GAIA: Generative Animatable Interactive Avatars with Expression-conditioned Gaussians

ZHENGMING YU, NVIDIA, USA and Texas A&M University, USA
TIANYE LI, NVIDIA, USA
JINGXIANG SUN, NVIDIA, China and Tsinghua University, China
OMER SHAPIRA, SEONWOOK PARK, MICHAEL STENGEL, and MATTHEW CHAN, NVIDIA, USA
XIN LI and WENPING WANG, Texas A&M University, USA
KOKI NAGANO and SHALINI DE MELLO, NVIDIA, USA

Fig. 1. Our method GAIA generates animation-ready Gaussian avatars. GAIA supports photorealistic novel view synthesis and individual control of identity and expression. With efficient generation and rendering, GAIA is readily available for interactive animation and editing.

3D generative models of faces trained on in-the-wild image collections have improved greatly in recent times, offering better visual fidelity and view consistency. Making such generative models animatable is a hard yet rewarding task, with applications in virtual AI agents, character animation, and telepresence. However, it is not trivial to learn a well-behaved animation model with the generative setting, as the learned latent space aims to best capture the data distribution, often omitting details such as dynamic appearance and entangling animation with other factors that affect controllability. We present **GAIA**: **G**enerative **A**nimatable **I**nteractive **A**vatars, which is able to generate high-fidelity 3D head avatars for both realistic animation and rendering. To achieve consistency during animation, we learn to generate Gaussians embedded in an underlying morphable model for human heads via a shared UV parameterization. For modeling realistic animation, we further design the generator to learn expression-conditioned details for both geometric deformation and dynamic appearance. Finally, facing an inevitable entanglement problem between facial identity and expression, we propose a novel two-branch architecture that encourages the generator to disentangle identity and expression. On existing benchmarks, GAIA achieves state-of-the-art performance in visual quality as well as realistic animation. The generated Gaussian-based avatar supports highly efficient animation and rendering, making it readily available for interactive animation and appearance editing.

CCS Concepts: • **Computing methodologies** → **3D imaging**; **Animation**; **Volumetric models**; • **Theory of computation** → **Adversarial learning**.

Additional Key Words and Phrases: 3D Gaussian Splatting, Neural Avatars, Avatar Animation

ACM Reference Format:
Zhengming Yu, Tianye Li, Jingxiang Sun, Omer Shapira, Seonwook Park, Michael Stengel, Matthew Chan, Xin Li, Wenping Wang, Koki Nagano, and Shalini De Mello. 2025. GAIA: Generative Animatable Interactive Avatars with Expression-conditioned Gaussians . In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721238.3730737

The work of Zhengming and Jingxiang was done during their internships at NVIDIA. Authors' addresses: Zhengming Yu, NVIDIA, USA and Texas A&M University, USA; Tianye Li, NVIDIA, USA; Jingxiang Sun, NVIDIA, China and Tsinghua University, China; Omer Shapira; Seonwook Park; Michael Stengel; Matthew Chan, NVIDIA, USA; Xin Li; Wenping Wang, Texas A&M University, USA; Koki Nagano; Shalini De Mello, NVIDIA, USA.

SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730737

1 INTRODUCTION

Generating photo-real animatable 3D faces is a crucial component in the modeling of humans with widespread applications in content creation, 3D video conferencing, and telepresence. It is non-trivial to create and animate 3D faces with realistic appearance using traditional graphics pipelines [Alexander et al. 2009; Deng and Noh 2008], as it often requires specialized artistic skill, manual adjustment of geometry and textures and highly curated multi-view reference data [Kirschstein et al. 2023; Saito et al. 2024]. Recent 3D-aware generative models (e.g., 3D GANs) that learn 3D faces from large diverse in-the-wild collections of 2D images are promising towards

improving photorealism [Chan et al. 2022] and, recently, 3D consistency [Kirschstein et al. 2024; Trevithick et al. 2024]. They also reduce the need for costly specialized data collection and provide improved generalization to in-the-wild operating conditions.

While 3D-aware GANs are paving the way towards fast unconditional or sparse-view conditioned [Trevithick et al. 2023] generation of highly photo-real 3D digital human faces, most solutions only allow for explicit control over the subject's identity (appearance and shape) through user-provided inputs. However, controlling facial expressions independently from identity remains largely unsolved. In this work, we address this fundamental challenge of conferring 3D-aware GANs for facial synthesis with the ability to explicitly and independently control *both* the subject's identity and expressions via intuitive user-provided controls, in other words, of creating *animatable* 3D-aware generative models for faces.

A few prior works [Bergman et al. 2022; Hong et al. 2022; Sun et al. 2023; Wu et al. 2022] have attempted to solve this problem. However, all employ either neural radiance fields (NeRF) [Mildenhall et al. 2021] or their efficient triplane representation [Chan et al. 2022] to model and animate faces. These scene representations while powerful at achieving photorealism, learn deformations implicitly and therefore do not allow precise and intuitive control of the face. Furthermore, their NeRF-based renderer is slow and hence requires upsampling of the low-resolution rendered images, compromising 3D consistency especially for details such as hair and wrinkles.

Several previous works [Chen et al. 2024; Deng et al. 2024b,a; Hong et al. 2022; Qian et al. 2024; Xu et al. 2024] achieve high quality avatar reconstruction and reenactment utilizing a reconstruction loss from multi-view or monocular videos. However, learning animatable 3D-aware GANs for faces from unstructured 2D image collections is a more challenging problem due to weaker supervision signals provided by the adversarial loss. 3D generative models typically learn a latent space that is representative of the training dataset. Walking on such a latent space results in images that encode not only different identities, but also illumination conditions, accessories, expressions and other deformations of the face, resulting in significant entanglement between these various factors. However, in tasks such as character animation, facial expressions often need to vary over time while other extraneous factors such as identity and illumination conditions should remain fixed. Hence, to achieve such controlled generation with 3D GANs, one must disentangle facial expression-related factors from the latent space. Additionally, it is well established in traditional computer graphics that modeling high-quality detailed animation such as wrinkles requires the modeling of dynamic textures conditioned on the character's animation state [Gotardo et al. 2018]. Yet, this fact is overlooked in current 3D-aware generative models for faces.

To address these challenges, we present **GAIA**: **G**enerative **A**nimatable **I**nteractive **A**vatar for high-fidelity 3D head avatar generation with controllable realistic animation and fast rendering. We identify that using an expressive morphable parametric model (FLAME) [Li et al. 2017] with its inbuilt explicit expression-controlled deformation, provides a strong and intuitive prior for modeling facial expressions. Hence in GAIA, we generate Gaussians, via an efficient StyleGAN architecture, that are embedded on an underlying FLAME model for human heads via a shared UV parametrization. We explicitly offset the Gaussian primitives [Kerbl et al. 2023] with FLAME-parameterized expression deformations to model facial expressions. The adoption of Gaussian splatting further improves fine-detail generation (such as hair), and view consistency of faces along with rendering speed. However, coarse deformations applied via FLAME only model deformations of the facial region. To model fine-grained high-fidelity animation for the entire head, inspired by traditional computer graphics, we further design our generator to learn expression-conditioned details for both geometric deformation and dynamic appearance by explicitly conditioning it on expression, besides identity and viewpoint. While this approach faithfully models the appearance of the training distribution, it invariably entangles identity and expression controls during generation. We disentangle them in two steps: (a) similar to the pose-conditioned discriminator introduced in EG3D [Chan et al. 2022], we adopt a dual discriminator that is expression and shape conditioned, and (b) we adopt a two-branch architecture to separately learn to apply identity and expression related residuals, trained with a multi-stage training procedure and regularization strategies. On various benchmarks, GAIA achieves state-of-the-art performance in visual quality as well as realistic animation, while maintaining 3D viewpoint consistency. The generated Gaussian-based avatars support highly efficient animation and rendering at interactive speed.

The contributions of this paper are as follows:

- We propose GAIA, an animatable 3D-aware generative model for high-fidelity 3D head avatar generation with controllable realistic animation and fast rendering.
- We achieve this goal by an expression-conditioned generation architecture that disentangles expression and identity controls and generates animatable Gaussians through the UV parametrization and joint articulations from a FLAME morphable model.
- We achieve state-of-the-art performance in visual quality as well as realistic animation and build a real-time interactive application demonstrating controllable character animation.
- We release our source code at https://research.nvidia.com/labs/amri/projects/gaia/.

2 RELATED WORK

In this section, we discuss related research on 3D generative adversarial networks and their applications to animating human avatars.

2.1 3D Generative Adversarial Networks

3D Generative Adversarial Networks (GANs) allow the learning of implicit or explicit 3D representations from unstructured collections of 2D images. Early works in 3D GANs defined implicit voxel-based representations and CNN-based neural renderers [Henzler et al. 2019; Nguyen-Phuoc et al. 2019, 2020; Niemeyer and Geiger 2021; Xue et al. 2022]. Neural Radiance Fields (NeRF) [Mildenhall et al. 2021] and its differentiable volume rendering method improved 3D consistency and photo-realism in later works, but with increased computational costs [Bergman et al. 2022; Cai et al. 2022; Chan et al. 2021; Deng et al. 2022; Gu et al. 2022; Schwarz et al. 2020]. The triplane representation was proposed in [Chan et al. 2022] to reduce computational complexity. Still, its super-resolution post-processing step resulted in inconsistent details which were

improved by training at higher-resolutions via patch-based discrimination [Skorokhodov et al. 2022; Xiang et al. 2023], distillation [Chen et al. 2023], and learnable ray sampling [Trevithick et al. 2024].

The introduction of 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] allowed fast and high-resolution rendering via a collection of explicit Gaussian primitives and a differentiable rasterization method. The explicit nature of 3DGS allows for higher multi-view consistency and association with known 3D geometries as demonstrated in GSM [Abdal et al. 2024] and GGHead [Kirschstein et al. 2024]. While GSM allowed human pose articulation using the underlying SMPL model [Loper et al. 2023], GGHead used the FLAME model [Li et al. 2017] only as a static template mesh, without the possibility to independently control the identity and expression of the face. In contrast, GAIA's expression-conditioned GAN approach captures fine details such as wrinkles with 3D consistency and also allows precise independent control of expressions, gaze and pose.

2.2 Neural Animatable Human Avatars

While 3D GANs can learn to capture 3D understanding from 2D images, it is challenging to simultaneously learn a method that allows faithful animation. Such animation capabilities are important for human avatars, as evidenced by the introduction of diverse models for controlling the shape and texture of the human face or body, such as morphable models (3DMM) [Blanz and Vetter 1999; Dai et al. 2020; Li et al. 2017; Loper et al. 2023; Ploumpis et al. 2020]. Early neural face animation works used 3DMMs [Thies et al. 2016] or keypoints [Wang et al. 2021; Zakharov et al. 2019] for driving face deformation. More recent works adopted the triplane representation and used a 3DMM [Chu and Harada 2024; Li et al. 2024], semantic maps [Sun et al. 2022], or implicitly learned representations [Deng et al. 2024b,a; Hong et al. 2022; Tran et al. 2024] to transfer facial expressions from one image to another while exhibiting good 3D consistency. Works such as Gaussian Head Avatars [Xu et al. 2024] and MonoGaussianAvatar [Chen et al. 2024] combined 3DMMs with Gaussian Splatting for improved visual fidelity and view consistency. Unlike 3D GANs, these methods were trained with precise ground-truth and image reconstruction losses. Several 3D GAN works for body articulation and animation have been proposed [Abdal et al. 2024; Hong et al. 2023; Noguchi et al. 2022], but few works have addressed the task of training an animatable 3D GAN for faces [Bergman et al. 2022; Sun et al. 2023; Wu et al. 2022].

Most related to our work is Next3D [Sun et al. 2023], which incorporates a 3DMM with neural textures. Next3D follows the design of EG3D [Chan et al. 2022] and adopts a super-resolution post-processing step, resulting in a lack of details such as wrinkles and low 3D consistency in the renderings. The architecture built on top of implicit representations makes it tricky to achieve fine-level animation control or efficient inference. In contrast, our approach GAIA takes advantage of the coarse yet reliable expression model from FLAME [Li et al. 2017], and learns to generate expressive dynamic details on top of it with expression-conditioned Gaussians and achieves photorealistic rendering, expressive animation, eye control, high 3D consistency as well as interactive rendering speeds.

3 METHOD

We illustrate an overview of our method, GAIA, in Fig. 2. GAIA is able to generate high-fidelity animation-ready head avatars. The key to achieving high-quality animation control is to properly decompose the generation procedure into a structure that is compatible with the animation process, i.e., factorizing control variables such as identity and expressions. Towards this goal, we first introduce an expression-conditioned Gaussian generation architecture, which disentangles identity and expression variations by two separate branches (Sec. 3.2). To animate the generated 3D Gaussians, we adopt a generalized skinning formulation (Sec. 3.3). We then propose strategies to effectively regularize the generated 3D Gaussians (Sec. 3.4) and a multi-stage adversarial training scheme to effectively train the generators on in-the-wild image datasets (Sec. 3.5).

3.1 Background

3.1.1 Gaussian Splatting. [Kerbl et al. 2023] proposed 3D Gaussian Splatting (3DGS), which models a 3D scene by a collection of Gaussian primitives. Each Gaussian $\mathbf{g}_i$ contains five attributes: $\mathbf{g}_i = \{\mathbf{p}_i, \mathbf{s}_i, \mathbf{q}_i, o_i, \mathbf{c}_i\}$, including the position of the Gaussian $\mathbf{p}_i \in \mathbb{R}^3$, scale vector $\mathbf{s}_i \in \mathbb{R}^3$, quaternion vector $\mathbf{q}_i \in \mathbb{R}^4$, opacity $o_i \in [0, 1]$ and spherical harmonic coefficients $\mathbf{c}_i$ for view-dependent appearance. Given all the Gaussians of the scene $\mathcal{G} = \{\mathbf{g}_i\}$ and camera parameters $\boldsymbol{\pi}$, 3DGS utilizes an efficient tile-based rasterizer $\mathcal{R}$ to render the 3D representation to an image,

$$\mathbf{I} = \mathcal{R}(\mathcal{G}, \boldsymbol{\pi}). \tag{1}$$

3.1.2 Head Morphable Models. FLAME [Li et al. 2017] is a differentiable function $\mathcal{F}$ that produces $N = 5023$ deformed vertices $\mathbf{V} \in \mathbb{R}^{N\times 3}$ given control variables shape $\boldsymbol{\beta}$, expression $\boldsymbol{\psi}$ and pose $\boldsymbol{\theta}$, i.e., $\mathbf{V} = \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta})$. The shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{300}$ captures the diversity of identity shapes. The expression parameter $\boldsymbol{\psi} \in \mathbb{R}^{100}$ controls the facial expression. FLAME further models the surface deformation due to bone activation of a human head, with pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ being $K = 5$ joint rotations (in axis angles) for the root joint, neck, jaw and the two eyeballs. The pose changes manifest by applying a Linear Blend Skinning (LBS) function $\mathcal{S}$ on the pre-skinning template vertices $\mathbf{V}_T$,

$$\mathbf{V} = \mathcal{S}(\mathbf{V}_T, \mathbf{J}, \boldsymbol{\theta}, \mathcal{W}), \tag{2}$$

where $\mathbf{J} \in \mathbb{R}^{K\times 3}$ are the activated 3D joint locations and $\mathcal{W}$ is the blendweight matrix that models the influence of each joint on all vertices. The pre-skinning vertices $\mathbf{V}_T$ are computed by applying a combination of linear components on the template mesh vertices $\bar{\mathbf{V}}$,

$$\mathbf{V}_T = \bar{\mathbf{V}} + \mathbf{B}_s\boldsymbol{\beta} + \mathbf{B}_e\boldsymbol{\psi} + \mathbf{B}_p f(\boldsymbol{\theta}), \tag{3}$$

where $\mathbf{B}_*$ are linear base tensors for shape (s), expression (e) as well as pose correctives (p) to prevent LBS artifacts. The linear coefficients for pose correctives are a transformed version of pose parameters $\boldsymbol{\theta}$. Please refer to [Li et al. 2017] for further details.

3.2 Expression-conditioned Gaussian Generation

3.2.1 UV-based Gaussian Attributes. To achieve photorealistic generation, we choose 3D Gaussians [Kerbl et al. 2023] as our representation. However, as a point-based representation, Gaussians are

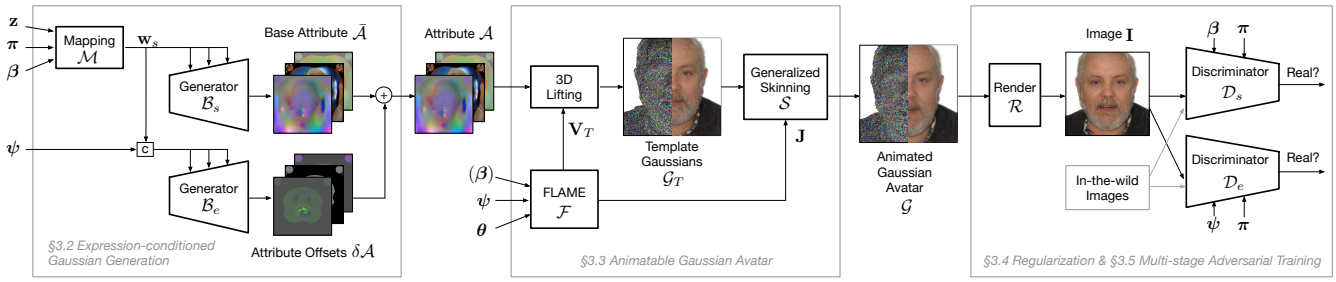


Fig. 2. **Overview.** GAIA generates animation-ready head avatar at high visual and animation fidelity by only learning from in-the-wild 2D images.

inherently unstructured, which poses challenges to properly regularize them under deformation. Similar to [Kirschstein et al. 2024], our model generates Gaussian attributes on UV maps $\mathcal{A}_\star \in \mathbb{R}^{T\times T\times D_\star}$ for all attribute categories[1] $\star \in \{p, s, q, o, c\}$, where T is the resolution of the UV maps and $D_\star$ is the dimension of Gaussian attribute $\star$. Generation on the UV maps takes advantage of existing powerful 2D generative backbones such as StyleGAN [Karras et al. 2020]. More specifically, we choose the UV parametrization that corresponds to the FLAME face model, which builds up a bridge between the 3D Gaussian primitives and the underlying morphable and articulate structure of the FLAME model. We will detail this in Sec. 3.3.

3.2.2 Conditioning Generator with Shape and Expression. Existing unconditional generative avatars [Kirschstein et al. 2024] model all variations of the face with a purely learned latent space $\mathbf{z}$ and camera conditions $\boldsymbol{\pi}$, which does not support factorized control for animating expression or editing identity. To support factorized animation control, we add FLAME shape (identity) $\boldsymbol{\beta}$ and expression $\boldsymbol{\psi}$ in the Gaussian attribute generator as additional conditioning variables. Another motivation of this design is to model expression-conditioned geometry and appearance changes, as these are crucial details to reach high realism (e.g., wrinkles) during animation.

We empirically find that adding shape and expression as additional conditioning variables helps the generator to better capture the data distribution. However, this comes with a cost that the generator tends to rely on the expression label to memorize the identity. This causes entangled generation where a change in expression parameters can affect the identity, which is a behavior clearly not acceptable for an animation system. We show these observations in Tab. 2 and Fig. 6 in Sec. 4.2.3. To prevent entanglement between identity and expression, we propose a novel expression-conditioned generation architecture, which consists of two separate branches to model identity and expression respectively, as shown in Fig. 2.

3.2.3 Two-Branch Architecture. We propose a two-branch generation architecture to decouple the generation processes of identity and expression. We design the identity branch to capture most of the geometry and appearance related to the person's identity. To produce the expression-dependent appearances and deformations (e.g., wrinkles), the expression branch learns to produce offsets based on the output from the identity branch, which mimics the dynamic displacement [Cao et al. 2015; Nagano et al. 2015] in traditional facial animation.

We generate identity-conditioned attributes with the shape branch. The shape branch firstly takes shape parameter $\boldsymbol{\beta}$ as well as latent code $\mathbf{z}$ and camera parameter $\boldsymbol{\pi}$ through the mapping network $\mathcal{M}$ to obtain an intermediate latent variable $\mathbf{w}_s$. Then we adopt a StyleGAN-style [Karras et al. 2020] backbone $\mathcal{B}_s$ to generate identity-related Gaussian attributes UV maps $\bar{\mathcal{A}}_\star$ for all Gaussian attribute categories $\star \in \{p, s, q, o, c\}$,

$$\mathbf{w}_s = \mathcal{M}(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\beta}), \quad \bar{\mathcal{A}}_\star = \mathcal{B}_s(\mathbf{w}_s). \tag{4}$$

We design another StyleGAN generator $\mathcal{B}_e$ to generate expression-conditioned attributes $\delta\mathcal{A}_\star$ based on the expression input $\boldsymbol{\psi}$ and the intermediate feature $\mathbf{w}_s$ from the shape branch,

$$\delta\mathcal{A}_\star = \mathcal{B}_e(\mathbf{w}_s, \boldsymbol{\psi}). \tag{5}$$

We obtain the final Gaussian attribute maps by applying the expression-conditioned attributes on the identity-conditioned attributes as offsets. In particular, we apply a binary UV mask of the face region $\mathbf{M}$ on $\delta\mathcal{A}_\star$ to further constrain the influence of the expression branch (e.g., facial expressions usually do not affect the hair regions). This masking, together with the regularization (Sec. 3.4) and multi-stage training scheme (Sec. 3.5), improves decoupling identity and expression.

$$\mathcal{A}_\star = \bar{\mathcal{A}}_\star + \mathbf{M} \odot \delta\mathcal{A}_\star. \tag{6}$$

3.3 Animatable Gaussian Avatar

Given the Gaussian attribute maps $\mathcal{A}_\star$ generated by the expression-conditioned generators, we lift the attributes to the 3D space and then animate them with a generalized skinning function.

3.3.1 Lifting Gaussian Attributes. Our Gaussian representation is embedded on the predefined UV parameterization of the FLAME head template. For animation, we need to lift the Gaussians into the canonical 3D space. Each valid texel $\mathbf{t}_i \in [0\,..\,T-1]^2$ corresponds to the i^{th} *template* Gaussian $\mathbf{g}_{T_i}$. All 3D Gaussian attributes $\star$ except for the positional and scale attributes are directly sampled from the corresponding UV-based attribute maps.

$$\star_{T_i} = \mathcal{I}(\mathcal{A}_\star, \mathbf{t}_i), \quad \forall \star \in \{q, o, c\} \tag{7}$$

where $\mathcal{I}(\cdot)$ is the bilinear sampling function on the UV map.

We regularize the range of the scale attributes by applying an additional non-linear transformation after the bilinear interpolation,

[1] Abbreviation of the Gaussian attribute categories: position (p), scale (s), rotation (q), opacity (o) and color (c).

similar to [Kirschstein et al. 2024],

$$\mathbf{s}_{T_i} = \exp(-s_{\max} - \text{softplus}(-(\mathcal{I}(\mathcal{A}_s, \mathbf{t}_i), -s_{\text{init}}) - s_{\max})). \quad (8)$$

The positional attribute map $\mathcal{A}_p$ is interpreted as offsets on the corresponding points on the morphable FLAME template (Eq. 3),

$$\mathbf{p}_{T_i} = \mathcal{I}'(\mathbf{V}_T, \mathbf{t}_i) + \gamma_p \cdot \tanh(\mathcal{I}(\mathcal{A}_p, \mathbf{t}_i)) \quad (9)$$

where $\mathcal{I}'(\cdot)$ is the bilinear sampling function on the mesh vertices. We apply a $\tanh(\cdot)$ transformation to the offsets to further limit their range. We empirically choose $s_{\max} = 3$, $s_{\text{init}} = 5$ and $\gamma_p = 0.25$. The generated Gaussian attributes $\mathcal{A}_\star$ (Eq. 7, 8 and 9) can be interpreted as additional corrective "blendshapes" as well as detailed appearance attributes, which are not modeled in a morphable model.

3.3.2 Generalized Skinning Function. Given the template Gaussians $\mathcal{G}_T$ computed by Eq. 7, 8 and 9, we then apply a generalized skinning function $\mathcal{S}$ to animate the Gaussians. The function takes a similar form of the original skinning function (Eq. 2) in FLAME. We bilinearly sample the original blendweight matrix $\mathcal{W}$ via the existing barycentric weights and produce a blendweight matrix $\mathcal{W}_g$, establishing the influence of the joints on all generated Gaussians,

$$\mathcal{G} = \mathcal{S}(\mathcal{G}_T, \mathbf{J}, \boldsymbol{\theta}, \mathcal{W}_g). \quad (10)$$

The generalized skinning function naturally inherits the existing FLAME facial rig, which improves realism and offers compatibility with existing graphics pipelines.

3.4 Regularization

We propose regularization strategies, which effectively encourage the generator to maintain reasonable animation behavior and facilitate the decoupling between identity and expression.

3.4.1 Expression-conditioned Generation. The entanglement issue between identity and expression often manifests as the expression parameters have an overly high influence on the generated results. To ensure that the expression-conditioned generator produces plausible attributes, we propose several regularization techniques. To constrain the influence of the expression branch, other than limiting the offsets in face region by a mask in Eq. 6, we disable the generated position and scale attribute offsets, as the expression-dependent details (e.g., wrinkle) are mostly a local effect. Furthermore, we apply L_2 regularization on the Gaussian attribute offsets $\delta\mathcal{A}_\star$,

$$\mathcal{L}_{\text{exp}} = \sum_{\star \in \{q,o,c\}} \|\mathbf{M} \odot \delta\mathcal{A}_\star\|_2 . \quad (11)$$

3.4.2 Inner Mouth. The inner mouth region is usually difficult to model as the oral cavity contains intricate internal structure such as the teeth and tongue. Similarly to FlashAvatar [Xiang et al. 2024], we stitch the inner mouth of the FLAME template mesh to prevent hole artifacts. We find that adding position and scale offsets in the mouth region helps to model the teeth, tongue and some light and some light interaction between them (e.g., shadows). We then only add position and scale offset in the mouth region.

3.4.3 Eyeballs. An important difference in design compared to existing methods is that our generated attributes are applied on a full-fledged and more detailed FLAME model with eyeball modeling. To maintain the original shape of eyeballs and reduce artifacts when the eyeballs are rotated, we apply a Total-Variation (TV) loss on the attributes maps of the eyeball region (by mask $\mathbf{M}_{\text{eyeballs}}$),

$$\mathcal{L}_{\text{uv}} = \sum_{\star} \text{TV}(\mathbf{M}_{\text{eyeballs}} \odot \mathcal{A}_\star). \quad (12)$$

3.4.4 All Gaussians. Similar to [Kirschstein et al. 2024], we also apply regularization terms on the final Gaussian attribute maps $\mathcal{A}_\star$ to constrain the evolution of Gaussians during training, resulting in better geometric and animation quality,

$$\mathcal{L}_p = \left\|\mathcal{A}_p\right\|_2, \mathcal{L}_s = \|\mathcal{A}_s\|_2, \mathcal{L}_o = \text{Beta}(\mathcal{A}_o), \quad (13)$$

where $\text{Beta}(\cdot)$ is the negative log-likelihood term of Beta(0.5, 0.5) distribution from [Lombardi et al. 2019].

3.5 Multi-Stage Adversarial Training

We design a multi-stage adversarial training scheme to effectively train the two-branch expression conditioned architecture.

3.5.1 Stage 1: Shape Only Training. In this stage, we only use the shape branch to generate the Gaussian attributes, i.e., $\mathcal{A}_\star = \bar{\mathcal{A}}_\star$. We pair the shape branch with a shape-conditioned discriminator $\mathcal{D}_s(\mathbf{I}; \boldsymbol{\beta}, \boldsymbol{\pi})$, where $\mathbf{I} = \mathcal{R}(\mathcal{G}(\bar{\mathcal{A}}_\star))$ is the rendered images of the shape-only generation. The generator $\mathcal{B}_s$ and the mapping function $\mathcal{M}$ is trained by the standard non-saturating GAN loss [Goodfellow et al. 2020] with R1 regularization [Mescheder et al. 2018].

$$\mathcal{L}^s_{\text{adv}} = \text{softplus}(-\mathcal{D}_s(\mathcal{R}(\mathcal{G}(\bar{\mathcal{A}}_\star)), \boldsymbol{\beta}, \boldsymbol{\pi})). \quad (14)$$

The total training loss is

$$\mathcal{L}^s_{\text{total}} = \mathcal{L}^s_{\text{adv}} + \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_o \mathcal{L}_o + \lambda_{\text{uv}} \mathcal{L}_{\text{uv}}. \quad (15)$$

To efficiently train the shape branch, we first train the generator at 256^2 rendering resolution as well as 256^2 UV resolution, with around 65K Gaussians. We then train the generator at 512^2 for both the rendering resolution and generated UV map resolution, which leads to around 262K Gaussians. When increasing the rendering resolution from 256 to 512, we add additional layers at both the generator and discriminator.

3.5.2 Stage 2: Joint Shape and Expression Training. We add the expression branch and train it along with the shape branch. We further design an expression-conditioned discriminator $\mathcal{D}_e(\mathbf{I}; \boldsymbol{\psi}, \boldsymbol{\pi})$, where $\mathbf{I} = \mathcal{R}(\mathcal{G}(\mathcal{A}_\star))$ is the rendered images of the full generation at 512^2 resolution,

$$\mathcal{L}^e_{\text{adv}} = \text{softplus}(-\mathcal{D}_e(\mathcal{R}(\mathcal{G}(\mathcal{A}_\star)), \boldsymbol{\psi}, \boldsymbol{\pi})). \quad (16)$$

The total training loss of this stage is

$$\mathcal{L}^e_{\text{total}} = \lambda_{d1} \mathcal{L}^s_{\text{adv}} + \lambda_{d2} \mathcal{L}^e_{\text{adv}} + \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_o \mathcal{L}_o + \lambda_{\text{uv}} \mathcal{L}_{\text{uv}} + \lambda_{\text{exp}} \mathcal{L}_{\text{exp}}. \quad (17)$$

We add additional convolution layers with zero initialization [Zhang et al. 2023] to minimize the influence of the expression branch at the beginning of the training.

4 EXPERIMENTS

4.1 Evaluation Settings

4.1.1 Datasets. We conduct our experiments on the FFHQ [Karras et al. 2019] dataset which contains 70K in-the-wild human face images. We follow EG3D [Chan et al. 2022] to crop the image to 512^2 resolution based on facial landmarks, and compute the camera

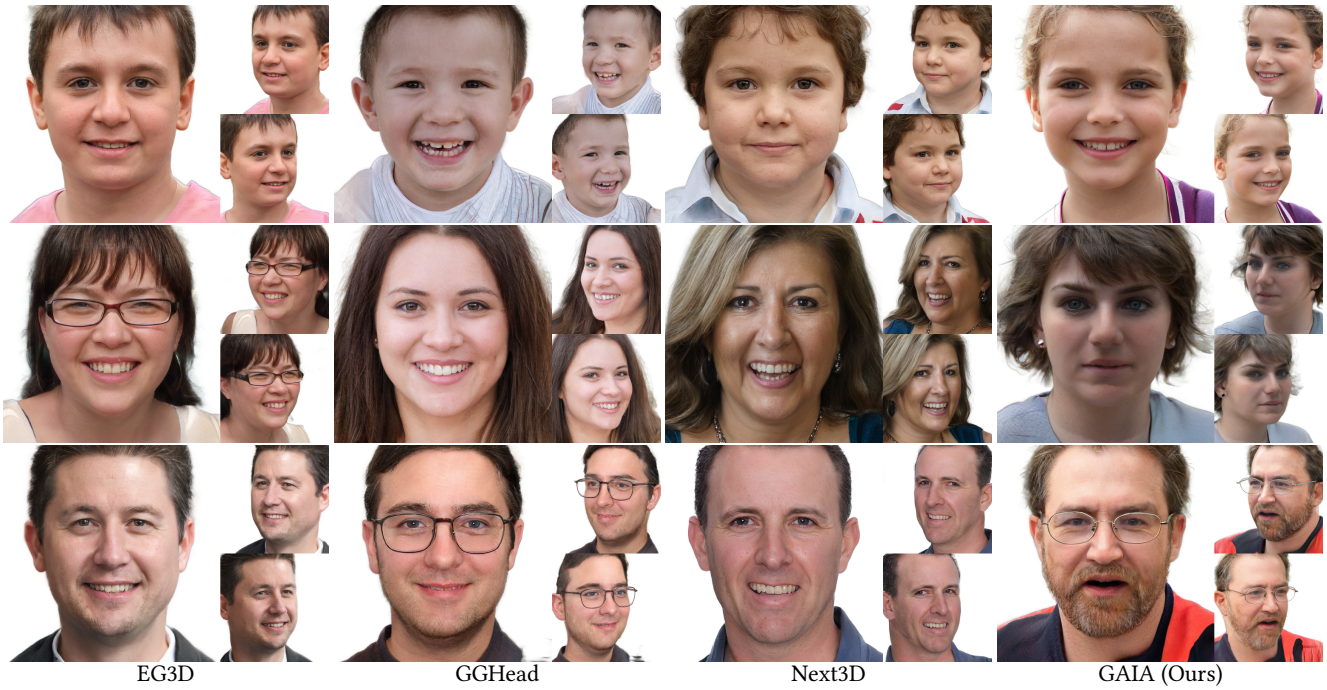


Fig. 3. **Qualitative Comparison on Novel View Synthesis.** We evaluate our generated image quality on random samples. Our method can generate comparable quality images and 3D-consistent results.

poses using 3DMM fitting [Deng et al. 2019]. We further deploy SMIRK [Retsinas et al. 2024] to estimate the FLAME [Li et al. 2017] shape and expression parameters. Finally, we use MODNet [Ke et al. 2022] to remove the background of the images.

4.1.2 Baselines. We compare our method against the following state-of-the-art (SotA) methods. (1) EG3D [Chan et al. 2022] is a SotA method that generates high-quality 3D faces based on a triplane representation and super-resolution network. (2) GGHead [Kirschstein et al. 2024] is a SotA method that synthesizes 3D heads by embedding 3D Gaussian on UV map of a static template. Note that EG3D and GGHead generate avatars based on learned latent code, which does not support animation with factorized control of identity and expression. (3) Next3D [Sun et al. 2023] is a SotA method that can generate animatable 3D faces with learnable neural texture and a super-resolution network to upsample the rendered image.

4.1.3 Metrics. We use Fréchet Inception Distance (FID) [Heusel et al. 2017] to measure image quality. We evaluate the faithfulness of the animation using the Average Expression Distance (AED), Average Pose Distance (APD), and Identity Consistency (ID). Note that Next3D measures the AED (AED-exp) on the FLAME expression parameters which does not include eyelid and jaw. In contrast, We additionally measure AED-eye and AED-jaw to evaluate the animation of eyelid and jaw. We evaluate identity consistency (ID) by calculating the similarity of the ArcFace features [Deng et al. 2019]. We provide additional evaluation details in the *Appendix* A.

4.1.4 Implementation Details. Our model is implemented in PyTorch [Paszke et al. 2019]. We use a batch size of 32 and adopt the learning rate and R1 gradient regularization [Mescheder et al. 2018] from EG3D [Chan et al. 2022]. We set the weight of R1 regularization to 1 in the first stage, and 2 in the second stage. We train stage 1 for 25M iterations at 256^2 resolution and 5M iterations at 512^2 resolution. we then train stage 2 for 4M iterations. In order to train smoothly, we increase λ_{d2} from 0 to 0.5 in 1M iterations and set $\lambda_{d1} = 1 - \lambda_{d2}$. We set the regularization weights as $\lambda_p = 0.1$, $\lambda_s = 0.05$, $\lambda_o = 1$, $\lambda_{uv} = 5$, and $\lambda_{exp} = 60$. The full training takes around 5 days on eight NVIDIA A100 GPUs. We use one-degree spherical harmonics in the generated Gaussians. We use the same extended FLAME model as in [Retsinas et al. 2024], with additional eyelids blendshapes, and extended expression parameters including the original expression, jaw and eyelid parameters. The blinking motion is mostly controlled by these eyelid blendshapes. For inner mouth, we assign blendweights for upper and lower teeth Gaussians following [Qian et al. 2024].

4.2 Results

We provide comparison results and ablation studies to demonstrate our method. We strongly recommend the reader to watch our *Supp. Video*[2] to better evaluate the photorealism of our results. Additional evaluations are provided in the *Appendix* A.

4.2.1 Comparison on Novel View Synthesis. In Fig. 3 and Tab. 1, we compare the novel-view rendering of our generated samples to the

[2]Please visit the project website: https://research.nvidia.com/labs/amri/projects/gaia/.

Table 1. **Quantitative Results on FFHQ Datasets**. As EG3D and GGHead do not support explicit animation control, we mark N/A for their metrics on animation quality. Our method has the best animation quality and comparable image quality with 3D-consistency without 2D super-resolution.

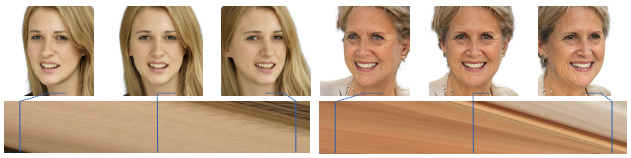
	Method	FID ↓	AED-exp ↓	AED-eye ↓	AED-jaw ↓	APD ↓	ID ↑
w/ SR	EG3D	3.28	N/A	N/A	N/A	N/A	N/A
	Next3D	3.18	0.93	0.149	0.046	0.031	0.74
w/o SR	GGHead	4.06	N/A	N/A	N/A	N/A	N/A
	Ours	3.85	0.53	0.083	0.040	0.027	0.72

Table 2. **Ablation Studies on FFHQ Datasets**. We analyze the effect of different designs of the generation architecture and deformation method.

Method	FID ↓	AED-exp ↓	AED-eye ↓	AED-jaw ↓	APD ↓	ID ↑
One Branch Uncond.	4.94	1.27	0.187	0.062	0.053	0.90
+ Shape Cond.	4.13	0.76	0.125	0.046	0.030	0.84
+ Expr. Cond.	3.81	0.52	0.076	0.039	0.029	0.42
+ Shape & Expr. Cond.	5.52	0.56	0.085	0.039	0.028	0.62
Naive Two-branch	4.21	0.53	0.084	0.041	0.026	0.61
Ours w/ Nearest Blendweight	8.65	0.58	0.095	0.039	0.027	0.71
Ours w/ Surface Field	12.36	0.74	0.136	0.040	0.032	0.76
Ours	3.85	0.53	0.083	0.040	0.027	0.72

Table 3. **Comparison on Inference Speed and Memory Consumption**. For inference time, we measure the average time to generate and render an avatar per frame, as well as the full inference time in frames per second (FPS). The time measurements are averaged over 500 frames. GAIA achieves higher efficiency in both time and memory compared to Next3D.

Method	Generation ↓	Render ↓	FPS ↑	Memory ↓
Next3D	36.61 ms	16.69 ms	18.76	701.3 MB
Ours	21.96 ms	1.09 ms	43.38	463.8 MB



Next3D GAIA (Ours)

Fig. 4. **Analysis of 3D Consistency.** We compare our 3D consistency with Next3D using Epipolar Line Images [Bolles et al. 1987]. Next3D produces noise and ripple artifacts while GAIA renders smoothly without flickering.

existing methods. Our method produces comparable high-quality images with the state-of-the-art methods. Quantitatively, both without the use of 2D super-resolution, GAIA achieves better FID than GGHead as shown in Tab. 1. EG3D and Next3D can generate high-quality images but suffer from 3D view-consistency due to the use of a 2D super-resolution network. To further investigate this, we follow GGHead to compare the Epipolar Line Images (EPI) [Bolles et al. 1987] with Next3D in Fig. 4. For view consistent images, the EPI should be smooth, whereas noise and ripple artifacts reveal the 3D inconsistencies. As shown in Fig. 4, Next3D produces more ripple and noise artifacts in the EPI while ours is smoother.

4.2.2 Comparison on Animation Quality. We compare our method with Next3D in terms of animation quality. As shown in Tab. 1, our method has better AED, APD, and comparable ID consistency compared to Next3D. It shows that our method can animate the avatar more accurately. Although Next3D has a better FID score, it suffers from low 3D consistency due to the use of a 2D super-resolution network. We further show in Fig. 5 that our method outperforms Next3D in terms of animation accuracy and quality. Specifically, Next3D cannot precisely control the eyelid, jaw and eyeball due to the implicit neural texture representation, while our method can accurately animate the avatar with wrinkles in the forehead. This is thanks to our learned animatable features in the expression branch. Furthermore, our method can also control the avatar with accurate eyeball motion due to our TV loss which is applied to the eyeball UV region.

4.2.3 Ablation Studies on Generation Architecture. In Tab. 2, we compare several alternative designs for generating Gaussian attribute maps. The baseline "One Branch Uncond." indicates that only one generator backbone is used, with only the default conditioning variables $\boldsymbol{\pi}$ and $\mathbf{z}$, while using the FLAME mesh during animation. We find that without the additional condition to the generator and discriminator, the model cannot accurately animate the generated avatar, thus this leads to unsatisfactory AED and APD. As shown in Fig. 6, the expressions of the driver are not captured, resulting in an inflated ID consistency score. Next, we add various combinations of additional conditioning variables to the generator. Adding the shape condition (i.e., stage 1 of GAIA) helps the generation and animation with better FID and AED, but the results still lack expression accuracy as shown in Fig. 6. We find that adding the expression condition improves the image quality and expression accuracy, but breaks ID consistency, i.e., the identity of the generated avatar changes during animation. We then try to add both shape and expression conditions on a one branch network to yield better ID consistency but this sacrifices animation accuracy. In order to achieve both high animation accuracy and ID consistency, we use a two-branch network with separate shape and expression conditioning. Without the proposed regularization schemes, the naive two-branch network can achieve accurate animation but lower ID consistency. With our regularization schemes applied on the expression branch, our final model is able to exhibit good animation accuracy and ID consistency.

4.2.4 Ablation Studies on Animation Formulation. We ablate different animation formulations. Nearest Blendweight method, used in [Zhao et al. 2023; Zheng et al. 2022, 2023; Zielonka et al. 2023], deforms the Gaussian points based on the nearest FLAME blendweights. Surface Field is the deformation method proposed in GNARF [Bergman et al. 2022]. As shown in Fig. 6, replacing our deformation method with these two alternative designs produces artifacts in the synthesized results while our method presents robustness.

4.2.5 Runtime and Interactive Animation. We compare the inference time and memory usage to Next3D in Tab. 3. Our method outperforms Next3D in terms of efficiency in both time and memory. Note that we can further accelerate to 52 FPS by caching the output of the generator's shape branch if the identity of the character is constant.

Fig. 5. **Qualitative Comparison on Animation Quality.** We estimate the FLAME parameters from several frames of video clips and use the parameters to animate the generated avatar. Our method shows more precise and detailed animation than Next3D especially wrinkles in forehead and eyeball motion. The two driver video clips are from Next3D [Sun et al. 2023] and IMAvatar [Zheng et al. 2022], respectively.

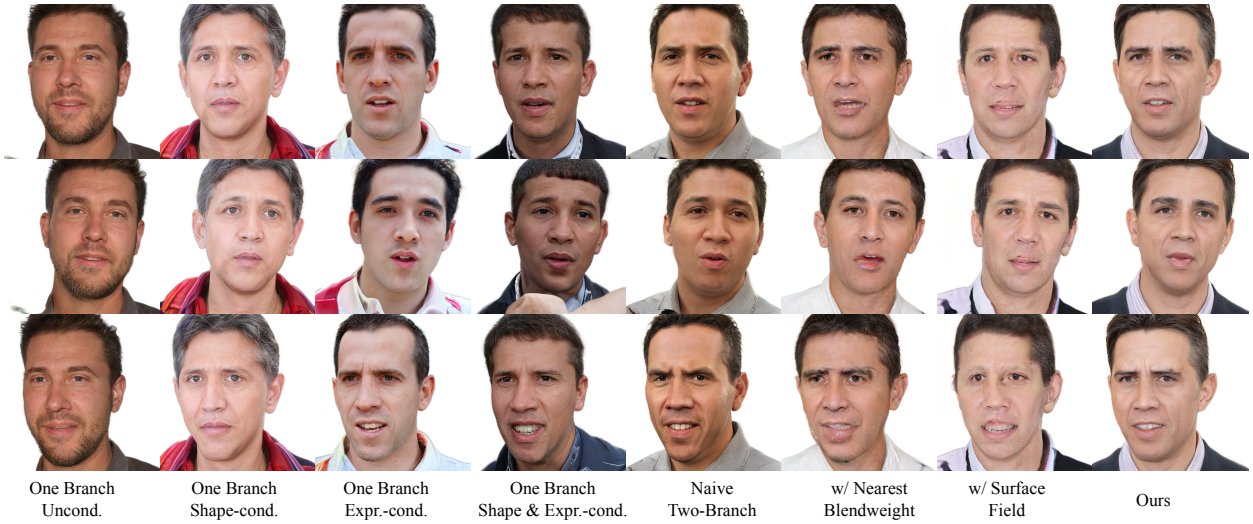


Fig. 6. **Ablation Study.** Our two-branch network with regularization has both good animation and ID consistency compared to other model designs and our deformation method is more robust without producing artifacts. The driver video clip is from Next3D [Sun et al. 2023].

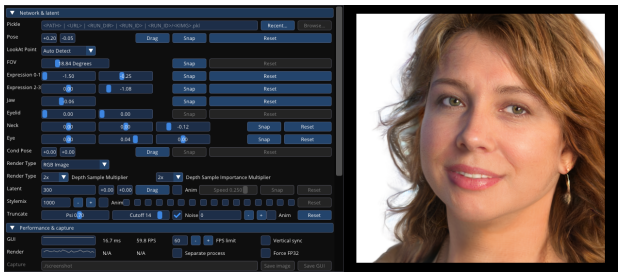


Fig. 7. **Real-time Interactive Animation and Editing.** GAIA supports efficient generation and rendering of photorealistic animation-ready avatars, with an interactive speed of 43 FPS on an NVIDIA RTX A6000 GPU.

As shown in Fig. 7, we develop an interactive viewer that supports real-time animation and editing, showcasing accurate control of expression, identity, and fast rendering possibilities of GAIA.

4.2.6 Limitations and Future Work. While GAIA is effective in the control of the expression and shape of 3D face avatars, it could be further extended to the full body. Similarly, modeling and allowing control of the tongue, face accessories and environmental illumination would further expand application possibilities. While our work allows animation over time, details such as hair and clothing will not deform based on the expectation of real-world physics. To allow physically-based modeling, research into 4D and physics-aware generative models may be useful. We discuss the future work on datasets as well as facial tracker in the *Appendix* A.

5 CONCLUSION

We present GAIA, an advanced 3D GAN framework designed for high-fidelity rendering with exceptional view consistency and interactive animation with intricate deformations, including wrinkles. We achieve this by incorporating a morphable FLAME model with animatable Gaussians through shared UV parameterization, and designing a two-branch generation architecture that learns to apply

shape and expression independently, allowing for disentanglement while animation. GAIA out-performs the state-of-the-art in animatable 3D GANs in both qualitative and quantitative measures, while allowing for animation in real-time, enabling interactive applications involving character animation and editing.

REFERENCES

Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. 2024. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9441–9451.

O Alexander, M Rogers, W Lambeth, J Chiang, W Ma, C Wang, and P Debevec. 2009. The digital emily project: Achieving a photoreal digital actor. *IEEE Computer Graphics and Applications* 30 (2009).

Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems* 35 (2022), 19900–19916.

V Blanz and T Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*. ACM Press, 187–194.

Robert C Bolles, H Harlyn Baker, and David H Marimont. 1987. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision* 1, 1 (1987), 7–55.

Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. 2022. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3981–3990.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.

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. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16123–16133.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5799–5809.

Xingyu Chen, Yu Deng, and Baoyuan Wang. 2023. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2338–2348.

Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. 2024. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.

Xuangeng Chu and Tatsuya Harada. 2024. Generalizable and Animatable Gaussian Head Avatar. *arXiv preprint arXiv:2410.07971* (2024).

Hang Dai, Nick Pears, William Smith, and Christian Duncan. 2020. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision* 128, 2 (2020), 547–571.

Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. 2024b. Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7119–7130.

Yu Deng, Duomin Wang, and Baoyuan Wang. 2024a. Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer. *arXiv preprint arXiv:2403.13570* (2024).

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

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

Zhigang Deng and Junyong Noh. 2008. Computer facial animation: A survey. In *Data-driven 3D facial animation*. Springer, 1–28.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical dynamic facial appearance modeling and acquisition. *ACM Transactions on Graphics (ToG)* 37, 6 (2018), 1–13.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *International Conference on Learning Representations*.

Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. 2019. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9984–9993.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Fangzhou Hong, Zhaoxi Chen, LAN Yushi, Liang Pan, and Ziwei Liu. 2023. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *International Conference on Learning Representations*.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.

Tero Karras, Samuli Laine, and Timo Aila. 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.

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

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1140–1147.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. 2024. GGHead: Fast and Generalizable 3D Gaussian Heads. *arXiv preprint arXiv:2406.09377* (2024).

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

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

Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. 2024. Generalizable one-shot 3D neural head avatar. *Advances in Neural Information Processing Systems* 36 (2024).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751* (2019).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge?. In *International conference on machine learning*. PMLR, 3481–3490.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Koki Nagano, Graham Fyffe, Oleg Alexander, Jernej Barbic, Hao Li, Abhijeet Ghosh, and Paul E Debevec. 2015. Skin microstructure deformation with displacement map convolution. *ACM Trans. Graph.* 34, 4 (2015), 109–1.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7588–7597.

Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. 2020. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems* 33 (2020), 6767–6778.

Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2022. Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision*. Springer, 597–614.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4142–4160.

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

George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 2024. 3D Facial Expressions through Analysis-by-Neural-Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2490–2501.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)* 42, 1 (2022), 1–13.
Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 130–141.
Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.
Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. 2022. Epigraf: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems* 35 (2022), 24487–24501.
Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. 2022. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (ToG)* 41, 6 (2022), 1–10.
Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20991–21002.
Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.
Phong Tran, Egor Zakharov, Long-Nhat Ho, Adilbek Karmanov, Ariana Bermudez Venegas, McLean Goldwhite, Aviral Agarwal, Liwen Hu, Anh Tran, and Hao Li. 2024. VOODOO XP: Expressive One-Shot Head Reenactment for VR Telepresence. *ACM Trans. Graph.* 43, 6, Article 253 (Nov. 2024), 26 pages. https://doi.org/10.1145/3687974
Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. *ACM Trans. Graph.* 42, 4, Article 135 (July 2023), 15 pages. https://doi.org/10.1145/3592460
Alex Trevithick, Matthew Chan, Towaki Takikawa, Umar Iqbal, Shalini De Mello, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2024. What You See is What You GAN: Rendering Every Pixel for High-Fidelity Geometry in 3D GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22765–22775.
Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.
Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. 2022. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems* 35 (2022), 36188–36201.
Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. 2024. FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1802–1812.
Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. 2023. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2195–2205.
Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2024. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.
Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. 2022. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18440–18449.
Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9459–9468.
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. 2023. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics* 43, 1 (2023), 1–16.
Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.
Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21057–21067.
Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4584.

A APPENDIX

A.1 One-shot Avatar Generation with GAN Inversion

With the trained GAIA generators, we adopt the Pivotal Tuning Inversion (PTI) [Roich et al. 2022] to obtain a person-specific 3D avatar given one input image. As shown in Fig. 8, our model can create photorealistic animatable avatars from real-world images.

Fig. 8. **One-shot Avatar Creation.** We use PTI [Roich et al. 2022] to fit an animatable 3D avatar from real-world images.

A.2 Effect of Regularization Parameters

We set the L_2 regularization weight λ_{exp} to various values and evaluate the effects of this hyperparameter. As shown in Tab. 4, with lower λ_{exp}, the model tends to perform better in expression (AED), but lose the consistency in identity (ID). We choose 60 as our final regularization weight as it leads to a better balance between animation accuracy and ID consistency.

Table 4. **Effect of Regularization Weights** λ_{exp}. We compare the quantitative results of different regularization weights of the expression branch.

λ_{exp}	FID ↓	AED-exp ↓	AED-eye ↓	AED-jaw ↓	APD ↓	ID ↑
30	3.95	0.52	0.081	0.040	0.026	0.69
40	4.60	0.52	0.080	0.040	0.026	0.69
50	4.02	0.53	0.082	0.040	0.027	0.70
60 (Ours)	3.85	0.53	0.083	0.040	0.027	0.72
70	4.01	0.53	0.082	0.040	0.027	0.70

A.3 Effect of the Two Generation Branches

In Fig. 9, we visualize the effects of our shape and expression branches. The results show that our expression-conditioned branch captures details (e.g., wrinkles) during animation.

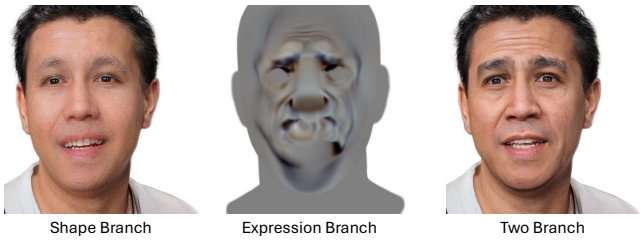


Fig. 9. **Visualization of Effects of the Two Generation Branches.** We visualized the animation with only the shape branch, the expression branch, and two branch. The results indicate that the expression branch captures details (e.g., wrinkles) during animation.

A.4 Results of Training at Higher Resolution

GAIA is able to produce rendering results at 1024^2 resolution. Based on our trained 512^2 resolution model, we finetune our model on 1024^2 resolution images for around 1M iterations. As shown in Tab. 5, we can achieve high-resolution rendering without sacrificing animation accuracy or ID consistency. We further show qualitative results on novel view synthesis results in Fig. 10 and animation in Fig. 11. Our model produces high quality results of novel-view synthesis and animation quality with 1024^2 resolution.

Table 5. **Quantitative Comparison of Training Resolutions**. We finetune our model (trained at 512^2 resolution) on datasets at 1024^2 resolution.

Resolution	FID ↓	AED-exp ↓	AED-eye ↓	AED-jaw ↓	APD ↓	ID ↑
Ours (512)	3.85	0.53	0.083	0.040	0.027	0.72
Ours (1024)	3.92	0.53	0.082	0.040	0.027	0.70

A.5 Additional Evaluation Details

For AED and APD, we randomly sample 10K camera poses and FLAME parameters to generate 10K images. Then we calculate the distance between the input FLAME parameters and the SMIRK [Retsinas et al. 2024] estimated FLAME parameters of these generated images. For identity consistency (ID), we follow Next3D [Sun et al. 2023] to randomly sample 1K identities, each identity with two randomly sampled poses and expression parameters. We then use a pre-trained ArcFace model [Deng et al. 2019] to calculate the similarity of the pair and report the average result.

A.6 Discussion on Animating GANs

Driving GANs with mapping networks for animation is an open problem, as it is tricky to ensure well-behaved animation while supporting intuitive and disentangled control with purely learned latent variables. A typical challenge is that the appearance and sometimes even the gender will change while editing expressions. Furthermore, only specific attribute editing (e.g., smile) is supported. GAIA supports disentangled expression control in FLAME space.

A.7 Additional Discussion on Future Work

The limited view angles, range of expressions and coverage on the mouth region presented in FFHQ can affect the range of rendering and animation. Exploring datasets of wider diversity is a promising direction. Our motion and animation transfer accuracy is bounded by SMIRK, which we used to extract animation parameters. We expect that future improvements in monocular face trackers will transfer to GAIA. Furthermore, as we trained the generator with RGB rendered images with a white background, some white regions (e.g., collar) learned to be transparent to still satisfy the discriminator. Training with RGBA rendering or random background could mitigate this issue.

A.8 Ethical Considerations

The creation of photorealistic, animatable head avatars raises important ethical concerns, particularly on misuse for identity theft, privacy violation, and deepfake-based misinformation. While our work advances the realism and controllability of such avatars, we strongly condemn any malicious or unauthorized use. We emphasize that this is an early step in the field and call for continued research on safeguards, media authentication, and responsible deployment to ensure positive impact on the society.

Fig. 10. **Novel View Synthesis Results at 1024^2 Resolution.** We use our tuned model at 1024^2 resolution to generate some random samples and visualize the novel view synthesis results at 1024^2 resolution.

Fig. 11. **Animation Results at 1024^2 Resolution.** We visualize the animation results of our tuned model in 1024^2 resolution. GAIA maintains high-quality animation at 1024^2 resolution. The two driver video clips are from Next3D [Sun et al. 2023] and IMAvatar [Zheng et al. 2022], respectively.