From Blurry to Believable: Enhancing Low-quality Talking Heads with 3D Generative Priors

Ding-Jiun Huang[1], Yuanhao Wang[1], Shao-Ji Yuan[1], Albert Mosella-Montoro[1], Francisco Vicente Carrasco[1], Cheng Zhang[2], Fernando de la Torre[1]
[1] Carnegie Mellon University [2] Texas A&M University

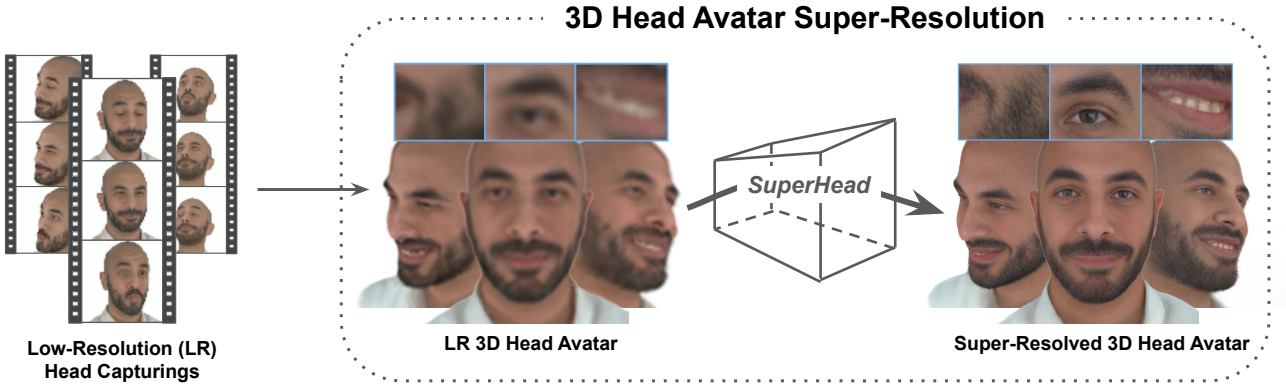

Figure 1. We present SuperHead, a new method for super-resolving low-resolution 3D head avatars. Given an low-resolution animatable avatar reconstructed from low-quality captures, SuperHead synthesizes high-fidelity geometry and detailed textures while ensuring multi-view and temporal consistency under diverse facial expressions. Unlike prior image- or video-based SR approaches, our method directly upsamples 3D avatars, enabling photorealistic animation and faithful identity preservation from degraded inputs.

Abstract

Creating high-fidelity, animatable 3D talking heads is crucial for immersive applications, yet often hindered by the prevalence of low-quality image or video sources, which yield poor 3D reconstructions. In this paper, we introduce SuperHead, a novel framework for enhancing low-resolution, animatable 3D head avatars. The core challenge lies in synthesizing high-quality geometry and textures, while ensuring both 3D and temporal consistency during animation and preserving subject identity. Despite recent progress in image, video and 3D-based super-resolution (SR), existing SR techniques are ill-equipped to handle dynamic 3D inputs. To address this, SuperHead leverages the rich priors from pre-trained 3D generative models via a novel dynamics-aware 3D inversion scheme. This process optimizes the latent representation of the generative model to produce a super-resolved 3D Gaussian Splatting (3DGS) head model, which is subsequently bound to an underlying parametric head model for animation. The inversion is jointly supervised using a sparse collection of upscaled 2D face renderings and corresponding depth maps, captured from diverse facial expressions and camera viewpoints, to ensure realism under dynamic facial motions. Experiments demonstrate that SuperHead generates avatars with fine-grained facial details under dynamic motions, significantly outperforming baseline methods in visual quality. The code will be made publicly available.

1. Introduction

High-fidelity, animatable 3D head avatars are increasingly important in augmented and virtual reality, telepresence, gaming, and digital entertainment, where realism is critical. However, capturing scan-quality geometry and dynamic textures with traditional pipelines requires costly, specialized equipment [3, 5, 8, 17], limiting its accessibility. As a result, much recent work reconstructs dynamic 3D heads from videos or images acquired with consumer devices [11, 16, 18, 22, 23, 31, 40, 42, 50, 51, 63, 65]. Yet such inputs often suffer from low resolution, poor lighting, and motion blur, producing avatars with blurry textures and artifacts that fall short of high-fidelity requirements.

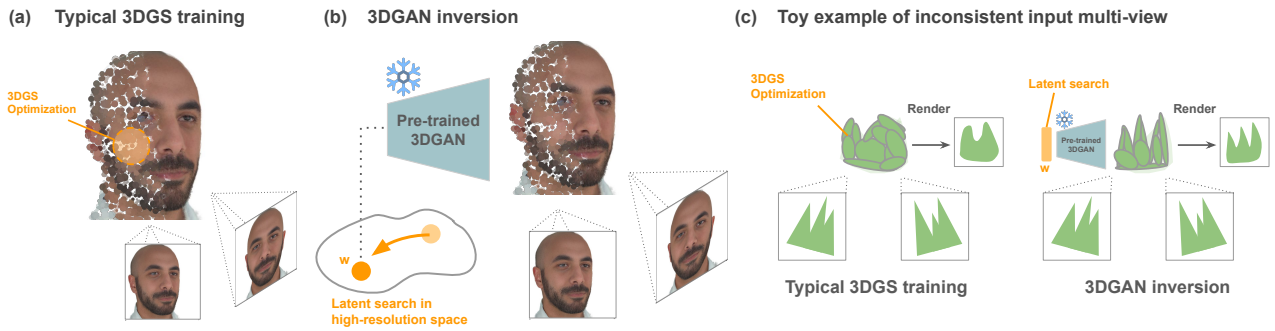

Figure 2. (a) Typical 3DGS process optimizes a head from multi-view images. (b) 3D GAN inversion aims to find a latent code whose generated 3D head best explains the given multi-view images. (c) When input multi-views are inconsistent, typical 3DGS tends to create an "averaged" result as a compromise among inconsistent inputs. 3D GAN inversion can generate high-frequency details because it searches in a pre-trained high-resolution space.

In this paper, we focus on enhancing the visual quality of low-resolution animatable head avatars, that is, 3D talking head models whose textures appear blurry and lack fine facial details, often as a result of reconstruction from low-quality input sources. While significant progress has been made in image, video, and static 3D super-resolution (SR), extending these advances to dynamic 3D heads presents unique challenges. Unlike static SR, this task demands simultaneously generating high-resolution geometry and textures, preserving temporal consistency across diverse expressions, and faithfully maintaining identity. Existing 3D SR approaches [21, 24, 39, 43, 47, 58] often rely on 2D priors, but applying them frame-by-frame leads to flickering, and video SR cannot ensure cross-view consistency, making them ill-suited for dynamic avatars.

Recent advances in 3D-aware generative models, particularly 3D GANs [4, 57, 59], have demonstrated strong 3D priors for reconstructing head avatars via GAN inversion. By projecting input images into the latent space of a 3D GAN, they can synthesize high-quality 3D heads from limited data. Inspired by this paradigm, we propose to extend 3D GAN inversion to the super-resolution of animatable 3D avatars. In Figure 2, we show the difference between typical 3DGS head reconstruction and 3D GAN inversion. We also provide an example to illustrate that 3D GAN inversion can well handle inconsistent multi-view. However, unlike static inversion, directly applying inversion to animatable avatars is much harder — super-resolved head must stay realistic and identity-preserving across diverse expressions.

To address these challenges, we propose a dynamics-aware 3D inversion framework for upsampling low-resolution animatable head avatars. The key idea is to leverage the strong 3D priors of generative head models while tightly coupling the inversion process with the underlying mesh structure (e.g., FLAME [36]). Instead of relying on single-frame information, we condition inversion on multi-view renderings and depth cues, ensuring cross-view consistency and faithful alignment. To further tackle animation, we extend this conditioning across multiple facial expressions, enabling the reconstructed avatar to preserve identity, detail, and realism even under dynamic motion.

We name our approach **SuperHead** and demonstrate its effectiveness on different benchmarks (i.e., NeRSemble [30], INSTA [69]). SuperHead works seamlessly across different head avatar models (e.g., GaussianAvatar [40], SplattingAvatar [42]), showing clear gains in fidelity over competing methods. Our dynamics-aware design ensures that the upsampled avatars remain consistent under diverse facial motions, a scenario where existing methods often struggle. Moreover, since GAN inversion is computationally lightweight, our method achieves these improvements without sacrificing efficiency, enabling fast inference. To the best of our knowledge, this is the first framework that directly addresses the task of super-resolving low-resolution animatable 3D head avatars and obtains competitive results.

2. Related Work

Animatable 3D Head Avatar. 3D head avatar modeling has long been an active area due to its broad applications. Early works rely on multi-view stereo for geometry reconstruction [3, 5, 8, 17], but require heavy computation and complex capture setups. More recent approaches employ volumetric representations [2, 16, 18, 20, 23, 44, 55, 68] or neural implicit functions [51, 56, 62, 63, 65], offering higher efficiency and reconstruction quality, often leveraging 3D Morphable Models [6, 36] for semantic control and identity–expression disentanglement. INSTA [69] deforms query points to a canonical space using FLAME [36], while Khakhulin et al. [29] learn non-rigid deformations on FLAME templates to recover dynamic details from in-the-wild videos. Another branch of methods [11, 64] explore 3D head avatar creation with few-shot images, avoiding the need of heavy video inputs. More recently, 3D Gaussian Splatting (3DGS) [28] has advanced novel view synthesis with superior fidelity and much faster rendering, and has become widely used for head avatar modeling

[10, 13, 31, 34, 37, 40, 42, 50, 54, 64, 67]. GaussianAvatars [40] binds 3D Gaussians to FLAME mesh surface to create animatable heads; SplattingAvatar [42] offsets Gaussians along surface normals to fit more complex geometry; and SurFhead [34] rigs 2D Gaussian surfels to parametric models for well-defined geometry. Our method is complementary to these designs and can be directly applied to different types of Gaussian-based head avatars to enhance their fidelity and robustness.

3D Super-Resolution. Despite progress in 3D reconstruction and generation, enhancing low-resolution 3D content remains challenging. One line of work [21, 24, 35, 47, 58] leverages pre-trained single-image super-resolution (SISR) models to upscale images rendered from 3D representations, with additional mechanisms to improve multi-view consistency. More recently, other methods [32, 39, 43] employ pre-trained video upsamplers [7, 53] to enhance 3D content through 2D video priors, offering stronger temporal coherence but struggling with long-range consistency under large viewpoint changes. In contrast, our approach achieves dynamic 3D super-resolution by directly enforcing both multi-view and temporal consistency, producing high-fidelity avatars robust to diverse expressions and motions.

Generative Head Avatar Reconstruction. Advances in 3D-aware generative models [1, 9, 19, 25, 31, 45] have gained attention for synthesizing high-quality 3D representations and consistent multi-view images, making them strong priors for head avatar reconstruction. Among them, GSGAN [25] generate 3D Gaussian-based static head through adversarial training with large amount of 2D head images. Several works [4, 15, 33, 46, 52, 57, 59] reconstruct and edit 3D faces from a single image via GAN inversion, projecting real images into the latent space of 3D-aware GANs to enable viewpoint-consistent synthesis and semantic control. Others target animatable avatars. AniFaceGAN [49] extends a static 3D-aware generator [12] with a deformation field mapping query points to canonical space, while Lin et al. [38] use 3D GAN inversion to animate portraits. Next3D [45] combines a tri-plane representation with a conditional 3D Morphable Model (3DMM) and applies PTI inversion [41], while InvertAvatar [61] improves reconstruction through incremental inversion across multiple frames. These works show the effectiveness of pre-trained 3D GANs for one-shot or few-shot avatar reconstruction. We extend this direction by leveraging 3D generative priors to upsample low-resolution animatable head avatars.

3. Preliminary

We use 3D GAN inversion with GSGAN [25] for 3D head reconstruction and adopt a representation where avatars are modeled as 3D Gaussian splats rigged to a parametric face model. In this section, we review both components.

3.1. 3D GAN Inversion

Given a pre-trained unconditional 3D GAN model G parameterized by weight g and an input image $\mathbf{I}$, the goal of 3D GAN inversion is to find a latent code $\mathbf{z}^*$ that accurately reconstructs $\mathbf{I}$ at the corresponding camera pose, while enabling content-consistent synthesis from novel viewpoints. This problem is commonly formulated as:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \mathcal{L}_{\text{recon}}(G(\mathbf{z}, \pi; g), \mathbf{I}), \tag{1}$$

where $\mathbf{z}$ is the latent representation in $\mathcal{W}^+$ space [26, 27] of the 3D GAN model, π is the corresponding camera matrix of the input image, and $\mathcal{L}_{\text{recon}}$ is a pixel-wise reconstruction or perceptual loss. However, optimizing only in the $\mathcal{W}^+$ space often fails to capture fine-grained facial details, resulting in suboptimal reconstructions. To address this issue, recent methods (e.g., [41]) propose a second-stage inversion in the parameter space by slightly finetuning the GAN generator, and achieves improved reconstruction results. This second-stage optimization is formulated as:

$$g^* = \arg\min_{g} \mathcal{L}_{\text{recon}}(G(\mathbf{z}^*, \pi; g), \mathbf{I}) \tag{2}$$

where $\mathbf{z}^*$ is fixed from the first stage. In our framework, we thus adopt this two-stage 3D GAN inversion approach for high-fidelity reconstruction. In this paper, we define "anchor image" as all images $\mathbf{I}$ that are used for inversion.

3.2. 3D Gaussian Avatar

Following GaussianAvatars [40], an avatar is represented by 3D Gaussian primitives anchored to the surface of a parametric face model such as FLAME [36]. Each primitive is defined in the local coordinate system of a mesh face by position $\boldsymbol{\mu}'$, rotation $\mathbf{r}'$, and scaling $\mathbf{S}'$, and transformed to global coordinates as:

$$\mathbf{r} = \mathbf{R}\mathbf{r}', \quad \boldsymbol{\mu} = s\mathbf{R}\boldsymbol{\mu}' + \mathbf{T}, \quad \mathbf{S} = s\mathbf{S}', \tag{3}$$

where $\mathbf{R}$ is the orientation of the triangle face, $\mathbf{T}$ is the mean of its three vertices, and s is a scale factor. The FLAME model is controlled by shape parameters β, pose parameters θ, expression parameters ψ, and static vertex offsets δ.

4. Our Approach

Overview. Figure 3 illustrates the SuperHead framework. The input is a low-resolution (LR) animatable 3D head model H, whose movement is driven by an underlying FLAME model M. Our goal is to synthesize an animatable head avatar $\hat{H}$ that exhibits a high level of realism and fidelity in various facial expressions, while preserving the identity of the subject. We start by synthesizing a static 3D Gaussian head in canonical space through multi-view

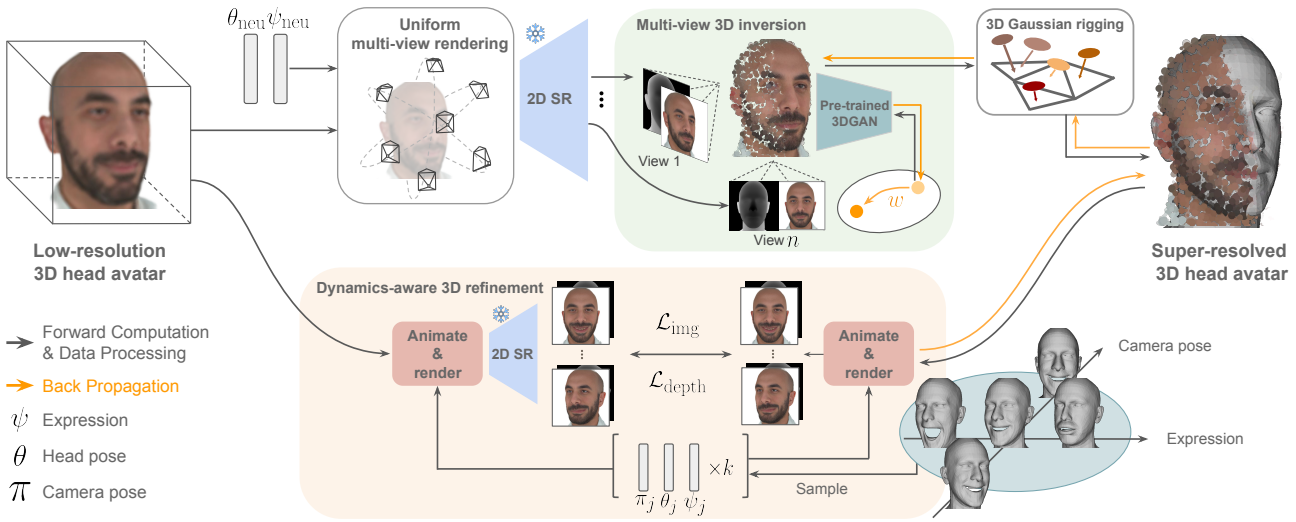

Figure 3. **Overview of SuperHead.** Given a low-resolution 3D head avatar driven by a morphable model, we first reconstruct static 3D head in the canonical space with multi-view 3D GAN inversion (Section 4.1). We then refine mesh geometry and rig 3D Gaussians onto mesh surface to enable animation (Section 4.2). We further include anchor images with diverse camera poses and expressions for dynamics-aware 3D refinement, ensuring the robustness of the 3D head model across viewing angles and complex facial motions (Section 4.3).

3D GAN inversion (Section 4.1), where multi-view renderings of a neutral expression ψ_{neu} from LR input model are sampled and super-resolved for the inversion. Then, we rig the static 3D Gaussian head to the FLAME mesh for animation (Section 4.2). Finally, to make the synthesized 3D head model robust under different facial motions, we perform dynamics-aware 3D refinement to get $\hat{H}$ by jointly optimizing 3D GAN inversion using a set of sampled and super-resolved anchor images $\{\hat{\mathbf{I}}_j\}_{j=1}^{k}$ (Section 4.3) rendered from the LR input model H with configurations $\{(\theta_j, \psi_j, \pi_j)\}_{j=1}^{k}$, where θ_j, ψ_j are the pose and expression parameters of FLAME and π_j is the camera parameter.

4.1. Multi-View 3D Inversion

Given a pre-trained 3D GAN G for generating Gaussian head avatars with parameters g, we optimize a latent code $\boldsymbol{w}^*$ such that the resulting 3DGS representation $\mathcal{G} = G(\boldsymbol{w}^*, g^*)$ can faithfully reconstructs a given set of n multi-view anchor images $\{\hat{\mathbf{I}}_i\}_{i=1}^{n}$.

Multi-view anchor images sampling. We collect $\{\mathbf{I}_i\}_{i=1}^{n}$ by sampling n multi-view renderings of a specific expression ψ_{neu} and FLAME mesh pose θ_{neu} from LR input model. The camera poses π_i of the multi-views are uniformly sampled on a sphere with radius r centered at the model. Note that we only sample views within an angle range that covers the frontal part of input head model. We manually select ψ_{neu} with a major consideration that ψ_{neu} should contain major occluded parts, e.g., teeth and eyeballs, so that G can generate these facial parts through the inversion. Then, we obtain high-resolution counterparts $\{\hat{\mathbf{I}}_i\}_{i=1}^{n}$ using an off-the-shelf image super-resolution model [66].

Multi-view 3D GAN inversion. We then reconstruct a static 3D Gaussian head in canonical space using the multi-view images $\{\hat{\mathbf{I}}_i\}_{i=1}^{n}$ and the canonical FLAME mesh $\hat{M}_{\theta_{\text{neu}},\psi_{\text{neu}}}$, defined by the neutral pose θ_{neu} and expression ψ_{neu}. The objective is to minimize the reconstruction loss:

$$\mathcal{L}_{\text{img}}^{i} = \left\| \hat{\mathbf{I}}_i' - \hat{\mathbf{I}}_i \right\|_2 + \lambda_{\text{p}} \mathcal{L}_{\text{pips}}(\hat{\mathbf{I}}_i', \hat{\mathbf{I}}_i), \tag{4}$$

where $\hat{\mathbf{I}}_i' = \mathcal{R}(G(\boldsymbol{w}, g); \pi_i)$ and $\mathcal{R}$ is a differentiable renderer. The object $\mathcal{L}_{\text{pips}}$ is the perceptual loss, and π_i is the camera parameter of $\hat{\mathbf{I}}_i$.

To enforce geometric alignment, we add depth supervision. Let D_i be the depth map of $\hat{M}_{\theta_{\text{neu}},\psi_{\text{neu}}}$ rendered at π_i, and m_i a mask excluding hair regions. The depth loss is

$$\mathcal{L}_{\text{depth}}^{i} = \left\| \left(\mathcal{R}_d(G(\boldsymbol{w}, g); \pi_i) - D_i \right) \cdot m_i \right\|_2, \tag{5}$$

where $\mathcal{R}_d$ is the differentiable depth renderer. Depth supervision is omitted on hair regions since FLAME does not model hair. The multi-view 3D GAN inversion loss is

$$\mathcal{L}_{\text{mv}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{L}_{\text{img}}^{i} + \lambda_{\text{d}} \mathcal{L}_{\text{depth}}^{i} \right). \tag{6}$$

The inversion jointly enforces image and depth consistency across views, ensuring the reconstructed 3D Gaussian head aligns photometrically and geometrically with the input.

4.2. 3D Gaussian Rigging

Following GaussianAvatars [40], we bind the reconstructed Gaussian head $\mathcal{G}$ to the underlying FLAME mesh for animation. However, the initial geometry of the low-resolution

head model H, driven by the FLAME mesh M, often deviates from the super-resolved images $\{\hat{\mathbf{I}}_i\}_{i=1}^n$, so directly binding $\mathcal{G}$ to the mesh may cause mismatching between 3D Gausians and corresponding facial parts. For example, 3D Gaussians of teeth could be bound to lips on the mesh. To correct these misalignments, we first perform mesh geometry refinement, then bind 3D Gaussians to the refined mesh.

Geometry refinement. Let $M_{\beta,\theta,\psi}$ denote the FLAME mesh parameterized by global shape β, pose θ, and expression ψ. For each image $\hat{\mathbf{I}}_{i\leq n}$, we detect 2D facial landmarks $\ell_i \in \mathbb{R}^{2\times m}$, where m is the number of landmarks. The corresponding 3D landmarks are extracted as $\mathbf{X}_i = S(M_{\beta,\theta_i,\psi_i}) \in \mathbb{R}^{3\times m}$, where $S(\cdot)$ is a landmark selection operator that selects the m predefined mesh vertices corresponding to facial landmarks. With camera projection $\Pi_i(\cdot)$, the optimized shape parameter is

$$\beta^* = \arg\min_{\beta} \sum_{i=1}^{n} \|\Pi_i(\mathbf{X}_i) - \ell_i\|_2 \,. \tag{7}$$

Here, β is shared globally across all anchor images, while $(\theta_i, \psi_i, \Pi_i)$ vary per image. The optimized shape β^* is fixed for the rest of the pipeline. See Figure 9 for illustrations.

3D Gaussian rigging. Following GaussianAvatars [40], we bind the reconstructed Gaussian head $\mathcal{G}$ to the underlying FLAME mesh. Each Gaussian primitive is associated with its nearest mesh face. To enable animation, the Gaussian parameters (i.e., position $\boldsymbol{\mu}$, rotation $\mathbf{r}$, and scale $\mathbf{S}$) are transformed into the local coordinate system of that face as

$$\mathbf{r}' = \mathbf{R}^{-1}\mathbf{r}, \boldsymbol{\mu}' = \frac{1}{s}\mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{T}), \mathbf{S}' = \frac{1}{s}\mathbf{S}^{-1}, \tag{8}$$

where $\mathbf{R}$ is the face orientation, $\mathbf{T}$ is the mean of its three vertices, and s is a scaling factor. After the FLAME mesh deforms (e.g., under expression changes), the local Gaussian parameters are mapped back to the global coordinate system using Equation 3.

4.3. Dynamics-Aware 3D GAN Refinement

Through multi-view 3D GAN inversion (Section 4.1) and 3D Gaussian rigging (Section 4.2), we can animate a high-resolution 3D head model. However, this setup has the following limitations: First, there is no guarantee of faithful reconstruction at camera poses far from $\pi_{i\leq n}$; Second, Some facial regions are occluded in the canonical view (e.g., eyelids, inner mouth), causing incomplete reconstruction; Finally, $\mathcal{G}$ is unstable under deformation, causing artifacts in large facial motions. To solve the above limitations, we utilize multi-expression renderings from LR input model to jointly optimize the 3D GAN inversion process.

Multi-expression anchor images sampling. To begin, we first augment $\hat{\mathbf{I}}$ by adding super-resolved multi-expression anchor images, getting a augmented image set of k images in total. We manually select various facial expressions that (1) each facial region (e.g., teeth, eyelids) is visible in at least one selected expression, and (2) the selected expressions capture a wide range of common facial deformations to improve robustness under stretch and compression. With the selected expressions, we sample and super-resolve multi-view renderings from LR input model following the same way in Section 4.1. Note that in the augmented image set, $\{\hat{\mathbf{I}}_i\}_{i=1}^n$ is solely used in multi-view 3D GAN inversion, while $\{\hat{\mathbf{I}}_j\}_{j=1}^k, k > n$ is used for dynamics-aware 3D GAN refinement. See Figure 4 for representative examples.

Figure 4. Example of sampled anchor images and the corresponding FLAME mesh used for 3D GAN inversion.

Joint optimization via multi-expression anchor images. We leverage images with different poses and expressions to jointly supervise 3D GAN inversion. Let $\mathcal{T}_j$ denote the deformation from the canonical mesh $\hat{M}_{\theta_{\text{neu}},\psi_{\text{neu}}}$ to $\hat{M}_{\theta_j,\psi_j}$. This yields transformed Gaussian heads $\mathcal{T}_j(\mathcal{G})$, enabling both image and depth supervision at every view. The per-view image reconstruction loss is

$$\mathcal{L}_{\text{img}}^j = \left\|\hat{\mathbf{I}}'_j - \hat{\mathbf{I}}_j\right\|_2 + \lambda_{\text{p}}\,\mathcal{L}_{\text{pips}}(\hat{\mathbf{I}}'_j, \hat{\mathbf{I}}_j), \tag{9}$$

where $\hat{\mathbf{I}}'_j = \mathcal{R}(\mathcal{T}_j(G(\boldsymbol{w}, g)); \pi_j)$ is the rendered image at view π_i. The per-view depth loss is

$$\mathcal{L}_{\text{depth}}^j = \left\|(D'_j - D_j)\cdot m_j\right\|_2, \tag{10}$$

where $D'_j = \mathcal{R}_d(\mathcal{T}_j(G(\boldsymbol{w}, g)); \pi_j)$ is the rendered depth map and m_i excludes hair regions. The total objective averages both terms across all views:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{k}\sum_{j=1}^{k}\left(\mathcal{L}_{\text{img}}^j + \lambda_{\text{d}}\,\mathcal{L}_{\text{depth}}^j\right). \tag{11}$$

Finally, we can have the final loss term in each optimization iteration by combining Equation 6 and Equation 11:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mv}} + \mathcal{L}_{\text{dyn}} \tag{12}$$

Following Section 3.1 in 3D GAN inversion, we adopt a two-stage training strategy: first optimize $\boldsymbol{w}$ with Equation 12, then fine-tune g with fixed $\boldsymbol{w}^*$.

Table 1. State-of-the-art comparison of upsampling animatable 3D head avatars on the NeRSemble and INSTA datasets. Image metrics (PSNR↑, SSIM↑, LPIPS↓) are computed frame-by-frame using an identical validation motion sequence on novel camera poses and expressions. We also report inference time for each methods, showing that our method achieves best quality with competitive efficiency.

Method	NeRSemble dataset [30]						INSTA dataset [69]			Time
	Self-Reenactment			Novel View Synthesis			Self-Reenactment			
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	min
GaussianAvatars (LR) [40]	18.56	0.811	0.302	18.72	0.822	0.249	19.79	0.837	0.220	—
Video-based SR [14]	21.91	0.840	0.254	21.18	0.833	0.219	23.01	0.850	0.158	30
SuperGaussian [43]	19.57	0.837	0.278	21.57	0.860	0.215	22.89	0.842	0.177	14
SR + GPAvatar [11]	18.29	0.768	0.244	18.14	0.810	0.228	20.15	0.823	0.156	3
SuperHead (ours)	22.21	0.850	0.238	22.76	0.871	0.213	23.76	0.864	0.135	5

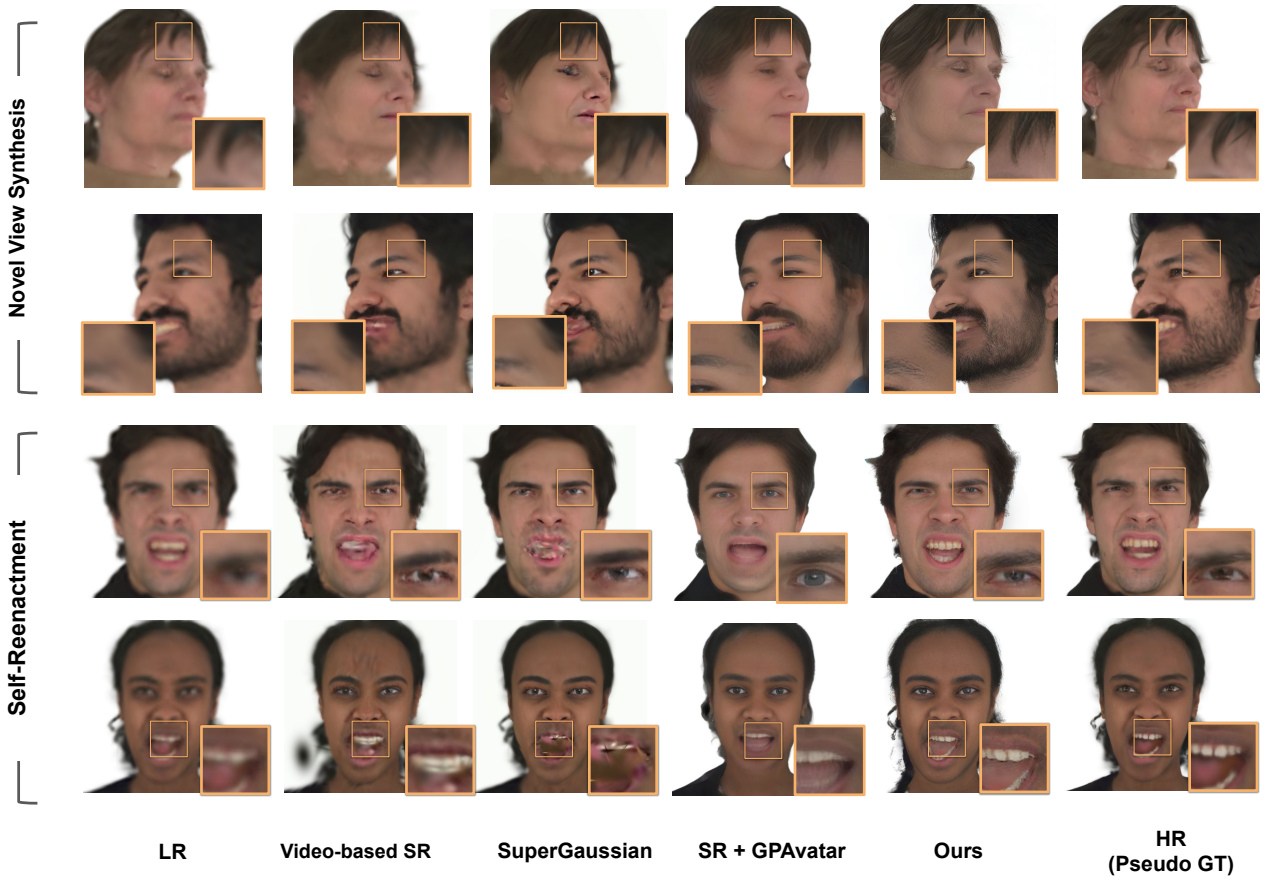

Figure 5. Qualitative comparisons on the NeRSemble dataset [30]. SuperHead synthesizes high-quality facial details across diverse expressions, clearly outperforming baselines and in some cases approaching the pseudo ground-truth head avatar.

5. Experiments

We validate SuperHead with diverse scenarios using different Gaussian avatar models. Section 5.1 describes the experimental setup, Section 5.2 presents the main results, and Section 5.3 provides ablations and analysis.

5.1. Setup

Datasets. We evaluate our method on two challenge datasets: NeRSemble [30] (multi-view video) and INSTA [69] (monocular video). For NeRSemble, a pseudo ground-truth animatable head avatar (HR avatar) was reconstructed using GaussianAvatars [40] from all multi-view recordings of a single animation sequence at a base resolution of 802×550. For INSTA, we train the HR avatar with monocular video. We use a down-sampling factor of 4 to create low-resolution videos. Each set of down-sampled videos was subsequently used to reconstruct a low-resolution head avatar (LR avatar) with GaussianAvatars, which serve as the input to our enhancement pipeline.

Figure 6. Qualitative results on INSTA dataset [69]. All methods are driven and rendered with novel camera poses and expressions.

Evaluation protocols. For NeRSemble, we consider two settings, both with baselines driven by the a validation sequence: (1) *Novel View Synthesis*, with novel camera poses $\pi_{\text{eval}} \notin \pi_{1\leq j\leq k}$. (2) *Self-Reenactment*, with novel expressions $\psi_{\text{eval}} \notin \psi_{1\leq j\leq k}$. We also show HR avatar as an upper-bound of 3D head super-resolution in qualitative results. For INSTA, which contains only monocular videos, we evaluate in the *Self-Reenactment* setting and report renderings with novel poses and expressions for all methods.

Metrics. We evaluate the quality of up-scaled animatable 3D head avatars using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [48], and Learned Perceptual Image Patch Similarity (LPIPS) [60].

Baselines. We compare SuperHead with the state-of-the-art methods: (1) *GaussianAvatars (LR)* [40]: We train it directly on low-resolution videos. (2) *Video-based SR*: Given a LR avatar and an animation sequence, we render multi-view video sequences and apply a video SR model [14] to upsample them independently. GaussianAvatars reconstructs the HR avatar from these upsampled videos. (3) *3D-based SR*: A naive baseline for upsampling an animatable 3D model is to upsample the static model in the canonical pose. Following this idea, we apply SuperGaussian [43] to upsample the static 3D Gaussian head under ψ_{neu} expression. The resulting HR head is then rigged to the original FLAME model and animated over the motion sequence. (4) *SR+GPAvatar*: We reconstruct a 3D head model using the few-shot method GPAvatar [11], with super-resolved multi-view renderings from the LR avatar as input.

Table 2. **Ablation study** on the key design components. Please see the corresponding qualitative analysis in Figure 8.

Setting	PSNR ↑	SSIM ↑	LPIPS ↓
SuperHead (ours)	22.15	0.841	0.243
(a) w/o multi-view inversion	19.86	0.812	0.283
(b) w/o multi-expr. inversion	21.18	0.824	0.269
(c) w/o depth supervision	21.74	0.829	0.251
(d) w/o FLAME refinement	22.03	0.838	0.246

Implementation details. We use pre-trained GSGAN [25] for 3D GAN inversion. We place cameras on a sphere of radius $r = 0.6$, with $\pi_{1\leq j\leq k}$ sampled within $\pm 45°$ across all axes and facing the LR input center. Each optimization runs 150 iterations with both multi-view inversion and dynamics-aware refinement. We use Adam with learning rate 0.04 for $\mathcal{W}^+$ inversion, followed by PTI [41] for the last 50 iterations with learning rate 0.002 and locality regularization to preserve GAN expressivity.

5.2. Main Results

We show quantitative results in Table 1. Our method outperforms all other baselines on both multi-view and monocular video datasets. We also report inference time for each enhancement method, and SuperHead shows great inference speed while achieving the best performance. Figure 5 and Figure 6 present qualitative comparisons between SuperHead and the baseline approaches. Other methods exhibit blurry textures and noticeable artifacts, likely due to inconsistencies arising from independently up-scaling multi-view

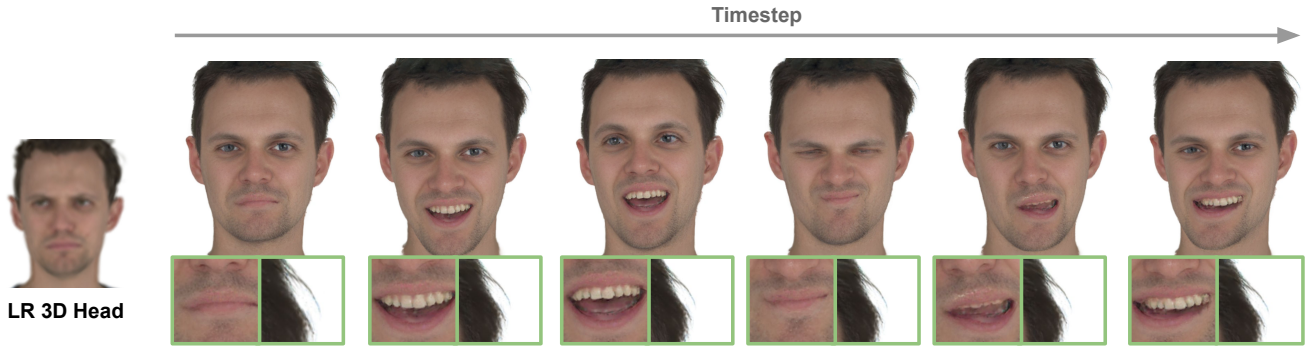

Figure 7. **Sequential frames of SuperHead**. SuperHead recovers fine details and facial components with multi-view/frame consistency.

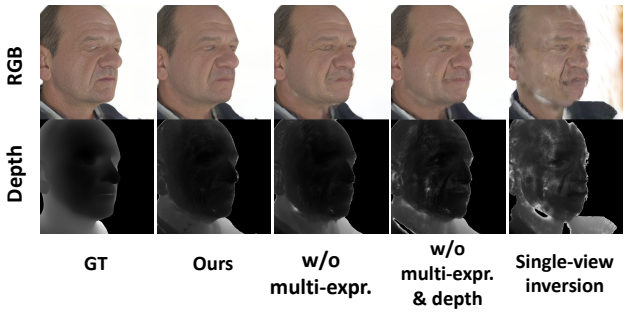

Figure 8. **Ablation studies.** We iteratively remove various design components, including multi-expression inversion, depth supervision, and multi-view inversion, and show the rendered image and depth maps of the output head avatar with novel expression and novel camera view. The results get progressively worse.

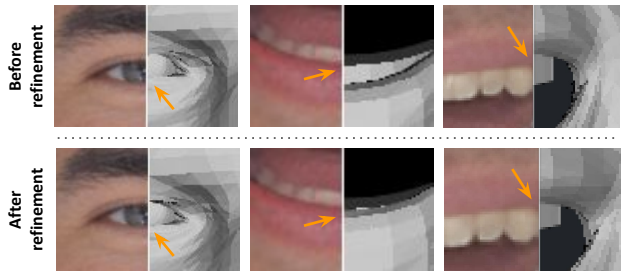

Figure 9. **Importance of geometry refinement.** Without refinement, anchor images misalign with the underlying geometry, such as the eye sockets, lower lip, and teeth. After refinement, the FLAME mesh aligns accurately with the anchor images.

inputs. In contrast, SuperHead synthesizes sharp textures and much richer facial details, e.g., facial hair and teeth. In some cases, the visual quality achieved by SuperHead is on par with and even surpass the pseudo ground-truth (GT) HR avatar in the synthesis of certain high-frequency details, e.g., eyebrows, underscoring the effectiveness of our approach in leveraging powerful generative 3D priors for the super-resolution task. We also show rendering sequence of SuperHead in 7, indicating that our method can recover high-quality and consistent facial details.

5.3. Ablation and Analysis

We conduct ablation studies to evaluate key design components and report results in Table 2 and Figure 8.

Multi-view inversion (cf. Section 4.1). When performing 3D GAN inversion with only single-view, the quality of novel views is poor with distorted geometry and appearance (Table 2-(a)). In contrast, with all components enabled, SuperHead produces high-fidelity reconstructions with smooth and coherent surfaces, demonstrating the complementary benefits of each design choice.

Multi-expression inversion and depth supervision (cf. Section 4.3). With only one expression, the 3D head often shows floating artifacts on novel expressions (Table 2-(b)). Similarly, omitting depth supervision can result in surfaces with holes (Table 2-(c)).

Geometry refinement. Before 3D Gaussian binding, we refine the mesh (cf. Section 4.2) to ensure Gaussians are rigged to the correct partsas, shown in Figure 9. This step also improves performance, as shown in Table 2-(d).

More results. We include videos, analysis on anchor image sampling, and results of comparability to other 3D avatar model (e.g., SplattingAvatar [42]) in the supplementary.

6. Discussion and Conclusion

We present SuperHead, a new framework for upsampling low-resolution animatable 3D head avatars. By leveraging upsampled multi-view images and depth cues across diverse expressions, SuperHead exploits 3D GAN priors to synthesize realistic 3D Gaussian heads with geometric consistency and temporal coherence. Extensive experiments show that SuperHead outperforms existing baselines in visual quality. We hope our method opens new possibilities for digital humans in VR, telepresence, and entertainment.

Limitation and future works. Our method still has challenges: it cannot handle dynamics of hair in head-shaking motion or recover 360 degree 3D head. Please be referred to the supplementary material for further discussion.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20950–20959, 2023. 3

[2] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. 2010. 1, 2

[4] Ananta R Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3055–3065, 2024. 2, 3

[5] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *ACM transactions on graphics (TOG)*, 26(3):33–es, 2007. 1, 2

[6] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 2

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[8] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 papers*, pages 1–10. 2010. 1, 2

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

[10] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[11] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 1, 2, 6, 7

[12] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10673–10683, 2022. 3

[13] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*, pages 459–476. Springer, 2024. 3

[14] Ruicheng Feng, Chongyi Li, and Chen Change Loy. Kalman-inspired feature propagation for video face super-resolution. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 6, 7

[15] Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, and Tony Tung. Vive3d: Viewpoint-independent video editing using 3d-aware gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4446–4455, 2023. 3

[16] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 1, 2

[17] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011. 1, 2

[18] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021. 1, 2

[19] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[20] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[21] Yuqi Han, Tao Yu, Xiaohang Yu, Di Xu, Binge Zheng, Zonghong Dai, Changpeng Yang, Yuwang Wang, and Qionghai Dai. Super-nerf: view-consistent detail generation for nerf super-resolution. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 3

[22] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. *arXiv preprint arXiv:2502.17796*, 2025. 1

[23] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[24] Xudong Huang, Wei Li, Jie Hu, Hanting Chen, and Yunhe Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8244–8253, 2023. 2, 3

[25] Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024. 3, 7

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[29] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2

[30] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2, 6

[31] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. *arXiv preprint arXiv:2406.09377*, 2024. 1, 3

[32] Hyun-kyu Ko, Dongheok Park, Youngin Park, Byeonghyeon Lee, Juhee Han, and Eunbyung Park. Sequence matters: Harnessing video models in 3d super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4356–4364, 2025. 3

[33] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023. 3

[34] Jaeseong Lee, Taewoong Kang, Marcel Buehler, Min-Jung Kim, Sungwon Hwang, Junha Hyung, Hyojin Jang, and Jaegul Choo. Surfhead: Affine rig blending for geometrically accurate 2d gaussian surfel head avatars. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[35] Jie Long Lee, Chen Li, and Gim Hee Lee. Disr-nerf: Diffusion-guided view-consistent super-resolution nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20561–20570, 2024. 3

[36] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 3

[37] Zhanfeng Liao, Yuelang Xu, Zhe Li, Qijing Li, Boyao Zhou, Ruifeng Bai, Di Xu, Hongwen Zhang, and Yebin Liu. Hhavatar: Gaussian head avatar with dynamic hairs. *arXiv e-prints*, pages arXiv–2312, 2023. 3

[38] Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 3d gan inversion for controllable portrait image animation. *arXiv preprint arXiv:2203.13441*, 2022. 3

[39] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024. 2, 3

[40] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1, 2, 3, 4, 5, 6, 7

[41] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 3, 7

[42] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 8

[43] Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J Mitra, Shenlong Wang, and Anna Frühstück. Supergaussian: Repurposing video models for 3d super resolution. In *European Conference on Computer Vision*, pages 215–233. Springer, 2024. 2, 3, 6, 7

[44] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7672–7682, 2022. 2

[45] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20991–21002, 2023. 3

[46] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. 3

[47] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 2, 3

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[49] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 3

[50] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 3

[51] Yunze Xiao, Hao Zhu, Haotian Yang, Zhengyu Diao, Xiangju Lu, and Xun Cao. Detailed facial geometry recovery from multi-view images by learning an implicit function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1, 2

[52] Yiran Xu, Zhixin Shu, Cameron Smith, Jia-Bin Huang, and Seoung Wug Oh. In-n-out: Face video inversion

and editing with volumetric decomposition. *arXiv preprint arXiv:2302.04871*, 3, 2023. 3

[53] Yiran Xu, Taesung Park, Richard Zhang, Yang Zhou, Eli Shechtman, Feng Liu, Jia-Bin Huang, and Difan Liu. Videogigagan: Towards detail-rich video super-resolution. *arXiv preprint arXiv:2404.12388*, 2024. 3

[54] Yuelang Xu, Zhaoqi Su, Qingyao Wu, and Yebin Liu. Gphm: Gaussian parametric head model for monocular head avatar reconstruction. 2024. 3

[55] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 2

[56] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[57] Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, et al. 3d gan inversion with facial symmetry prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2023. 2, 3

[58] Youngho Yoon and Kuk-Jin Yoon. Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12428–12438, 2023. 2, 3

[59] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2437–2447, 2023. 2, 3

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[61] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 3

[62] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20343–20352, 2022. 2

[63] Mingwu Zheng, Haiyu Zhang, Hongyu Yang, and Di Huang. Neuface: Realistic 3d neural face rendering from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16868–16877, 2023. 1, 2

[64] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, Guidong Wang, and Xu Lan. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 2, 3

[65] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[66] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 4

[67] Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *ECCV*, 2024. 3

[68] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*, 2022. 2

[69] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 2, 6, 7

From Blurry to Believable: Enhancing Low-quality Talking Heads with 3D Generative Priors

Supplementary Material

In this supplementary material, we provide additional details and results omitted in the main text.

A. Contribution and Limitations

Main contribution. While many previous works have explored super-resolution (SR) in 2D content, e.g., images, or 3D representation static representation, e.g., 3D Gaussian, super-resolution in dynamic 3D representation remains an unexplored topic. The main challenge lies in the fact that 2D SR not only struggles with multi-view but also temporal inconsistencies, when up-sampling a dynamic 3D representation. Our method addresses this challenge by performing multi-view and multi-expression 3D GAN inversion, ensuring that the synthesized 3D head preserves high-frequency details even when the up-sampled anchor images are inconsistent. To the best of our knowledge, this is the first attempt on super-resolution of dynamic 3D avatar representation.

Limitation. The major limitation is that 3D GAN cannot synthesize complete 3D head, i.e., it struggles to generate back of a human head. The main reason is that 3D GAN is trained on FFHQ [2], which consists only of frontal human faces. Building a large-scale human face dataset that includes views of back head is a possible way to extend 3D GAN's ability of synthesizing back views of human heads. As shown in Figure S1, while GSGAN [1] can synthesize frontal views of high-fidelity details, it struggles to synthesize the back of the human head.

Figure S1. 3D GAN struggles to synthesize the back of a human head. We rotate a synthesized head before camera to show quality gap between views of frontal and back of a head.

B. Additional Implementation Details

We adopt GSGAN [1] as our 3D GAN backbone. To make 3D GAN robust towards side views of a 3D head and hairstyles, we processed FFHQ [2] by cropping the image source to include full head in the image. Then, we fine-tuned the GSGAN checkpoint on the re-cropped FFHQ dataset. Figure S3 shows that the fine-tuned 3D GAN can not only synthesize finer details on facial parts but also accurate hairstyles. All of our experiments, including the 3D GAN fine-tuning, were conducted on a RTX A6000 GPU.

C. Additional Results and Analyses

Additional qualitative results. We show additional visual comparison of various baselines introduced in the main paper in Figure S3. Our method demonstrates superior capability in recovering detailed facial expressions, e.g., corner of the mouth, but also accurate geometry of the hair.

Table S1. SuperHead achieves identical performance when applying to SplattingAvatar [5] on INSTA dataset [6], proving SuperHead's generalizability to enhance diverse 3D avatar models.

Setting	PSNR ↑	SSIM ↑	LPIPS ↓
SplattingAvatar (LR) [5]	19.24	0.825	0.251
SuperHead + SplattingAvatar [5]	23.04	0.834	0.167
SuperHead + GaussianAvatars [4]	23.76	0.864	0.135

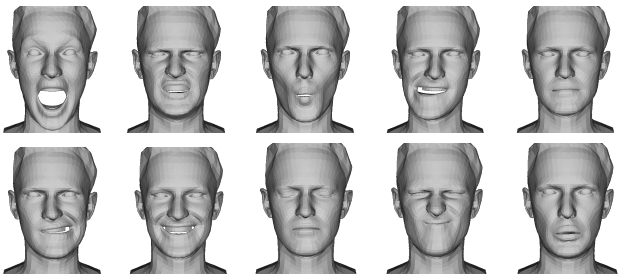

Figure S2. Expressions we used to sample anchor images.

Comparability to other 3D avatar model. We further evaluate our method on an alternative 3D avatar model to demonstrate its generalizability. Specifically, we adopt SplattingAvatar [5], which, similar to GaussianAvatars [4], rigs 3D Gaussians onto the FLAME mesh with a learnable normal offset to the surface. The upsampling procedure follows the same pipeline as with GaussianAvatars: we first sample and enhance anchor images from a SplattingAvatar trained on low-resolution captures, and then perform multi-view inversion along with dynamics-aware 3D refinement to optimize a 3D Gaussian head rigged on the underlying FLAME mesh. The results are reported in Table S1. We compare SplattingAvatar (LR), a 3D head model trained on low-resolution captures, with SuperHead + SplattingAvatar, the corresponding upsampled 3D head model. We show that our method successfully enhances low-quality 3D head

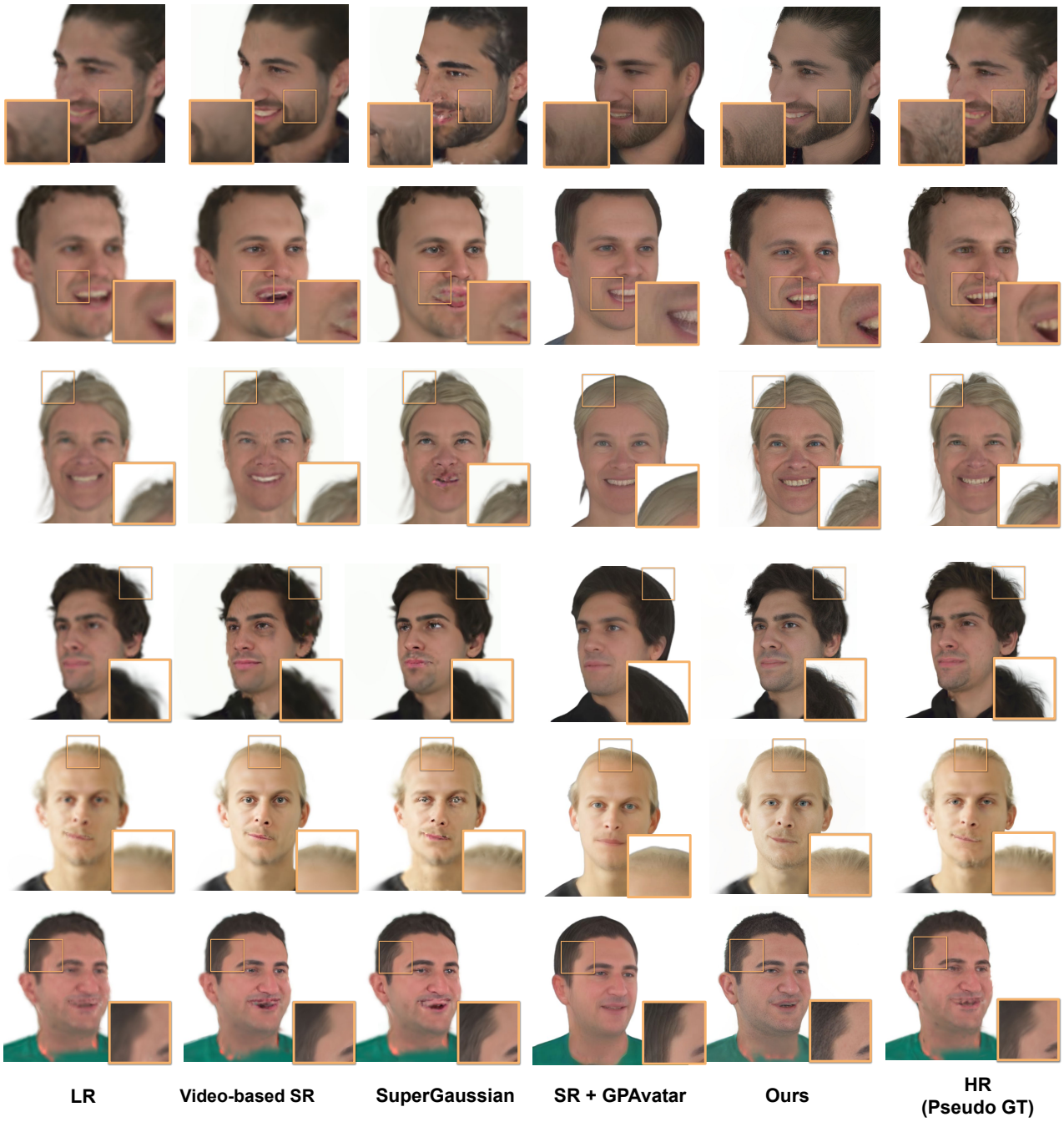

Figure S3. Additional qualitative results on the NeRSemble dataset [3] and INSTA [6]. In addition to zooming in facial parts of results, we also show the holistic view of upsampled 3D avatar, indicating that our method can not only enhance facial expressions but also details such as hair strands. Please zoom in to check details.

models across different design choices, thereby demonstrating its strong generalizability.

Anchor image sampling As mentioned in Section 4.3 of the main paper, we perform dynamics-aware 3D GAN refinement to improve the synthesized 3D head under different expressions and motions. For this purpose, we carefully select a set of expressions to form an "expression pool", from which we sample anchor images with different camera poses. We found that a set of 10 expressions is sufficient to achieve good performance. We show the expressions we use throughout our experiments in Figure S2. The selected expressions cover a range of facial motions, from screaming to smiling and eye-closing.

D. Ethical and Societal Impacts

Our work improves the quality of 3D head avatar reconstruction, which has potential benefits in areas such as telecommunication and digital content creation. At the same time, we are aware of possible risks, including issues of privacy, misuse for non-consensual content, and bias in representation. We emphasize the importance of developing and applying such techniques responsibly and with appropriate safeguards. Furthermore, like many generative methods, reconstruction results may contain certain biases if not carefully addressed. We believe it is important for future research and deployment of these techniques to be guided by principles of responsible AI, including fairness, transparency, and safeguards against malicious use.

References

[1] Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024. 1

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[3] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2

[4] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1

[5] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[6] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 1, 2