ID-Sculpt: ID-aware 3D Head Generation from Single In-the-wild Portrait Image

Jinkun Hao[1], Junshu Tang[1], Jiangning Zhang[2], Ran Yi[1*], Yijia Hong[1], Moran Li[2], Weijian Cao[2], Yating Wang[1], Chengjie Wang[1, 2], Lizhuang Ma[1*]

[1]Shanghai Jiao Tong University
[2]Youtu Lab, Tencent
{leo_hao, tangjs, ranyi, hyj542682306, wyating_0929, lzma}@sjtu.edu.cn;
{vtzhang, moranli, weijiancao, jasoncjwang}@tencent.com;

Abstract

While recent works have achieved great success on image-to-3D object generation, high quality and fidelity 3D head generation from a single image remains a great challenge. Previous text-based methods for generating 3D heads were limited by text descriptions and image-based methods struggled to produce high-quality head geometry. To handle this problem, we propose a novel framework, **ID-Sculpt**, to generate high-quality 3D heads while preserving their identities. Our work incorporates the identity information of the portrait image into three parts: 1) geometry initialization, 2) geometry sculpting, and 3) texture generation stages. Given a reference portrait image, we first align the identity features with text features to realize ID-aware guidance enhancement, which contains the control signals representing the face information. We then use the canny map, ID features of the portrait image, and a pre-trained text-to-normal/depth diffusion model to generate ID-aware geometry supervision, and 3D-GAN inversion is employed to generate ID-aware geometry initialization. Furthermore, we use ID-aware guidance to calculate ID-aware Score Distillation (ISD) for geometry sculpting. For texture generation, we adopt the ID Consistent Texture Inpainting and Refinement which progressively expands the view for texture inpainting to obtain an initialization UV texture map. We then use the ID-aware guidance to provide image-level supervision for noisy multi-view images to obtain refined texture maps. Extensive experiments demonstrate that we can generate high-quality 3D heads with accurate geometry and texture from a single in-the-wild portrait image.

Project page — https://jinkun-hao.github.io/IDSculpt/

1 Introduction

3D head generation aims to create high-quality 3D digital assets of human heads that align with user preferences, utilizing various input data, which has significant applications in film character creation, gaming, online meetings, education, etc. Conventional methods rely on multi-view geometric consistency (Kirschstein et al. 2023; Zheng et al. 2023a; Munkberg et al. 2022; Yariv et al. 2021) or statistical 3D face prior model (Bai et al. 2023; Zheng et al. 2023b; Grassal et al. 2022) to achieve high-quality 3D head generation. The former necessitates multi-view images or several minutes of monocular video, which are not applicable to a single in-the-wild portrait image. Moreover, 3D face shape prior models typically focus on the face region, failing to accurately reconstruct the entire head shape and being unable to handle elements such as long hair and clothing. An alternative approach involves employing Generative Adversarial Networks (GANs) to accomplish 3D head generation, which is constrained by limited training data and typically produces frontal-facing faces. Moreover, they struggle to preserve identity consistency when presented with images of large head poses. Recently, text-to-image (Ramesh et al. 2021, 2022; Rombach et al. 2022; Saharia et al. 2022; Han et al. 2023; Pei et al. 2024) generation models trained on large-scale image datasets have achieved significant success. Some approaches employ SDS loss (Poole et al. 2022) to ensure similarity between the generated 3D model and images, facilitating image-to-3D (Melas-Kyriazi et al. 2023; Tang et al. 2023; Liu et al. 2023c,b; Long et al. 2023) or text-to-3D (Poole et al. 2022; Lin et al. 2023; Sanghi et al. 2022; Wang et al. 2022, 2024). However, when employing these methods for customized 3D head generation, the absence of identity information makes it difficult to reproduce the geometric details and texture realism in the portrait image.

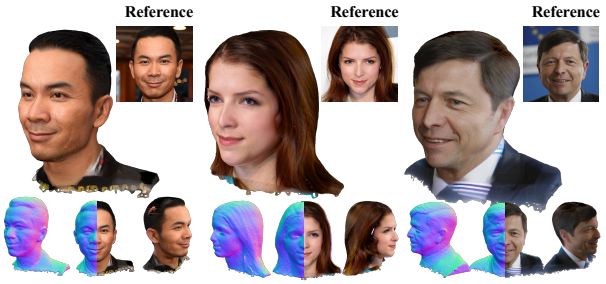

Figure 1: ID-Sculpt can generate a high-quality 3D head with detailed face geometry and photo-realistic texture given an in-the-wild portrait image.

Through analyzing previous methods, we summarize the key challenges from the three stages in the optimization-based 3D head generation approach: ***1)*** Geometry Initialization: Previous image-to-3D methods used an ellipsoid as the

*Corresponding author.

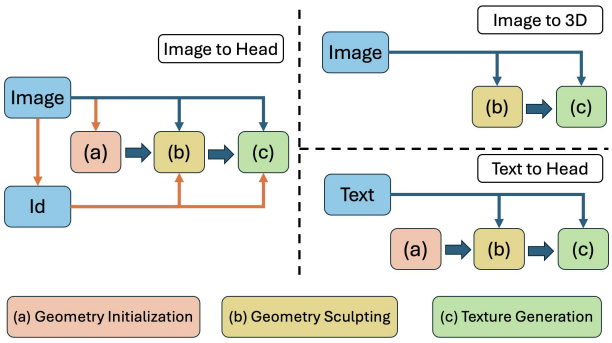

Figure 2: Comparison of previous optimization methods (Right) with our method (Left). Our method leverages the portrait image to get a better initial mesh and utilizes extracted identity information to enhance geometric and texture optimization.

initial geometry (Chen et al. 2023) and could not accurately reconstruct the geometry of the head. On the other hand, text-to-head methods used a unified FLAME head (Han et al. 2024; Liu et al. 2023a) as the geometric prior, which lacks personalized information and cannot generate individualized features that deviate from the FLAME head, such as long hair. ***2)*** Geometry Sculpting: Although image-to-3D methods attempted to use view-dependent diffusion (Liu et al. 2023b)to guide geometry generation, the resulting geometry lacked the generalization ability to the reasonable head geometry. Besides, text-to-3D methods use facial image description text to generate geometry, which can result in a deviation from the identity of the reference image. ***3)*** Texture Generation: Previous methods used VSD (Wang et al. 2024) to mitigate color over-saturation issues based on SDS. However, the generated textures still tend to have excessive shading and unrealistic texture (Huang et al. 2023).

To address the challenges mentioned above, we incorporate identity information into the three stages of 3D head generation compared with previous methods in Figure 2. Firstly, we extract Identity embedding using an ID feature extractor and align them with text embedding using a pre-trained ID adapter with decoupled cross-attention. This, along with landmark-guided ControlNet, enables ID diffusion and provides **ID-aware guidance**, allowing us to incorporate the identity and facial layout information from the reference image into the geometry and texture generation stages. Secondly, we propose **ID-aware geometry supervision**. By extracting canny maps containing detailed information from the portrait image, we combine canny-guided ControlNet and ID features with off-the-shelf text-to-normal and text-to-depth methods (Huang et al. 2023) to generate high-quality geometric constraint for the geometry generation stage. Additionally, we obtain **ID-aware initialization** using 3D GAN inversion. In the geometry sculpting stage, we optimize the 3D head using ID-aware guidance with ID-aware Score Distillation (ISD) loss. In the texture generation stage, we first use Progressive Texture Inpainting to generate a UV-textured map for the uncolored 3D head mesh. Then we leverage ID diffusion to denoise the rendered image from random viewpoints and further refine the texture by calculating the loss between the rendered and denoised images.

In this paper, we propose a new method for generating high-quality 3D heads from single in-the-wild portrait images. The main contributions of this paper are as follows:

- A novel pipeline for generating high-quality 3D head models from a single in-the-wild face image.
- A method that incorporates identity information into the geometry initialization, geometry generation, and texture generation stages, significantly improving the identity details of the generated 3D heads.
- An ID Consistent Texture Inpainting and Refinement method, which uses Progressive Texture Inpainting and refinement method to achieve realistic texture generation.

2 Related Works

2.1 Text to 3D Generation

Recent advances in image generative models make it possible to synthesize diverse high-fidelity 3D objects from text prompts. Efforts such as DreamFusion (Poole et al. 2022) explore text-to-3D generation by leveraging a Score Distillation Sampling (SDS) loss shows impressive results. Subsequently, Magic3D (Lin et al. 2023) increases the rendering resolution of generated 3D objects. Fantasia3D (Chen et al. 2023) models the appearance via the BRDF modeling. ProlificDreamer (Wang et al. 2024) proposes Variational Score Distillation (VSD) to mitigate oversaturation problems. Although many works have solved the problem of texture oversaturation to some extent, it is still difficult to generate realistic geometry and texture.

2.2 Image to 3D Generation

Image to 3D generation has also developed rapidly. Zero-1-to-3 (Liu et al. 2023b) first proposes view-dependent diffusion, which focuses on learning camera viewpoints within]diffusion models, enabling zero-shot novel view synthesis. GeNVS (Chan et al. 2023) leverages pixel-aligned features from input views to add 3D awareness to 2D diffusion models. On the other hand, some optimize-based image-to-3D methods, such as Magic123 (Qian et al. 2023) and Make-it-3D (Tang et al. 2023), incorporate 3D priors from Zero-1-to-3 and employ coarse-to-fine two-stage optimization. Subsequent works (Sun et al. 2023; Yu et al. 2023) further refining and enhancing the quality of generated 3D models. However, these methods are limited by the ability of view-dependent diffusion, which cannot accurately generate the geometry from the image.

2.3 3D Head Generation

Recently, various approaches (Grassal et al. 2022; Gafni et al. 2021; Zheng et al. 2023b; Xu et al. 2023) have explored reconstructing 3D head avatars from monocular face videos using different 3D representations. However, these methods necessitate several minutes of video input and struggle to reconstruct less prominent parts in videos. In addition, methods generating heads from 3D GANs like EG3D (Chan et al. 2022) proposes a tri-plane representation that can efficiently

render high-quality facial images. Panohead (An et al. 2023) advances EG3D by enabling full-head 3D face generation. Rodin (Wang et al. 2023) uses diffusion models to generate 3D heads. Nonetheless, they are constrained by the inherent limitations of implicit representations and cannot produce high-quality surface geometry.

Text-based human head generation methods like Headsculpt (Han et al. 2024) and HeadArtist (Liu et al. 2023a) integrate facial landmarks to solve layout issues like the Janus face problem. Humannorm (Huang et al. 2023) refines diffusion models using 3D human body data for better normal and depth guidance. Despite these, texture saturation in these methods can impede the attainment of photo-realistic results. In contrast, our method incorporates identity information in all generation stages, which can generate 3D head with great texture realism and identity consistency.

3 Method

In this paper, we propose ID-Sculpt, a novel optimization-based method for high-fidelity 3D head generation from a single portrait image. The overall pipeline of ID-Sculpt is shown in Figure 3. We fully leverage identity-aware priors to our generation process through **ID-aware Initialization and Guidance** (§3.1). The generation process consists of two stages: **Geometry Sculpting** (§3.2) and **ID-Consistent Texture Inpainting and Refinement** (§3.3).

3.1 ID-aware Initialization and Guidance

Given a facial image $\boldsymbol{x}_{\text{ref}}$ as input, our goal is to generate a high-fidelity 3D head model parameterized with θ, that preserves the identity and appearance of the portrait image. We adopt an optimization-based pipeline based on the pretrained text-to-image diffusion model $p_t(\boldsymbol{x}_t|y)$, and formulate the optimization process for generating a 3D head as follows:

$$\begin{aligned} &\min \begin{cases} \mathbb{E}_{t,c}[\mathrm{D}_{\mathrm{KL}}(q_t^\theta(\boldsymbol{x}_t|\mathbf{c})\|p_t(\boldsymbol{x}_t|y))], & \mathbf{c}=\mathbf{c}_{\text{rand}} \\ D(\boldsymbol{x}_{\text{ref}},\ g(\theta,\mathbf{c})), & \mathbf{c}=\mathbf{c}_{\text{ref}} \end{cases} \\ &\text{s.t. } \theta=\{\theta|\boldsymbol{x}_{\text{ref}},\ \theta_0\}, \quad \theta_0=\text{Initial Mesh}, \end{aligned} \tag{1}$$

where $p_t(\boldsymbol{x}_t|y)$ is the pretrained text-to-image diffusion model and $q_t^\theta(\boldsymbol{x}_t|\mathbf{c})$ is the distribution at time t of the forward diffusion process starting from the rendered image $g(\theta,\mathbf{c})$ with the camera $\mathbf{c}$. $\mathbf{c}_{\text{ref}}$ is the reference camera pose. $D(\cdot)$ denotes the distance function, and θ_0 is the initial 3D mesh provided by users.

For 3D full head generation task, the direct description of the input facial image, often cannot fully capture the variations of different facial features. Besides, pretrained 2D view-dependent diffusion models (*e.g.,* Zero-123 (Liu et al. 2023b)) are trained on general objects, which cannot generalize to human heads. Therefore, our key insight is to leverage identity knowledge and inject state-of-the-art identity priors into the mesh initialization and diffusion guidance.

ID-aware Initialization. The initialization is crucial for 3D generation. According to the optimization goal in Eq.(1), a good initial 3D mesh θ_0 should possess the approximate geometric shape of the face in the given image $\boldsymbol{x}_{\text{ref}}$. Some approach uses FLAME (Feng et al. 2021) to obtain the initial 3D head shape. However, this approach cannot generate personalized facial features and does not describe the geometric shape of hair. To address this issue, we employ a 3D-aware GAN model (An et al. 2023) trained on a comprehensive 360° head dataset to obtain the initial 3D head mesh θ_0. Specifically, given the input facial image $\boldsymbol{x}_{\text{ref}}$, we first perform inversion to find latent code z and then perform Pivotal Tuning Inversion (PTI) (Roich et al. 2022) to obtain a 3D representation that better matches the input image. Through 3D GAN inversion, we can obtain an initial 3D head mesh that corresponds to the input facial image, denoted as:

$$\theta_0 = \text{3D-GAN}(z_0|\boldsymbol{x}_{\text{ref}}). \tag{2}$$

ID-aware Guidance. In order to obtain finer information about the facial image, we extract high-dimensional identity features I_r from the reference image as additional conditions beyond text prompt, to guide the denoising process of diffusion models. We use a facial recognition model (Deng et al. 2019) trained on large-scale facial datasets as the identity encoder to extract robust identity features. Additionally, we utilize a light projection network to map the extracted ID feature to an enhanced ID feature f_{ID} aligned with text embedding, and insert the ID feature into diffusion with decoupled cross-attention (Ye et al. 2023). In this way, we construct an ID-guided diffusion model p^{id}, which will be utilized in our ID-aware Score Distillation loss to provide identity guidance for a specific person.

Identity-enhanced prompt only extracts general identity information, but cannot provide the information of 3D facial layout. For example, a person has different mouth shapes when laughing and crying, but the identity features are the same. To address this issue, we utilize a pretrained facial shape predictor (Guo et al. 2020) to obtain 3D facial landmarks of the reference image $\boldsymbol{x}_{\text{ref}}$, and align the 3D facial landmarks to the initial 3D head mesh.

Given a camera pose $\boldsymbol{c}$, we project the extracted 3D facial landmarks onto the 2D plane, and obtain a facial landmark image $L(\mathbf{c})$, which provides the facial component layout information from that viewpoint. We then apply a landmark-guided ControlNet (Zhang, Rao, and Agrawala 2023) to introduce the facial layout information into diffusion guidance.

By incorporating three types of information, *i.e.,* text prompt y (obtained from $\boldsymbol{x}_{\text{ref}}$ by BLIP-2 (Li et al. 2023)), identity feature f_{ID}, and the projected facial landmark image $L(\mathbf{c})$, we propose **ID-aware Score Distillation (ISD)** loss, which can be formulated as:

$$\mathcal{L}_{\text{ISD}} = \mathbb{E}_{t,c}[D_{\text{KL}}(q_t^\theta(\boldsymbol{x}_t|\mathbf{c})\|p_t^{\text{id}}(\boldsymbol{x}_t|y, f_{ID}, L(\mathbf{c}), \mathbf{c}))]. \tag{3}$$

Compared to SDS, ISD introduces more ID-preserved supervision that is aligned with the reference portrait image.

ID-aware Geometry Supervision. In the geometry sculpting stage, the reference image $\boldsymbol{x}_{\text{ref}}$ in RGB space cannot be directly used as the supervision for geometry. A common method is to use normal maps and depth maps that contain geometric information for supervision, which can be predicted based on the reference image in RGB space. Therefore, it is crucial to obtain high-quality geometric representations $\boldsymbol{X}_{\text{ref}}^{\text{geo}}$ from the reference image in RGB space.

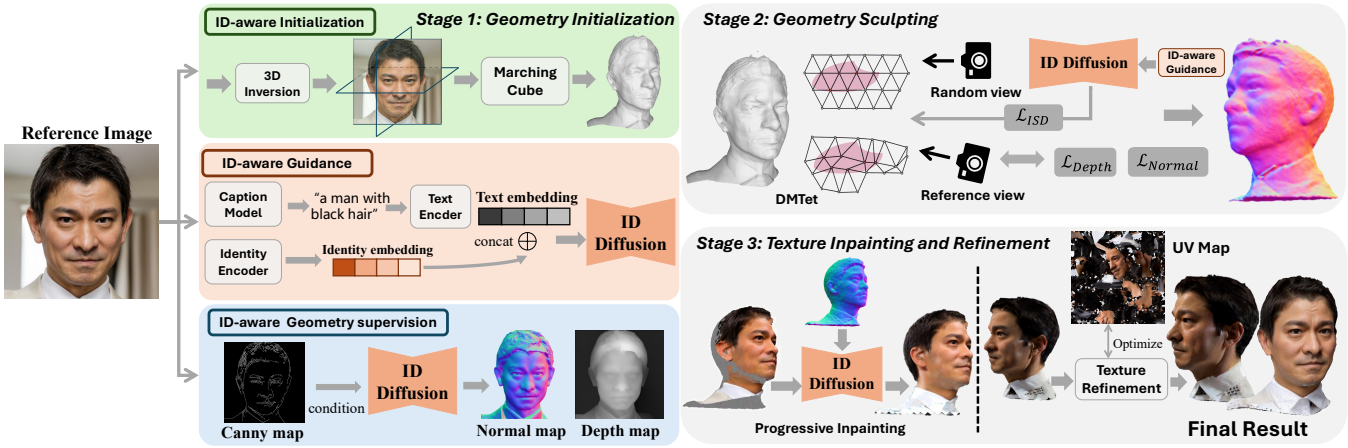

Figure 3: Overview of our proposed ID-Sculpt framework. Given a reference image, ID-Sculpt leverages 3D GAN inversion for improved identity-related geometry initialization in stage 1. In stage 2, ID-aware guidance and ID-aware geometry supervision are used to intergrate identity information into the geometry sculpting process. In stage 3, ID Consistent Texture Inpainting and Refinement is applied to generate a high-quality head texture, where we first use the inpainting method to generate a rough texture and then use image-level ID-aware supervision for texture refinement. With these methods, we can generate high-quality 3D head models with consistent identities from a single in-the-wild face image.

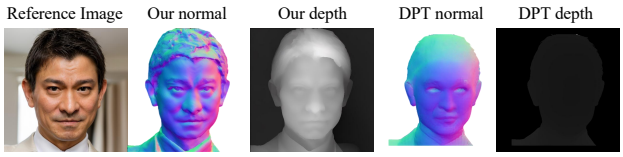

Figure 4: Comparison between our normal map and depth map and those generated by DPT(Ranftl, Bochkovskiy, and Koltun 2021; Ranftl et al. 2020).

Previous methods (Eftekhar et al. 2021) directly predicted depth maps and normal maps from images but failed to capture some geometric details in the reference images, which could result in flat geometries, as shown in Fig 4. Different from previous methods, we first use an edge extractor to extract high-frequency facial geometric details and obtain a Canny image as the condition. Then we utilize pre-trained Text-to-Normal and Text-to-Depth diffusion models (Huang et al. 2023) with ControlNet to generate normal maps and depth maps. Additionally, to better preserve identity information, we also add ID features into the diffusion process. Through these approaches, high-quality geometric supervision in the reference view can be formulated as:

$$\boldsymbol{X}_{\text{ref}}^{\text{geo}} = p_{\text{geo}}^{\text{id}}(\boldsymbol{x}_t|y, f_{\text{ID}}, \text{Canny}(\boldsymbol{x}_{\text{ref}})), \tag{4}$$

where $\boldsymbol{X}_{\text{ref}}^{\text{geo}} = \{\boldsymbol{X}_{\text{ref}}^{\text{normal}}, \boldsymbol{X}_{\text{ref}}^{\text{depth}}\}$ and $p_{\text{geo}}^{\text{id}} = \{p_{\text{normal}}^{\text{id}}, p_{\text{depth}}^{\text{id}}\}$. In this way, the 3D head model can obtain geometric details that are as rich as the information in the RGB image of the reference view.

3.2 Geometry Sculpting

During the geometry sculpting stage, we aim to obtain a fine-grained 3D geometry, that is consistent with the input image under the reference view, and consistent across different viewpoints. We first convert the initial 3D mesh obtained from ID-aware initialization into DMTet (Shen et al. 2021) representation to facilitate high-resolution details and high-quality surface. Specifically, we use the normals rendered from 3D-GAN as the pseudo ground truth normals $\mathcal{N}_{\mathbf{c}}$. For the DMTet object to be optimized, we use Marching Tetrahedra to generate the normal map $\hat{\mathcal{N}}_{\mathbf{c}}$. To polish the overall surface from multiple viewpoints, we uniformly sample camera poses $\mathbf{c}$ distributed in the 3D-GAN space, and use L_2 distance between normal maps as the loss function, defined as $\mathcal{L}_{\text{norm}} = \|\hat{\mathcal{N}}_{\mathbf{c}} - \mathcal{N}_{\mathbf{c}}\|_2$.

Geometry Refinement. During the geometry refinement stage, we use **ID-aware Geometry Supervision** to ensure the generation of fine geometry under the reference view, while supplementing the information with the prior of the diffusion model under other views, as shown in Eq.(1). Under the reference view, we calculate the reference loss from the head foreground mask, normal map, and depth map of the reference image, which is defined as:

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{depth}}, \tag{5}$$

where the foreground mask M is obtained through face parsing, and both the mask loss L_{mask} and normal loss L_{normal} are defined as the $\mathcal{L}_2$ distance between the rendered mask and normal map, and the reference mask and normal map. For the depth loss L_{depth}, we use the negative Pearson correlation to alleviate the problem of depth scale mismatch.

For other views that are randomly sampled around the 3D head, we use the ISD loss L_{ISD} in Eq.(3) to guide the optimization of the geometry, based on the ID-enhanced diffusion prior. Finally, the Marching Tetrahedra algorithm is applied to extract the mesh.

3.3 ID-Consistent Texture Inpainting and Refinement

During the texture generation stage, our goal is to rapidly generate realistic textures for the 3D head mesh. Previous textures generated from SDS/VSD loss (Chen et al. 2023; Wang et al. 2024) may become over-saturated and inconsistent with the color tone of the reference image. Therefore, we design an inpainting and refinement process to quickly produce high-quality UV texture maps.

Progressive Texture Inpainting. With an uncolored 3D head surface mesh M with vertices V and triangular faces F generated in the geometry generation stage, denoted as $M = (V, F)$, we initialize a UV texture map $\mathbf{T}_0$ corresponding to the mesh coordinates. Starting from the reference view $\mathbf{c}_0$, we first back-project the reference portrait image onto the 3D mesh. To maintain consistency with the reference image texture, we sample a camera trajectory $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n\}$ (where $\mathbf{c}_i$ is the i-th viewpoint, and n is the number of sampled viewpoints) to progressively generate texture. The camera trajectory starts from the reference image viewpoint, gradually moves the camera pose to the back view of the head, and finally generates the texture for the top of the head that has not been generated yet.

During the inpainting process, at a specific viewpoint $\mathbf{c}_k$, given the texture map $\mathbf{T}_{k-1}$ that has been textured in $k-1$ previous viewpoints, we can render an RGB image I_k with partially colored regions, as well as a mask $\mathbf{m}_{k-1}$ indicating the colored (visible) regions. To better inpaint the uncolored areas, we render the normal map $\boldsymbol{x}_k^{normal}$ from the 3D mesh as the geometric condition at the $\mathbf{c}_k$ viewpoint, and then use the normal-conditioned diffusion model to generate the texture for the uncolored areas, which is formulated as:

$$\hat{I}_k = p_0^{\text{id}}(\boldsymbol{x}_0 | y, f_{ID}, \boldsymbol{x}_k^{normal}). \tag{6}$$

The generated image $\hat{I}_k$ at the viewpoint $\mathbf{c}_k$ will be projected back onto the 3D mesh, resulting in the inpainted texture map $\hat{\mathbf{T}}_k$. And the uncolored region $(1 - \mathbf{m}_{k-1})$ in UV texture $\mathbf{T}_{k-1}$ is updated as follows:

$$\mathbf{T}_k = \mathbf{m}_{k-1} \odot \mathbf{T}_{k-1} + (1 - \mathbf{m}_{k-1}) \odot \hat{\mathbf{T}}_k, \tag{7}$$

where $\mathbf{T}_k$ is the updated UV texture map after k step inpainting. With this progressive inpainting method, we can gradually inpaint the textures from other viewpoints starting from the reference portrait image.

Texture Refinement. In the texture refinement stage, we aim to solve the texture inconsistency problem that appears in texture inpainting and further enhance the detail and realism of the generated texture.

We optimize the UV texture map by utilizing the ID-guided diffusion model and minimizing the distance between the rendered images and the refined images.

Specifically, for a rendered image $\boldsymbol{x}_0$ from a sampled viewpoint, we add Gaussian noise with t diffusion steps and use the ID-guided diffusion model p^{id} to restore it. Moreover, we add rendered normals $\boldsymbol{x}^{normal}$ from the corresponding viewpoint as additional geometric information conditions during the diffusion denoising process. This yields a refined image $\hat{\boldsymbol{x}}_0$ in the respective viewpoint which we will use as supervision to optimize the texture map:

$$\hat{\boldsymbol{x}}_0 = p_0^{\text{id}}(\boldsymbol{x}_0 | \boldsymbol{x}_t, y, f_{ID}, \boldsymbol{x}^{normal}). \tag{8}$$

After obtaining the refined image $\hat{\boldsymbol{x}}_0$, we calculate the pixel-level MSE loss $\mathcal{L}_{\text{MSE}}$ between $\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0$ to further refine the texture map. Besides, we add a perceptual loss $\mathcal{L}_{\text{percep}}$ to enhance the texture detail information and maintain style similarity. We also calculate the MSE loss $\mathcal{L}_{\text{MSE}}^{\text{ref}}$ and perceptual loss $\mathcal{L}_{\text{percep}}^{\text{ref}}$ between x_0 from the reference view and the reference image $\boldsymbol{x}_{\text{ref}}$. By combining these supervisions, we obtain the final loss $\mathcal{L}_{\text{refine}}$ for the texture refinement stage:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{percep}} + \mathcal{L}_{\text{MSE}}^{\text{ref}} + \mathcal{L}_{\text{percep}}^{\text{ref}}. \tag{9}$$

4 Experiments

In the experimental section, we start with the implementation details, followed by quantitative and qualitative comparisons between our method and others. Finally, we analyze the role of each module in our method.

4.1 Implementation Details

Our ID-Sculpt is built upon the open-source project Threestudio (Guo et al. 2023). In the geometry stage, we use SD1.5 (Rombach et al. 2022) as the base diffusion model. During the texture generation stage, we employ Realistic Vision 4.0 as the base diffusion model. In the geometry generation stage, we perform 5,000 optimization iterations. During the texture generation stage, we first perform texture inpainting from 15 ordered viewpoints, followed by 400 steps of texture refinement. Our experiments are conducted on a single V100 GPU, with a batch size set to 1. For each 3D head, the total optimization time is approximately 40 minutes.

Baselines. We conducted extensive comparisons with four other single-view 3D generation methods (Magic123 (Qian et al. 2023), DreamCraft3D(Sun et al. 2023), Wonder3D (Long et al. 2023)) and a single-view human head generation method (Panohead (An et al. 2023)). For Magic123, DreamCraft3D, and Wonder3D, we generated the 3D models strictly following the official open-source code and parameters. For Panohead, we followed the official guidance and obtained results using the GAN inversion method. In addition, we compared our method with the state-of-the-art text-to-3D human head generation methods HumanNorm (Huang et al. 2023), for it has excellent generation results and good reproducibility.

Datasets. We conducted experiments on a subset of facial images from the FFHQ (Karras, Laine, and Aila 2019) dataset, a high-quality in-the-wild facial dataset known for its rich diversity in age, ethnicity, and image backgrounds, as well as significant variations in facial attributes. This dataset presents a challenging task due to its unconstrained environment. From this dataset, We selected 106 portrait images with unoccluded faces to evaluate our model and baselines.

Evaluation metrics. We employ LPIPS (Zhang et al. 2018) for fidelity measurement, CLIP-I (Ye et al. 2023) score and ID (An et al. 2023) score to measure the multi-view consistency and identity preservation.

Figure 5: Qualitative evaluation with image-to-3D methods. Our method surpasses previous approaches in terms of the facial detail in head geometry as well as the realism of the texture. The zoom-in result of eye/mouth region and facial geometry shows that our method has clearer structural details and more reasonable geometric structure.

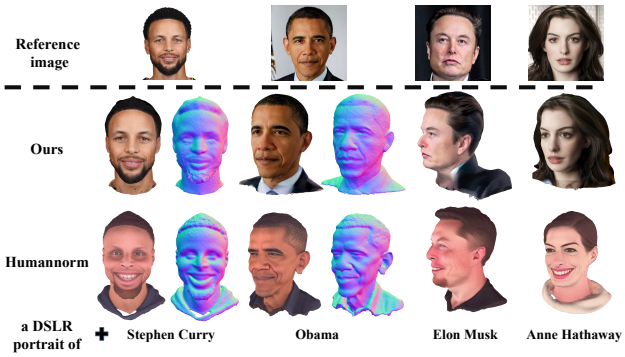

Figure 6: Qualitative evaluation with text-to-head method. Our method has a clear advantage in the realism.

4.2 Qualitative Evaluation

Qualitative evaluation with image-to-3D methods. In Fig. 5 we show the comparisons between our method and four other image-to-3D methods. These four images are selected from the in-the-wild face image dataset FFHQ and include faces of different ages and genders. For the geometry, the 3D generation methods based on optimization (Magic123 and Dreamcraft3D) can't restore the stereoscopic shape of the human head. Wonder3D can generate basic human head shapes but generally suffers from overly flat facial features in side views, and the facial surface lacks geometry detail. Besides, Panohead produces strange protrusions in the eye area. In contrast, our method, after introducing the identity information to the geometry sculpting stage, can restore fine geometry on the face while ensuring a high-quality overall head shape. For texture results, Dreamcraft3D and Wonder3D cannot generate images of the head from other angles. Wonder3D's texture quality drops significantly on the side (i.e., the side rear view of the head in the first row of Wonder3D) and cannot produce high-resolution textures. Panohead struggles to generate high-fidelity structural details in areas like the eye and mouth region. After using ID-enhanced inpainting and refinement processes, our method can generate more photorealistic textures. In summary, our method can produce high-quality head geometry and realistic textures while fully restoring the facial information from the reference images.

Qualitative evaluation with text-to-head method. In Fig. 6, we show a comparison of our method with the current SOTA text-to-head method HumanNorm. Since text description cannot fully describe in-the-wild face information, we choose examples of celebrities so that the text-to-head method can accurately generate heads based on the celebrity names. For generated geometry, HumanNorm tends to generate heads with exaggerated structures, whereas our method generates far more realistic geometries. In terms of texture, the heads generated by HumanNorm have over-saturated colors. In contrast, our method fully utilizes both the reference image information and the identity information of the person depicted to generate more realistic faces.

4.3 Quantitative Evaluation

For each generated head model, we render images from four consistent viewpoints: left-frontal, left, right-frontal, and right. In Tab. 1, we present the quantitative comparisons

Methods	LPIPS ↓	CLIP-I ↑	ID ↑	User Study ↑
Magic123	0.367	0.4447	0.0245	1.11
PanoHead	0.045	0.6600	0.2737	4.09
Wonder3D	0.135	**0.6860**	0.2729	3.13
DreamCraft3D	0.068	0.6074	0.2526	1.95
ID-Sculpt (Ours)	**0.039**	0.6605	**0.4451**	**4.72**

Table 1: Quantitative comparisons on head generation task. We compute LPIPS under the reference view, CIIP-I and ID score under novel views.

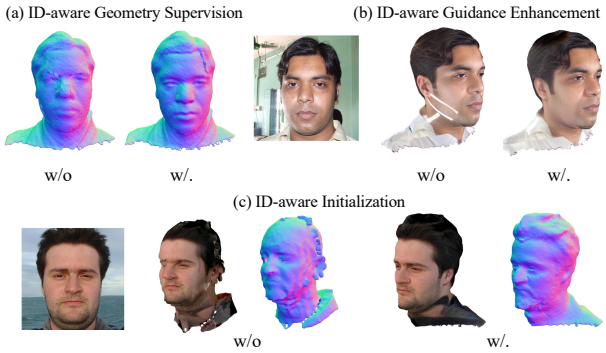

Figure 7: Visual ablation results of our ID-aware Initialization and Guidance.

with the previous SOTA methods. Our method outperforms previous approaches in terms of the ID metric and LPIPS score, indicating its effectiveness in capturing facial features and demonstrating high identity consistency across various perspectives in the analysis of novel viewpoints.

User Study. To validate our model's robustness and quality, we conducted a user study with 15 random examples. Participants ranked five methods' outputs based on quality from two random viewpoints. Our model ranked highest on average across 480 responses from 32 participants.

4.4 Ablation Study

Effectiveness of ID-aware Initialization and Guidance. We present the visual result in Fig.7 result to show the effectiveness of our ID-aware Initialization and Guidance method. (a) By replacing the normal map obtained through ID-aware Geometry Supervision with the normal map extracted from DPT(Ranftl, Bochkovskiy, and Koltun 2021; Ranftl et al. 2020), we can observe that without ID-aware Geometry Supervision, the generated geometry loses many facial details. (b) When removing the ID-aware Guidance Enhancement in the texture generation stage, we can see that the faces lose the constraints imposed by identity information and easily produce artifacts for large angles (e.g., side faces) without sufficient reference image information. (c) We compare our ID-aware Initialization results by using Flame as the initialization geometry. As Flame aligns only with the facial landmarks, the head portions other than the face in the images are difficult to align with the initialization geometry, leading to difficulty in geometry optimization.

Effectiveness of ID-Consistent Texture Inpainting and Refinement. As shown in Fig.8, without (a) Texture refinement=, relying solely on inpainting can lead to discontinuous boundaries between each inpainting step, which is particularly noticeable at the intersection of the ears and hair. Without (b) Progressive Inpainting to complete the texture and instead use a view-surrounding approach, incoherent texture colors will be generated (e.g. yellow part at the neck).

Figure 8: Visual ablation results of our ID-Consistent Texture Inpainting and Refinement. Our method effectively reduces the artifacts in texture generation.

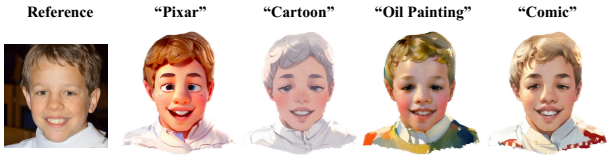

Figure 9: 3D head with different styles. With the stylized diffusion model, we can achieve diversified stylized 3D head generation while maintaining identity.

4.5 Application

ID-consistant Texture Stylization. With the capability of stylized text-to-image diffusion models, we achieve stylized 3D head generation. Due to the use of ID-aware guidance in the texture generation stage, the generated stylized heads can maintain the identity information of the character in the portrait image, as shown in Fig 9.

5 Conclusion

We propose ID-Sculpt, a novel method to generate a 3D head from a single in-the-wild portrait image. Our method significantly improves ID consistency and geometric details by adding identity knowledge to the initialization, guidance, and geometry supervision part of the optimization process. Moreover, we design a two-stage ID-consistent texture inpainting and refinement method, which leads to coherent and photo-realistic head texture. ID-Sculpt outperforms previous methods in both geometry and texture quality, demonstrating the ability to generate high-fidelity 3D heads. These advantages combined with stylized 3D head generation bring great help for personalized 3D head asset generation.

Limitations and future work. Although we can generate high-quality 3D heads from a portrait image, the optimization-based approach takes the whole generation a long time (40 minutes). In the future, we will work on accelerating the speed of 3D head generation.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 72192821, 62302297, 62472282), Shanghai Sailing Program (22YF1420300), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), the Fundamental Research Funds for the Central Universities (project number: YG2023QNA35).

References

An, S.; Xu, H.; Shi, Y.; Song, G.; Ogras, U. Y.; and Luo, L. 2023. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20950–20959.

Bai, Z.; Tan, F.; Huang, Z.; Sarkar, K.; Tang, D.; Qiu, D.; Meka, A.; Du, R.; Dou, M.; Orts-Escolano, S.; et al. 2023. Learning Personalized High Quality Volumetric Head Avatars from Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16890–16900.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Chan, E. R.; Nagano, K.; Chan, M. A.; Bergman, A. W.; Park, J. J.; Levy, A.; Aittala, M.; De Mello, S.; Karras, T.; and Wetzstein, G. 2023. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4217–4229.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22246–22256.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Eftekhar, A.; Sax, A.; Malik, J.; and Zamir, A. 2021. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10786–10796.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2022. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18653–18664.

Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Guo, Y.-C.; Liu, Y.-T.; Shao, R.; Laforte, C.; Voleti, V.; Luo, G.; Chen, C.-H.; Zou, Z.-X.; Wang, C.; Cao, Y.-P.; and Zhang, S.-H. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

Han, X.; Cao, Y.; Han, K.; Zhu, X.; Deng, J.; Song, Y.-Z.; Xiang, T.; and Wong, K.-Y. K. 2024. Headsculpt: Crafting 3d head avatars with text. *Advances in Neural Information Processing Systems*, 36.

Han, Y.; Zhang, J.; Zhu, J.; Li, X.; Ge, Y.; Li, W.; Wang, C.; Liu, Y.; Liu, X.; and Tai, Y. 2023. A Generalist FaceX via Learning Unified Facial Representation. *arXiv preprint arXiv:2401.00551*.

Huang, X.; Shao, R.; Zhang, Q.; Zhang, H.; Feng, Y.; Liu, Y.; and Wang, Q. 2023. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*.

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

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4): 1–14.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, H.; Wang, X.; Wan, Z.; Shen, Y.; Song, Y.; Liao, J.; and Chen, Q. 2023a. HeadArtist: Text-conditioned 3D Head Generation with Self Score Distillation. *arXiv preprint arXiv:2312.07539*.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023b. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9298–9309.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023c. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2023. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*.

Melas-Kyriazi, L.; Laina, I.; Rupprecht, C.; and Vedaldi, A. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8446–8455.

Munkberg, J.; Hasselgren, J.; Shen, T.; Gao, J.; Chen, W.; Evans, A.; Müller, T.; and Fidler, S. 2022. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8280–8290.

Pei, G.; Zhang, J.; Hu, M.; Zhai, G.; Wang, C.; Zhang, Z.; Yang, J.; Shen, C.; and Tao, D. 2024. Deepfake Generation and Detection: A Benchmark and Survey. *arXiv preprint arXiv:2403.17881*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; et al. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision Transformers for Dense Prediction. *ArXiv preprint*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1): 1–13.

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 conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sanghi, A.; Chu, H.; Lambourne, J. G.; Wang, Y.; Cheng, C.-Y.; Fumero, M.; and Malekshan, K. R. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18603–18613.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6087–6101.

Sun, J.; Zhang, B.; Shao, R.; Wang, L.; Liu, W.; Xie, Z.; and Liu, Y. 2023. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*.

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22819–22829.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

Wang, T.; Zhang, B.; Zhang, T.; Gu, S.; Bao, J.; Baltrusaitis, T.; Shen, J.; Chen, D.; Wen, F.; Chen, Q.; et al. 2023. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4563–4573.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Xu, Y.; Chen, B.; Li, Z.; Zhang, H.; Wang, L.; Zheng, Z.; and Liu, Y. 2023. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. *arXiv preprint arXiv:2312.03029*.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, W.; Yuan, L.; Cao, Y.-P.; Gao, X.; Li, X.; Quan, L.; Shan, Y.; and Tian, Y. 2023. Hifi-123: Towards high-fidelity one image to 3d content generation. *arXiv preprint arXiv:2310.06744*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, M.; Zhang, H.; Yang, H.; and Huang, D. 2023a. NeuFace: Realistic 3D Neural Face Rendering From Multi-View Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16868–16877.

Zheng, Y.; Yifan, W.; Wetzstein, G.; Black, M. J.; and Hilliges, O. 2023b. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21057–21067.