges, we construct **a new high-quality dataset** containing 227 sequences of digital human portraits captured from 96 different perspectives, totaling 21,792 frames, featuring high-quality facial texture details. To further improve performance, we propose a novel multi-view diffusion model named **ID-TS diffusion model**, which integrates identity and expression information into the two-stage multi-view diffusion process. The low-resolution stage ensures structural consistency of heads across multiple views, while the high-resolution stage preserves facial detail fidelity and coherence. Finally, we propose **an enhanced feed-forward Gaussian avatar reconstruction method** that optimizes the network on multi-view images of each single subject, significantly improving 3D facial texture details. Extensive experiments demonstrate that our method achieves robust performance across challenging scenarios, while showcasing broad applicability across numerous downstream tasks.

1 Introduction

Synthesizing full-head 3D models with arbitrary expressions has widespread applications in game design, AR/VR, video conferencing, *etc*. Traditional approaches rely on statistical 3D face models (Blanz and Vetter 1999; Feng et al. 2021; Wang et al. 2024) to generate high-fidelity 3D heads. However, they are typically limited to the facial region and often fail to reconstruct the full head, missing important features such as hair, glasses, and accessories. Other works (Deng, Wang, and Wang 2024; Chu et al. 2024; Chu and Harada 2024; Ye et al. 2024; Taubner et al. 2025) have proposed methods for reconstructing 4D dynamic head models from single portrait images, achieving promising results. Nonetheless, these approaches share the same limitation that their focus remains primarily on the facial area, resulting in incomplete full-head reconstructions. This significantly constrains their applicability in downstream tasks.

In recent years, numerous single-image-to-3D reconstruction methods (Long et al. 2024; Liu et al. 2023; Shi et al. 2023; Wang and Shi 2023; Xu et al. 2024) based on multi-view diffusion models have achieved remarkable success, which inspires us to explore their potential for 3D human head generation. While these models are trained on large-scale 3D datasets to generate multi-view images for 3D object reconstruction, they often lack sufficient geometric and textural consistency when the input is a human face. This discrepancy becomes particularly problematic because human faces require not only geometric consistency but also the preservation of subtle features such as expression, identity, and accessories. Upon reviewing existing image-to-3D methodologies, the following primary limitations can be identified: 1) Most image-to-3D diffusion models are trained on large-scale 3D datasets (Deitke et al. 2023b,a), which often lack high-quality head data and suffer from a limited number of meshes, resulting in poor quality reconstructions. Besides, existing facial datasets (Martinez et al. 2024; Pan et al. 2023; Kirschstein et al. 2023; Bühler et al. 2024) solely focus on frontal facial regions while neglecting critical side and rear head perspectives for full-head reconstruction. 2) Existing image-to-3D diffusion models still struggle to maintain identity and expression consistency for human head generation. This is particularly challenging because human faces require much higher consistency compared to other objects. 3) Prior works struggle to ensure consistent facial feature details across multi-view portrait images, *e.g.*, the human eyes may remain highly sensitive to subtle discrepancies among different views. 4) Previous multi-view diffusion models employed suboptimal 3D reconstruction approaches that failed to preserve high-fidelity texture details in human head modeling.

To address these challenges, we enhance the existing image-to-3D methodology with several key innovations. First, we construct a **3D digital human head dataset** with diverse attributes (*e.g.*, hairstyle, skin color, age, gender, accessory, expression, *etc*.) to fine-tune the multi-view diffusion model and 3D reconstruction model. This dataset is constructed using a high-quality digital human head engine, offering rich facial texture details and enabling the simulation of complex texture variations during facial expression

*Corresponding authors.

Figure 1: Given a single head image, our method can generate multi-view images with rich facial details. And the high-fidelity 3D Gaussian head of the identity is then reconstructed using the multi-view images. Our method also generalizes well to multiple downstream tasks, such as frontal face prediction, stylized head generation, and facial expression animation.

changes. Equally crucial is our ability to freely configure camera parameters, which facilitates the acquisition of appropriate multi-view images for training purposes. Second, we propose an **identity-guided two-stage diffusion model**, called ID-TS diffusion model, which is specifically designed for high-fidelity human head generation. To address the limitation of existing methods in preserving facial identity and expression information, we propose the identity-aware diffusion framework that incorporates ID embeddings to guide the generation process, significantly enhancing the model's capability in identity preservation. Toward consistent and high-fidelity facial details, we adopt a two-stage diffusion process: the low-resolution stage addresses shape consistency across multiple viewpoints of the head, while the high-resolution stage refines facial details by integrating outputs from the low-resolution stage, thereby further improving the coherence of facial features. Finally, we employ an **enhanced feed-forward Gaussian avatar reconstruction method** for high-quality 3D human head reconstruction, while jointly optimizing the feed-forward network. Specifically, in order to enhance the detail preservation and spatial coherence of 3D Gaussian heads, we fine-tune the network with our dataset and optimize each subject with the 16-frame multi-view outputs from ID-TS diffusion model. Experimental results demonstrate that our method can generate vivid 3D human heads, even under challenging conditions such as profile-view inputs, stylized head models, and complex facial expressions.

In summary, the contributions of our work are as follows:

- We propose a high-quality human head dataset consisting of 227 sequences of digital human portraits captured from 96 different perspectives, totaling 21,792 frames.
- We introduce an identity-guided two-stage diffusion model (ID-TS diffusion model), which significantly improves multi-view identity consistency and expression preservation while achieving finer facial detail synthesis.
- We present an enhanced feed-forward Gaussian avatar reconstruction method that significantly improves the fidelity of 3D human head details.

2 Related Works

3D Full-Head Generation: 3D reconstruction of human head from a single image has been a hot topic for an extended period, resulting in numerous high-quality contributions to the field. PanoHead (An et al. 2023) builds upon the EG3D (Chan et al. 2022) and achieves full head synthesis based on GAN (Goodfellow et al. 2014). However, since its triplane representation is implicit, the expressive capability of their model is somewhat limited. ID-Sculpt (Hao et al. 2025) uses ID-aware guidance for 3D head reconstruction from a single image and achieves relatively favorable results; but, its reliance on the head shape prior provided by PanoHead may lead to potential shape artifacts. Rodin (Wang et al. 2023) and RodinHD (Zhang et al. 2024a) incorporate a combination of the diffusion model and the triplane representation. However, their methods fail to generate fine details and exhibit lower similarity with the input image. FaceLift (Lyu et al. 2024) conducts training on multi-view diffusion within a large-scale avatar dataset, aiming to generate high-fidelity 3D heads. However, its input images are restricted to the frontal views of identities, which poses a constraint on the data diversity and potential application scenarios.

Multi-View Diffusion Models: Multi-View diffusion builds upon text-to-image diffusion (Rombach et al. 2022), leveraging the strong 2D priors of diffusion models and achieving significant advancements through training on large-scale 3D datasets. Most methods focus their innovations on maintaining multi-view consistency. For instance, Zero123++ (Shi et al. 2023) fixes the camera viewpoint and employs a linear noise addition strategy, Wonder3D (Long et al. 2024) enhances consistency by controlling the generation of multi-view normal maps and RGB images, while ImageDream (Wang and Shi 2023) aligns text-to-3D diffusion

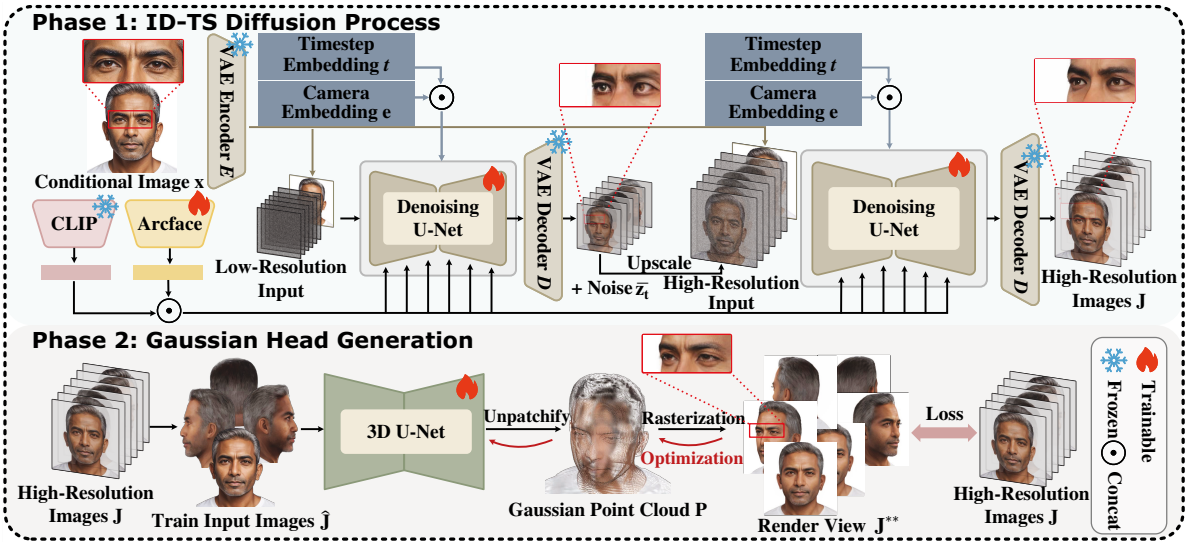

Figure 2: Overview of our inference pipeline. **Phase 1**: In the low-resolution stage, we embed camera pose **e** and noise step **t** via positional encoding, concatenate them, and feed them into the U-Net's residual blocks. The conditional image **x** is encoded into the latent space of the VAE encoder **E**, concatenated with noise, and processed alongside CLIP and ArcFace embeddings via cross-attention in the transformer block, generating multi-view images with accurate head shapes and coarse facial texture. In the high-resolution stage, we upsample the previous outputs, do element-wise addition with latent noise, and denoise them, outputting comprises multi-view images with high-fidelity texture details. **Phase 2**: With front/left/back/right images as inputs and remaining frames as supervision for 3D U-Net optimization, we finally yield a high-quality Gaussian head **P**.

model (Shi et al. 2024) capabilities to ensure consistency in image-to-3D synthesis. However, their performance in multi-view facial synthesis is suboptimal, and the number of generated views is highly limited, which constrains the effectiveness of downstream 3D facial reconstruction tasks.

3D Reconstruction: Reconstruction algorithms have made significant progress since the advent of NeRF (Mildenhall et al. 2020), which implicitly encodes a 3D scene into a neural network and renders it using volumetric functions. And there are also many methods (Poole et al. 2023; Wang and Shi 2023) that apply NeRF for Score Distillation Sampling (SDS). Other works (Hong et al. 2024) apply triplane for 3D scene encoding and utilize MLP to decode 2D images. However, the implicit representation restricts their applications, especially in object reconstruction. Recently, 3D Gaussian Splatting (Kerbl et al. 2023) has demonstrated promising reconstruction results with the representation of Gaussian point cloud. This approach can be utilized for modeling dynamic scenes or objects, and its rendering speed is also fast. Nowadays, feed-forward Gaussian generation methods (Tang et al. 2024; Zhang et al. 2024b) further improved multi-view consistency and surface smoothness in sparse view inputs, by integrating neural networks into Gaussian point clouds generation process and training in large-scale 3D dataset. However, the Gaussian point clouds generated through such a single forward pass often exhibit deficiencies in fine details, particularly for unseen objects.

3 Methodology

3.1 Problem Formulation

Given a single image $\mathbf{x} \in \mathrm{R}^{3\times H\times W}$ of an identity **I** and the corresponding rough camera elevation **e**, we first encode this information into our ID-TS diffusion model. In the low-resolution stage, we generate multi-view images $\bar{\mathbf{J}} \in \mathrm{R}^{N_s\times 3\times h\times w}$, N_s denotes image number, $h = H/\sigma$, $w = W/\sigma$, and σ denoting the downsampling ratio. In the high-resolution stage, we upsample $\bar{\mathbf{J}}$ and mix it with the random noise $\bar{\mathbf{z}}_t \in \mathrm{R}^{N_s\times 3\times H\times W}$. The denoising process finally outputs multi-view images $\mathbf{J} \in \mathrm{R}^{N_s\times 3\times H\times W}$ around the identity **I**. The camera positions of these output images are represented as $\pi \in \mathrm{R}^{N_s\times 2} = \{(\mathbf{e}_i, \mathbf{a}_i)\}_{i=1}^{N_s}$, where $\mathbf{e}_i$ corresponds to the elevation of each camera, equal to the input elevation **e**, and $\mathbf{a}_i$ represents the azimuth. After that, we optimize feed-forward Gaussian generation network to reconstruct the 3D head model using 4 of the multi-view images $\hat{\mathbf{J}} \in \mathrm{R}^{4\times 3\times H\times W}$ and the corresponding camera parameters $\hat{\pi} \in \mathrm{R}^{4\times 2} = \{(\mathbf{e}_j, \mathbf{a}_j)\}_{j=1}^{4}$ as inputs and rest images as supervisory data. Finally, we get Gaussian point cloud $\mathbf{P} \in \mathrm{R}^{(4\times H'\times W')\times 14}$, where H' and W' represent network output shape, the last dimension "14" includes 3-channel colors, 3-channel scale factors, 3-channel position coordinates, 4-channel rotation factors and 1-channel opacity factor. The total number of Gaussian points is $4 \times H' \times W'$.

3.2 Digital Human Head Dataset

Existing large-scale 3D datasets for multi-view diffusion models suffer from several limitations: low-quality textures, insufficient mesh polygon counts, and a notable scarcity of

Figure 3: Different expressions and an accessory in 3D digital human models. Diverse facial expressions are sampled to enrich our dataset, along with an accessory to validate the fitting capability of our model.

3D human head models. Furthermore, available public head datasets typically only provide limited viewpoint variations. For the purpose of getting high-fidelity 3D full-head models, we construct a digital human head dataset, which includes different genders, ages, races, *etc*. Moreover, we provide high-fidelity texture details to facilitate network training. In total, we construct 227 sequences of 3D digital human portraits. **For detailed construction procedures, please refer to the supplementary material.**

To train our neural network, we sample multi-view images from the 3D models of digital human heads. Specifically, we render 1024×1024 resolution images from $N = 100$ realistic human head models. These renderings cover $N_e = 6$ elevation angles of $[-10°, 0°, 10°, 20°, 30°, 40°]$, azimuth angles at $N_a = 16$ positions around the head, resulting in $N_v = 96$ unique viewpoints per head. Additionally, each avatar features $N_f = 10$ appearances with 9 different expressions and 1 accessory, as shown in Fig. 3.

3.3 ID-TS Diffusion Model

Multi-View image generation is a critical step in our method, providing the essential high-quality, identity-preserving, and appearance-consistent face images from different perspectives for 3D head reconstruction. We train our ID-TS diffusion model which takes one single image $\mathbf{x}$ as input and outputs spatially consistent multi-view images $\mathbf{J}$. The entire process is illustrated in the higher segment of Fig. 2.

Architecture: ID-TS diffusion model consists of a two-stage diffusion model, and each diffusion model is initialized from Stable Video Diffusion (Blattmann et al. 2023). We adapt it into a multi-view generative framework, drawing inspiration from (Voleti et al. 2024; Yang et al. 2024). For each diffusion model, the architecture contains VAE encoder $\mathbf{E}$, decoder $\mathbf{D}$ and denoising U-Net. Each U-Net block includes a residual module and two transformer modules, equipped with a temporal attention layer and a spatial attention layer. The input condition image $\mathbf{x}$ is processed in two ways. First, it is encoded into a latent space state and then concatenated with the noisy latent sequence $\mathbf{z}_t$ for denoising. Second, $\mathbf{x}$ is used as guidance for the cross-attention mechanism with $\mathbf{z}_t$. This guidance ensures the denoising U-Net generates content aligned with the input identity. Typically, CLIP (Radford et al. 2021) embeddings are used as guidance to provide semantic information for general images. Besides, we additionally employ identity guidance, utilizing ArcFace (Deng et al. 2019), a network specifically designed for facial representation, to better preserve facial identity. At the same time, the camera pose $\mathbf{e}$ is integrated with timestep $\mathbf{t}$ and fed into the residual block, ultimately producing frames of multi-view surround images. The fine-tuning objective is defined as follows:

$$\min_{\theta} \mathrm{E}_{z\sim\varepsilon(x),t,\epsilon\sim N(0,1)}||\epsilon - \epsilon_\theta(\mathbf{z}_t; \mathbf{t}, \mathbf{e}, \mathbf{x})||_2^2, \quad (1)$$

θ represents the network parameters, ϵ is the scheduled Gaussian noise added during training, and ϵ_θ is the predicted noise. The camera pose $\mathbf{e}$ controls the vertical viewpoint of the multi-view images. After the latent diffusion process, the latent state is decoded by the VAE decoder $\mathbf{D}$ to generate the multi-view images.

Identity Guidance: Previous approaches (Liu et al. 2023; Shi et al. 2023; Long et al. 2024; Wang and Shi 2023) leverage CLIP as a conditional guidance mechanism. However, since CLIP is trained solely for image-text alignment, it lacks the capability to effectively comprehend and represent more sophisticated facial attributes such as identity information and expression variations. To provide better facial information guidance for diffusion models, we leverage ArcFace (Deng et al. 2019) for improved cross-attention signals. Notably, ArcFace also serves as a powerful classifier; thus, we fine-tune the network on datasets specifically designed for facial expressions and identities to further enhance facial representation. Through this finetuning process, the ArcFace embedding is capable of distinguishing between different expressions of the same identity while providing consistent representations across different viewing angles.

Two-Stage Diffusion Process: While single-stage diffusion models effectively capture global structural information in early denoising steps and refine local details in later stages (Balaji et al. 2022), such an approach still exhibits limitations in the domain of multi-view human head generation. Specifically, it shows deficiencies in maintaining detailed feature coherence (particularly in eyes and teeth) and produces suboptimal texture clarity. Inspired by DS-VTON (Sun et al. 2025), we introduce a dual-scale framework that explicitly disentangles head structure maintenance and facial texture enhancement. We first introduce a low-resolution stage to guide the generation process. The proposed process suppresses high-frequency information to guide the diffusion model toward generating multi-view consistent head structures while providing coarse facial landmark priors (e.g., eyes, nose, mouth) for geometric guidance. During training, we downsample ground truth images $\mathbf{J}^*$ to $h \times w$ and these images are encoded with the latent state as $\mathbf{z}_0$, produced by the VAE encoder $\mathbf{E}$. Then we add noise to the images, predict and remove it using a denoising U-Net, and finally decode the denoised output through the VAE Decoder $\mathbf{D}$ to reconstruct low-resolution images $\bar{\mathbf{J}}$. Then, a high-resolution diffusion model is introduced for final results generation. In this stage, we extend a residual-based denoising strategy that predicts the residual between high-resolution images and their low-resolution

counterparts. This enables the model to focus specifically on texture restoration, building upon the structural alignment achieved in the first stage. Specifically, we encode ground truth images $\mathbf{J}^*$ into the latent space using VAE encoder $\mathbf{E}$ and denote them as $\bar{\mathbf{z}}_0$. Meanwhile, we upsample the first-stage results $\bar{\mathbf{J}}$ to $H \times W$, then encode them into the latent space as $\bar{\mathbf{j}}$. Next, we blend $\bar{\mathbf{j}}$ and Gaussian noise ϵ in a predefined ratio to synthesize the noisy input for denoising. Under this formulation, for each latent noisy sequence $\bar{\mathbf{z}}_t$ the forward and reverse diffusion processes become:

$$\bar{\mathbf{z}}_t = \sqrt{\alpha_t}\bar{\mathbf{z}}_{t-1} + \sqrt{1-\alpha_t}(\xi_1 \cdot \epsilon + \xi_2 \cdot \bar{\mathbf{j}}), \quad (2)$$

$$\bar{\mathbf{z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\bar{\mathbf{z}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_{\bar{\theta}}(\bar{\mathbf{z}}_t, \mathbf{t}, \mathbf{e}, \mathbf{x})) + \sigma_t \mathbf{z}, \quad (3)$$

where ξ_1 and ξ_2 represent the noise weights, $\bar{\epsilon}_{\bar{\theta}}$ represent the predicted noise in this stage, other notations follow DDPM(Ho, Jain, and Abbeel 2020). Through this residual-guided denoising process, we obtain texture-enhanced high-resolution multi-view images $\mathbf{J}$, providing high-quality inputs for downstream reconstruction tasks.

Objective Functions: To improve generalization, we add identity loss to our loss function. The identity loss computes the cosine distance between the latent features of $\mathbf{J}$ and the ground truth $\mathbf{J}^* \in \mathrm{R}^{3\times H\times W}$ extracted from ArcFace network. It is defined as:

$$\mathcal{L}_{ID}(\mathbf{J}, \mathbf{J}^*) = 1 - \frac{1}{N_s}\sum_{i}^{N_s}\frac{j_i \cdot j_i^*}{||j_i||||j_i^*||}, \quad (4)$$

where j_i, j_i^* represent features of $\mathbf{J}$ and $\mathbf{J}^*$. We finally employ identical loss functions for both stages, including MSE loss, perceptual loss (Zhang et al. 2018), and ID loss. The overall loss function is defined as:

$$\mathcal{L}_{\mathcal{M}}(\mathbf{J}, \mathbf{J}^*) = \mathcal{L}_{MSE} + \lambda_p\mathcal{L}_{perc} + \lambda_i\mathcal{L}_{ID}, \quad (5)$$

where λ_p and λ_i are loss weights.

3.4 3D Gaussian Avatar Reconstruction

We reconstruct a 3D Gaussian avatar from the multi-view images $\mathbf{J}$ and their corresponding camera positions π. Although existing feed-forward Gaussian point cloud generation methods (Tang et al. 2024; Zhang et al. 2024b) have achieved remarkable success in object reconstruction from sparse-view inputs, they still struggle to faithfully recover facial textures and fine-grained details in human head reconstruction. To address this, we utilize a feed-forward Gaussian reconstruction approach (Tang et al. 2024) for enhanced reconstruction. The entire process is illustrated in the lower segment of Fig. 2.

Architecture: The method proposes a 3D U-Net for Gaussian point cloud generation. Each block in this U-Net comprises a residual layer and a 3D self-attention layer (two spatial dimensions and one dimension across input images). Given 4 input images $\hat{\mathbf{J}}$ and corresponding camera parameters $\hat{\pi}$, the method first calculates each camera's world coordinate origin $o_n \in \mathrm{R}^3$ and ray direction $d_n \in \mathrm{R}^3$ for each pixel, and then the method concatenates $\hat{\mathbf{J}}$, the corresponding Plücker ray embedding $o_n \times d_n$ and the ray direction d_n in the last channel as the 3D U-Net's inputs. Finally, the U-Net outputs a feature map shaped as [4, H', W', 14], and by flattening the first three channels, a Gaussian point cloud $\mathbf{P}$ is generated.

Single Subject Optimization: Building upon the 3D U-Net architecture, we fine-tune the network on our digital human head dataset to optimize performance. However, empirical evaluation indicated that the one-step generation results exhibit deficiencies in both facial texture details and multi-view consistency. To address this issue, we leverage the availability of multi-view images from our upstream pipeline to perform individualized optimization for each subject. Specifically, we choose multi-view input configuration comprising 4 equidistant perspectives $\hat{\mathbf{J}}$ (frontal, right, posterior, and left views) to generate a Gaussian point cloud representation $\mathbf{P}$. Subsequently, we utilize the camera parameters associated with $\mathbf{J}$ to render synthesized images $\mathbf{J}^{**} \in \mathrm{R}^{N_s\times 3\times H\times W}$, which are then iteratively optimized against the corresponding multi-view reference images from ID-TS diffusion process. The final loss function aligns with the ID-TS model, comprising MSE loss, perceptual loss, and ID loss. With this optimization strategy, we significantly enhance the quality of 3D reconstruction for each subject, achieving accurate facial details and multi-view consistency.

4 Experiments

4.1 Experimental Settings

During the multi-view stage, our model generates $N_S = 16$ multi-view head images with an initial learning rate of 1e-6 and a cosine annealing scheduler, setting the loss function coefficients $\lambda_p = \lambda_i = 1.0$; the low-resolution stage operates at $h \times w = 256 \times 256$ resolution and trains for 20h on 8×A6000 GPUs, while the high-resolution stage uses $H \times W = 512 \times 512$ resolution with a mix ratio $\xi_1 = 0.4, \xi_2 = 0.6$ of low-resolution inputs and noise(mixing ratio experiment is shown in supplementary material), training for 4 days on 8×A6000 GPUs; for the Gaussian reconstruction network, we set the learning rate to 1e-5 and the loss function coefficients $\lambda_p = \lambda_i = 1.0$, outputting a $4 \times H' \times W' \times 14 = 4 \times 128 \times 128 \times 14$ tensor after 2 days of training on 8×A6000 GPUs. We also perform per-subject 3D U-Net optimization (LR=1e-6) on a single A6000 GPU; it takes 400 s in total, with seed 42 for reproducibility.

Datasets. Our experiments are conducted on two distinct datasets for comprehensive evaluation. The first dataset is the NeRSemble v2 (Kirschstein et al. 2023), where we select 14 subjects, each featuring 2 distinct expressions. The second dataset is our newly constructed digital human head dataset, comprising previously unseen faces and expressions, where we select 10 subjects, each with 5 distinct expressions. For quantitative evaluation, we randomly sample 16 multi-view facial images, while for qualitative assessment, the test viewpoints are not constrained to frontal faces. As for ablative analyses, we only use digital human head because of the viewpoint restriction.

Baselines. For the evaluation of the entire pipeline, from a single image to a 3D head representation, we selected single-image-to-3D-head methods that provide pub-

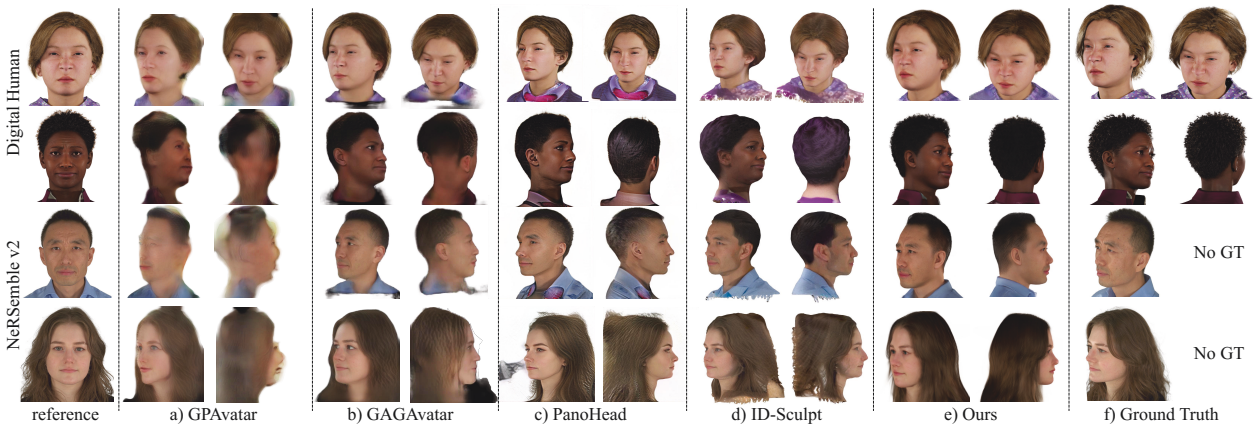

Figure 4: Qualitative comparison. Visualization results demonstrate that our method significantly outperforms existing approaches in capturing fine-grained details, including head geometry, facial feature consistency, expression texture, and gaze direction.

Dataset	Method	LPIPS↓	PSNR↑	CSIM↑	DreamSim↓	User Study↑
Digital Human	GPAvatar	0.7452 ± 0.0050	10.1993 ± 0.0773	0.6461 ± 0.0127	0.3424 ± 0.0092	0.0
	GAGAvatar	0.2960 ± 0.0081	13.2312 ± 0.1907	0.8808 ± 0.0078	0.1611 ± 0.0050	0.0
	PanoHead	0.3230 ± 0.0101	13.0568 ± 0.5874	0.8603 ± 0.0132	0.1453 ± 0.0080	0.0650
	ID-Sculpt	0.3145 ± 0.0111	12.5297 ± 0.5861	0.8483 ± 0.0142	0.2080 ± 0.0159	0.0215
	Ours	**0.2315 ± 0.0054**	**20.4287 ± 0.2082**	**0.9189 ± 0.0062**	**0.0804 ± 0.0031**	**0.9135**
NeRSemble v2	GPAvatar	0.7081 ± 0.0024	10.8568 ± 0.1521	0.5336 ± 0.0130	0.2740 ± 0.0056	0.0
	GAGAvatar	0.3329 ± 0.0090	13.2822 ± 0.3308	0.8424 ± 0.0132	0.1370 ± 0.0040	0.0215
	PanoHead	0.3649 ± 0.0120	13.3809 ± 0.3605	0.8379 ± 0.0118	0.1128 ± 0.0072	0.0655
	ID-Sculpt	0.3284 ± 0.0085	13.7613 ± 0.4523	0.8293 ± 0.0090	0.1587 ± 0.0105	0.0
	Ours	**0.3147 ± 0.0099**	**16.0120 ± 0.3365**	**0.8519 ± 0.0115**	**0.1043 ± 0.0033**	**0.9130**

Table 1: Quantitative evaluation of single image to 3D head. We conduct a comprehensive evaluation of our method against 4 state-of-the-art 3D head generation approaches across two distinct datasets, employing five diverse metrics. Experimental results demonstrate that our approach consistently outperforms all competing methods, achieving superior performance across all evaluation criteria.

licly available implementations, including GPAvatar (Chu et al. 2024)(ICLR 2024), GAGAvatar (Chu and Harada 2024)(NeurIPS 2024), PanoHead (An et al. 2023)(CVPR 2023), and ID-Sculpt (Hao et al. 2025)(AAAI 2025).

Evaluation Metric. We compared each generated frame with its corresponding ground truth frame using LPIPS (Zhang et al. 2018), PSNR, DreamSim (Fu et al. 2023), and cosine similarity of identity embeddings (CSIM) (Deng et al. 2019). These metrics collectively assess both pixel-level accuracy and identity preservation. Additionally, we conducted a questionnaire-based user study, selecting four test cases from each comparison category and asking participants to choose the best generation method. We then analyzed the distribution of preferences based on the proportion of responses.

4.2 Single Image to 3D head Evaluation

We evaluated the whole pipeline from single head image input to 3D full-head representation. We observe that varying camera coordinate systems across different methods lead to inconsistent head sizes and positions in the sampled multi-view images. To mitigate this, we align the head keypoints of images generated by different methods, minimizing potential testing bias. Quantitative results are shown in Tab. 1, and qualitative comparison is presented in Fig. 4. Among these methods, GPAvatar struggles with large-angle profile generation, GAGAvatar exhibits errors in gaze direction, PanoHead lacks fidelity in clothing and expression preservation, while ID-Sculpt suffers from inaccurate shape estimation. Both experiments demonstrate that our method achieves the best results in generating realistic head shapes and fine-grained facial texture details. Additional experiments and dynamic video comparisons are provided in the supplementary material.

4.3 Ablative Analyses

For ablative analyses, we first evaluate the effect of ID guidance at the low-resolution stage. Then, we examine the impact of incorporating ID guidance in ID-TS diffusion and using a two-stage diffusion process. Finally, we test different 3D point cloud generation strategies, and we define that "O" represents training with 3D object dataset, "H" repre-

Figure 5: Applications for downstream tasks. We evaluate our method on three downstream tasks: frontal view prediction from profile views, stylized head generation, and facial animation, demonstrating its versatility across diverse application scenarios.

Category	Ablation	LPIPS↓	PSNR↑	CSIM↑	DreamSim↓
Low-Resolution Stage 256x256	Fine-tuned	0.2816 ± 0.0080	18.6968 ± 0.3564	0.8860 ± 0.0087	0.1217 ± 0.0068
	+ID	**0.2770 ± 0.0075**	**18.9627 ± 0.2307**	**0.8961 ± 0.0100**	**0.1106 ± 0.0063**
ID-TS Diffusion 512x512	Fine-tuned	0.2446 ± 0.0063	17.5717 ± 0.4647	0.8799 ± 0.0098	0.0938 ± 0.0064
	+ID	0.2414 ± 0.0036	18.2606 ± 0.1940	0.9039 ± 0.0062	0.0870 ± 0.0030
	+TS	0.2092 ± 0.0040	19.1919 ± 0.4311	0.9164 ± 0.0072	0.0603 ± 0.0026
	+ID+TS	**0.2028 ± 0.0034**	**19.9974 ± 0.2503**	**0.9261 ± 0.0071**	**0.0585 ± 0.0022**
Gaussian Head Generation	O	0.2980 ± 0.0044	15.6821 ± 0.0846	0.8578 ± 0.0067	0.1413 ± 0.0032
	O+H	0.2556 ± 0.0048	18.3844 ± 0.1326	0.9005 ± 0.0039	0.1036 ± 0.0039
	O+H+S	**0.2315 ± 0.0054**	**20.4287 ± 0.2082**	**0.9189 ± 0.0062**	**0.0804 ± 0.0031**

Table 2: Quantitative ablation results for our three core contributions. Both ID guidance and the two-stage diffusion process show higher scores, and per-subject optimization presents better results among different reconstruction strategies.

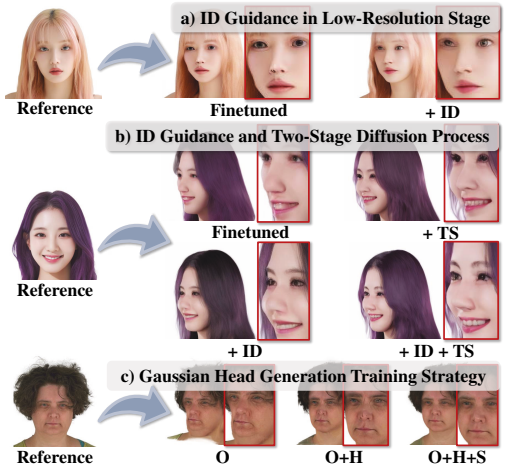

Figure 6: Visual results for ablation study. Visualization demonstrates that our innovation significantly enhances the quality of multi-view and 3D head model reconstruction.

sents fine-tuning with 3D digital human dataset, "S" represents optimizing in single subject. Quantitative experiments are performed solely on the digital head dataset, as shown in Tab. 2. And visual comparison results are presented in Fig. 6. Experimental results demonstrate that identity guidance preserves facial fidelity, while the two-stage diffusion process enhances texture details. Our single-head optimization reduces Gaussian splatting artifacts, improving 3D reconstruction quality.

4.4 Applications

We demonstrate three representative downstream applications enabled by our 3D reconstruction framework, with qualitative results visualized in Fig. 5.

Frontal View Prediction. Frontal view prediction can facilitate applications such as suspect facial reconstruction and access control enhancement in face recognition. Our approach employs randomly sampled viewpoint conditions during multi-view training process, enabling the model to generate high-quality frontal face images even from extreme profile views.

Stylized Generation. Stylized 3D head generation enables the creation of virtual 3D avatars and facilitates applications in film production. We propose a framework that constructs high-fidelity 3D anime-style characters by first translating 2D images into stylized artwork and then integrating the processed head into our model for 3D reconstruction.

Facial Animation. The construction of 4D head avatars enables applications in virtual video conferencing, virtual streaming, and related domains. We achieve 4D head avatar synthesis by sampling video outputs from our 3D head model and leveraging a 2D facial expression transfer framework (Guo et al. 2024).

5 Conclusion

Although our method achieves promising results, limitations remain in image resolution due to hardware constraints and in generalization to unseen accessories, which we aim to address through improved hardware and dataset expansion.

Acknowledgments

This research is supported, in part, by the National Natural Science Foundation of China (Grant No. 62302295), the Shanghai Municipal Science and Technology Major Project, China (Grant No. 2021SHZDZX0102), the Startup Fund for Young Faculty at SJTU (Grant No. 22X010503821) and the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, P.R. China.

References

An, S.; Xu, H.; Shi, Y.; Song, G.; Ogras, Ü. Y.; and Luo, L. 2023. PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360°. In *CVPR, 2023*, 20950–20959. IEEE.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *CoRR*, abs/2211.01324.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, 187–194. USA: ACM Press/Addison-Wesley Publishing Co. ISBN 0201485605.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; Jampani, V.; and Rombach, R. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *CoRR*, abs/2311.15127.

Bühler, M. C.; Li, G.; Wood, E.; Helminger, L.; Chen, X.; Shah, T.; Wang, D.; Garbin, S. J.; Orts-Escolano, S.; Hilliges, O.; Lagun, D.; Riviere, J.; Gotardo, P. F. U.; Beeler, T.; Meka, A.; and Sarkar, K. 2024. Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures. In Igarashi, T.; Shamir, A.; and Zhang, H. R., eds., *SIGGRAPH Asia 2024 Conference Papers, 2024*, 29:1–29:12. ACM.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; Mello, S. D.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; Karras, T.; and Wetzstein, G. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR, 2022*, 16102–16112. IEEE.

Chu, X.; and Harada, T. 2024. Generalizable and Animatable Gaussian Head Avatar. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *NeurIPS, 2024*.

Chu, X.; Li, Y.; Zeng, A.; Yang, T.; Lin, L.; Liu, Y.; and Harada, T. 2024. GPAvatar: Generalizable and Precise Head Avatar from Image(s). In *ICLR, 2024*. OpenReview.net.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; VanderBilt, E.; Kembhavi, A.; Vondrick, C.; Gkioxari, G.; Ehsani, K.; Schmidt, L.; and Farhadi, A. 2023a. Objaverse-XL: A Universe of 10M+ 3D Objects. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS, 2023*.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023b. Objaverse: A Universe of Annotated 3D Objects. In *CVPR, 2023*, 13142–13153. IEEE.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE CVPR, 2019*, 4690–4699. Computer Vision Foundation / IEEE.

Deng, Y.; Wang, D.; and Wang, B. 2024. Portrait4D-V2: Pseudo Multi-view Data Creates Better 4D Head Synthesizer. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV, 2024*, volume 15075 of *Lecture Notes in Computer Science*, 316–333. Springer.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8).

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS, 2023*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *NeurIPS, 2014*, 2672–2680.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168*.

Hao, J.; Tang, J.; Zhang, J.; Yi, R.; Hong, Y.; Li, M.; Cao, W.; Wang, Y.; Wang, C.; and Ma, L. 2025. ID-Sculpt: ID-aware 3D Head Generation from Single In-the-wild Portrait Image. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI, 2025*, 3383–3391. AAAI Press.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS, 2020*.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2024. LRM: Large Reconstruction Model for Single Image to 3D. In *ICLR, 2024*. OpenReview.net.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139:1–139:14.

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023. NeRSemble: Multi-View Radiance Field Reconstruction of Human Heads. *ACM Trans. Graph.*, 42(4).

Liu, R.; Wu, R.; Hoorick, B. V.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. In *ICCV, 2023*, 9264–9275. IEEE.

Long, X.; Guo, Y.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.; Habermann, M.; Theobalt, C.; and Wang, W. 2024. Wonder3D: Single Image to 3D Using Cross-Domain Diffusion. In *CVPR, 2024*, 9970–9980. IEEE.

Lyu, W.; Zhou, Y.; Yang, M.-H.; and Shu, Z. 2024. FaceLift: Single Image to 3D Head with View Generation and GS-LRM. arXiv:2412.17812.

Martinez, J.; Kim, E.; Romero, J.; Bagautdinov, T. M.; Saito, S.; Yu, S.; Anderson, S.; Zollhöfer, M.; Wang, T.; Bai, S.; Li, C.; Wei, S.; Joshi, R.; Borsos, W.; Simon, T.; Saragih, J. M.; Theodosis, P.; Greene, A.; Josyula, A.; Maeta, S.; Jewett, A.; Venshtain, S.; Heilman, C.; Chen, Y.; Fu, S.; Elshaer, M.; Du, T.; Wu, L.; Chen, S.; Kang, K.; Wu, M.; Emad, Y.; Longay, S.; Brewer, A.; Shah, H.; Booth, J.; Koska,

T.; Haidle, K.; Andromalos, M.; Hsu, J.; Dauer, T.; Selednik, P.; Godisart, T.; Ardisson, S.; Cipperly, M.; Humberston, B.; Farr, L.; Hansen, B.; Guo, P.; Braun, D.; Krenn, S.; Wen, H.; Evans, L.; Fadeeva, N.; Stewart, M.; Schwartz, G.; Gupta, D.; Moon, G.; Guo, K.; Dong, Y.; Xu, Y.; Shiratori, T.; Prada, F.; Pires, B.; Peng, B.; Buffalini, J.; Trimble, A.; McPhail, K.; Schoeller, M.; and Sheikh, Y. 2024. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *NeurIPS, 2024*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *ECCV, 2020*, volume 12346 of *Lecture Notes in Computer Science*, 405–421. Springer.

Pan, D.; Zhuo, L.; Piao, J.; Luo, H.; Cheng, W.; Wang, Y.; Fan, S.; Liu, S.; Yang, L.; Dai, B.; Liu, Z.; Loy, C. C.; Qian, C.; Wu, W.; Lin, D.; and Lin, K. 2023. RenderMe-360: A Large Digital Asset Library and Benchmarks Towards High-fidelity Head Avatars. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS, 2023*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR, 2023*. OpenReview.net.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *ICML, 2021*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF CVPR, 2022*, 10684–10695.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

Shi, Y.; Wang, P.; Ye, J.; Mai, L.; Li, K.; and Yang, X. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *ICLR, 2024*. OpenReview.net.

Sun, X.; Hong, Y.; Zhan, J.; Lan, J.; Zhu, H.; Wang, W.; Zhang, L.; and Zhang, J. 2025. DS-VTON: High-Quality Virtual Try-on via Disentangled Dual-Scale Generation. *arXiv preprint arXiv:2506.00908*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. LGM: Large Multi-view Gaussian Model for High-Resolution 3D Content Creation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV, 2024*, volume 15062 of *Lecture Notes in Computer Science*, 1–18. Springer.

Taubner, F.; Zhang, R.; Tuli, M.; and Lindell, D. B. 2025. CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusion Models. In *CVPR, 2025*, 5318–5330. Computer Vision Foundation / IEEE.

Voleti, V.; Yao, C.; Boss, M.; Letts, A.; Pankratz, D.; Tochilkin, D.; Laforte, C.; Rombach, R.; and Jampani, V. 2024. SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image Using Latent Video Diffusion. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV, 2024*, volume 15059 of *Lecture Notes in Computer Science*, 439–457. Springer.

Wang, P.; and Shi, Y. 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; and Guo, B. 2023. RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *CVPR, 2023*, 4563–4573. IEEE.

Wang, Z.; Zhu, X.; Zhang, T.; Wang, B.; and Lei, Z. 2024. 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. In *CVPR, 2024*, 1672–1682. IEEE.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models. *CoRR*, abs/2404.07191.

Yang, H.; Chen, Y.; Pan, Y.; Yao, T.; Chen, Z.; Ngo, C.; and Mei, T. 2024. Hi3D: Pursuing High-Resolution Image-to-3D Generation with Video Diffusion Models. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM, 2024*, 6870–6879. ACM.

Ye, Z.; Zhong, T.; Ren, Y.; Yang, J.; Li, W.; Huang, J.; Jiang, Z.; He, J.; Huang, R.; Liu, J.; Zhang, C.; Yin, X.; Ma, Z.; and Zhao, Z. 2024. Real3D-Portrait: One-shot Realistic 3D Talking Portrait Synthesis. In *ICLR, 2024*. OpenReview.net.

Zhang, B.; Cheng, Y.; Wang, C.; Zhang, T.; Yang, J.; Tang, Y.; Zhao, F.; Chen, D.; and Guo, B. 2024a. RodinHD: High-Fidelity 3D Avatar Generation with Diffusion Models. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV, 2024*, volume 15072 of *Lecture Notes in Computer Science*, 465–483. Springer.

Zhang, K.; Bi, S.; Tan, H.; Xiangli, Y.; Zhao, N.; Sunkavalli, K.; and Xu, Z. 2024b. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV, 2024*, volume 15080 of *Lecture Notes in Computer Science*, 1–19. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE CVPR, 2018*, 586–595. Computer Vision Foundation / IEEE Computer Society.