# PANOLAM : LARGE AVATAR MODEL FOR GAUSSIAN FULL-HEAD SYNTHESIS FROM ONE-SHOT UNPOSED IMAGE

**Anonymous authors**
Paper under double-blind review


Figure 1: PanoLAM creates high-fidelity Gaussian full-heads with one-shot unposed images in seconds.

## ABSTRACT

We present a feed-forward framework for Gaussian full-head synthesis from a single unposed image. Unlike previous work that relies on time-consuming GAN inversion and test-time optimization, our framework can reconstruct the Gaussian full-head model given a single unposed image in a single forward pass. This enables fast reconstruction and rendering during inference. To mitigate the lack of large-scale 3D head assets, we propose a large-scale synthetic dataset from trained 3D GANs and train our framework using only synthetic data. For efficient high-fidelity generation, we introduce a coarse-to-fine Gaussian head generation pipeline, where sparse points from the FLAME model interact with the image features by transformer blocks for feature extraction and coarse shape reconstruction, which are then densified for high-fidelity reconstruction. To fully leverage the prior knowledge residing in pretrained 3D GANs for effective reconstruction, we propose a dual-branch framework that effectively aggregates the structured spherical triplane feature and unstructured point-based features for more effective Gaussian head reconstruction. Experimental results show the effectiveness of our framework towards existing work.

## 1 INTRODUCTION

Reconstructing head avatars from a single image is an important area of research within computer graphics and computer vision. The capability to generate high-fidelity head avatars opens up new possibilities in areas such as 3D telepresence, video conferencing, filmmaking, gaming, and augmented/virtual reality (AR/VR). However, efficiently creating digital head avatars that are high-fidelity and efficient in rendering is challenging, and various studies have tried to resolve this problem.

One line of work uses 3D Generative Adversarial Networks (3D-GANs) (Chan et al., 2022a; An et al., 2023) for the unconditional generation of 3D head avatars, which learn 3D generation models from 2D

images paired with camera poses. Unlike 2D-GANs (Karras et al., 2019) for single-view generation, these approaches can generate face images with control over viewing angles. The pioneering work EG3D (Chan et al., 2022a) introduces a triplane representation in Cartesian space to enable novel view synthesis after generation. The following works (An et al., 2023; Li et al., 2024) allow a wider range of viewing angles from the generated NeRF space and make 360° images synthesis possible. However, these GAN-based frameworks require time- and computation-consuming GAN-inversion (Ko et al., 2023) or test-time optimization (Roich et al., 2023) for image-conditioned generation during inference. Furthermore, the accurate camera pose of the input image is required for joint optimization of these inversion approaches. Rodin (Wang et al., 2023a) and its follow-up (Zhang et al., 2024a) instead utilize diffusion models to generate triplanes of the head avatars. However, the multi-step diffusion process during inference still requires minutes of optimization for each case. Moreover, the costly ray marching hinders the rendering resolution of these NeRF-based frameworks. Though 2D super-resolution upsampling is proposed to improve the rendering efficiency and quality, it compromises 3D consistency.

In this work, we propose a feed-forward framework to generate a full-head Gaussian avatar given a single *unposed* image. This enables fast reconstruction and rendering of a 3D avatar during inference. However, this task presents certain challenges. **1)** The first hurdle lies in the lack of large-scale 3D head datasets for network training. To tackle this problem, we propose to leverage the prior knowledge in trained 3D GANs and sample a large-scale, diverse dataset from EG3D (Chan et al., 2022a) and SphereHead (Li et al., 2024). We label and remove invalid cases manually to improve the quality of the dataset. **2)** The second challenge involves how to efficiently and effectively reconstruct the high-fidelity Gaussian head from a single unposed image. To overcome this, we propose two main designs for our framework. **i)** Firstly, to improve the network efficiency, we propose a coarse-to-fine reconstruction mechanism for high-fidelity Gaussian full-head reconstruction. Specifically, sparse points extract rich image features, craft the coarse shape, and are then upsampled for high-fidelity dense Gaussian head reconstruction. By leveraging the topology prior in the FLAME model, we also introduce an efficient network-free upsampling mechanism for features and point clouds densification. **ii)** Secondly, to improve the effectiveness of the reconstruction process, we propose a dual-branch framework that hybridizes the unstructured point features and the structured spherical triplane features for the reconstruction of a Gaussian head. To distill the spherical triplane knowledge from the pretrained 3D GAN, we analyze its original feature aggregation manner and propose an efficient mechanism to effectively aggregate multi-layer features from the spherical triplane.

The principal contributions of our work can be summarized as follows:

- We propose PanoLAM, a large avatar model for *Gaussian full-head* reconstruction from *a single unposed* image.
- We propose a coarse-to-fine reconstruction mechanism for efficient and high-fidelity Gaussian full-head reconstruction. Built upon the topology prior in the FLAME model, we introduce an efficient network-free upsampling mechanism for points and features densification.
- We propose a dual-branch framework that hybridizes the representation of point and the spherical triplane prior from 3D GAN. An efficient aggregation mechanism is also proposed to effectively extract multi-layer features from the triplane.
- We propose a large-scale, diverse synthesis dataset for 3D avatar reconstruction and generation. To our knowledge, ours is the largest and most diverse 3D head avatar dataset.

## 2 Related Work

**3D Head Modeling Models.** Various 3D representations have been applied for 3D head modeling, including mesh, Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020), Signed Distance Functions (SDFs), and Gaussian Splatting (Kerbl et al., 2023). Mesh-based modelings (Blanz & Vetter, 1999; Paysan et al., 2009; Li et al., 2017) are dedicated to representing 3D human heads with parametric textured meshes and deform the mesh model to improve the topology (Khakhulin et al., 2022; Liao et al., 2025). NeRF-based works (Yu et al., 2023; Ma et al., 2023; Li et al., 2023a; Ki et al., 2024; Bai et al., 2023a; Park et al., 2021a; Zielonka et al., 2023; Zhang et al., 2024b; Zhao et al., 2024; Bai et al., 2023b; Guo et al., 2021; Gao et al., 2022; Park et al., 2021b; Athar et al., 2022; Hong et al., 2021; Tretschk et al., 2021; Gafni et al., 2021) utilize NeRF for 3D head modeling and get more realistic

rendering. To improve geometry quality, SDF-based methods (Zheng et al., 2025; 2022; Zanfir et al., 2024; Canela et al., 2024) are also introduced, where each 3D position is assigned with the signed distance to the nearest surface. For more efficient rendering, the point-based (Zheng et al., 2023) and Gaussian Splatting-based (Qian et al., 2024; Xu et al., 2024; Saito et al., 2024; Ma et al., 2024; Dhamo et al., 2024; Chen et al., 2024) frameworks are also proposed. However, these frameworks usually require minutes to hours of optimization from videos or multiview images with estimated camera pose for each person before usage, limiting their capability of scaling up and applications that require fast reconstruction. Instead, our work introduces a framework that can generate a Gaussian *full-head* avatar from *a single unposed image* in *a single forward pass* within *seconds*.

**3D Avatar Generative Models.** One line of generative models utilizes 3D-aware Generative Adversarial Networks (GANs) (Chan et al., 2022a) to synthesize view-consistent images. Early approaches (Nguyen-Phuoc et al., 2020; 2019; Gadelha et al., 2017; Szabó et al., 2019; Shi et al., 2021; Liao et al., 2020) employ explicit 3D representations like meshes and voxel grids, while more recent studies (Chan et al., 2022a; An et al., 2023) utilize implicit representations for better image quality. EG3D (Chan et al., 2022a) and its follow-ups (An et al., 2023; Li et al., 2024) utilize 3D triplane representation with GANs to generate heads that are capable of novel view synthesis. However, these methods are mainly designed for unconditional generation, and require time- and computation-consuming techniques like GAN inversion (Sun et al., 2022; Chan et al., 2022b; Ko et al., 2023) and test-time optimization (Roich et al., 2023) for image-conditioned generation. The slow optimization procedure during inference and the sacrifice in multi-view consistency hinder their real-world application. Trevithick et al. (2023); Bhattarai et al. (2024); Yuan et al. (2023); Chen et al. (2023) distill an encoding network from EG3D for efficient image-to-triplane generation, but can only generate views near the front face. Instead of using GAN-based frameworks, more recent works (Wang et al., 2023a; Zhang et al., 2024a) utilize diffusion models to generate triplanes of head avatars. However, the multi-step diffusion process is still slow and computation-intensive during inference. Several works (Chu et al., 2024; Abdal et al., 2024; Sun et al., 2023; Kirschstein et al., 2025; Deng et al., 2024; Li et al., 2023b) also introduce generalizable animatable avatar generation, but cannot reconstruct full heads. Moreover, due to the slow rendering speed of NeRF, these approaches require 2D super-resolution networks to enhance image detail, causing view inconsistencies. To enhance rendering quality and speed, recent works (He et al., 2025; Chu & Harada, 2024; Kirschstein et al., 2024; Tang et al., 2024) also utilize Gaussian Splatting (Kerbl et al., 2023) as the 3D representation for generation. LAM (He et al., 2025) introduces a feedforward framework for single-shot animatable Gaussian head generation that can be animated and rendered in real-time on various platforms. However, it requires estimation of FLAME parameters for reconstruction, which is time-consuming, and the generated head can only be rendered from limited viewing angles due to the limited viewing angles in the training 2D videos. Instead, this work can reconstruct the 3D *Gaussian full-head* given a single *unposed* image *in a single forward pass*, thanks to our large-scale 3D training datasets and our coarse-to-fine and dual-branch network design.

# 3 METHODOLOGY

## 3.1 OVERVIEW

As shown in Fig. 2, given an unposed reference image, our framework generates the Gaussian full-head in an efficient coarse-to-fine and joint point-triplane manner, where dense point cloud and point features are densified from sparse ones and fused with the spherical triplane features for dense Gaussian head regression. As shown in the upper left part of Fig. 2, in the coarse stage of the point branch, sparse vertices from the FLAME vertices interact with the extracted multi-level image features by stacked cross-attention modules for feature extraction and coarse Gaussian point reconstruction. The dense reconstruction stage leverages barycentric weights from FLAME subdivision to densify the coarse Gaussian point position and features for dense Gaussian points regression. To fully leverage the prior knowledge residing in the spherical triplane of 3D GAN, as shown in the bottom branch of Fig. 2, a spherical triplane branch is also introduced, and a novel efficient multi-layer aggregation mechanism is proposed to aggregate the structured features from the spherical triplane using the densified point cloud, which are then concatenated with the unstructured point features for high-fidelity Gaussian head reconstruction. To optimize the spherical-triplane branch and distill the prior knowledge, in the bottom right of the framework, a frozen SphereHead decoder is utilized to decode the feature to images for supervision during training.

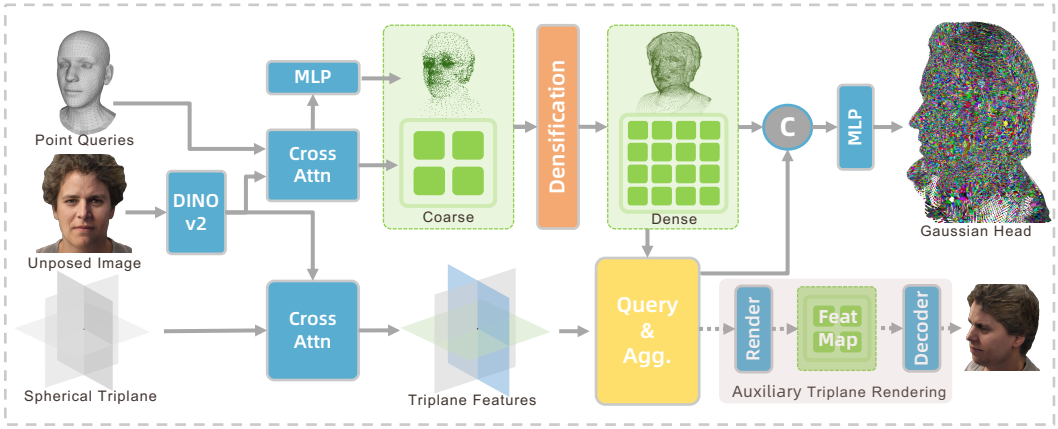


Figure 2: **Overall Framework.** Given an unposed head image as input, PanoLAM involves two branches to achieve single-pass 3D Gaussian head reconstruction: a point-based transformer for *coarse-to-fine* point shape reconstruction and point features extraction, and a spherical triplane transformer to distill prior knowledge from 3D GAN. Features from the two branches are concatenated for high-fidelity Gaussian head regression.

### 3.2 Coarse-to-fine Gaussian Head Avatar Generation

The number of Gaussian points is one key factor that can affect the fidelity of the reconstructed Gaussian head, especially for the head avatar that contains sharp details like hair strands, mustaches, and wrinkles. The total number of Gaussian points should be large enough to produce these details. However, the large number of Gaussian points is hard to optimize during training, especially in our feed-forward pipeline, where the iterative densification operation in the traditional 3D Gaussian Splatting training pipeline is not applicable. To resolve this problem, unlike previous works (He et al., 2025; Zou et al., 2024) that directly regresses the dense point cloud, we introduce a coarse-to-fine mechanism. Specifically, sparse Gaussian points are first reconstructed for coarse shape reconstruction and then densified to dense Gaussians for high-fidelity details crafting. There are two advantages in the proposed design. Firstly, the sparse number of points enables efficient cross-attention with the image features for textures and shape features extraction, saving large computation and memory costs compared with a large number of points. Secondly, the sparse point can be well supervised and optimized for large deformation to carve the shape of the head, especially in regions that require large deformation from the original FLAME model, e.g., long hairs. Dense Gaussian points upsampled from these correctly deformed sparse points can focus more on texture details crafting, which are much easier to optimize.

**Coarse Gaussian Head Generation** To fully leverage the head shape prior that resides in the FLAME model, we initialize the coarse Gaussian point with sparse vertices (5,023 points) on the neutral FLAME model in the canonical space. Our target is to regress the deformation offset for shape refinement of regions that FLAME cannot model, e.g., long hairs, glasses, and caps. To supervise the shape deformation with RGB input images, Gaussian attributes for each point are also regressed for differentiable Gaussian Splatting. To get started, we assign positional embedding to each point and utilize learnable multi-layer perceptron (MLP) modules to project the channel of features into the token channel $C_t$ of transformers as: $F_{P_0} = MLP(\gamma(V))$, where $V$ is the spatial position of each vertice and $\gamma$ the L_frequency sinusoidal encoding as in NeRF. To extract the texture and shape information in the given unposed image, we utilize the pre-trained DinoV2 (Oquab et al., 2024) for feature extraction. Inspired by (Ranftl et al., 2021; He et al., 2025), we fuse features derived from both shallow and deep layers to obtain both local and global image features $F_I$. Specifically, an MLP fuses the $\{5, 12, 18, 24\}$ layers of features in DinoV2 into $C_t$ token channels of features as:

$$F_I = MLP(\mathcal{C}(F_{D_5}, F_{D_{12}}, F_{D_{18}}, F_{D_{24}}), \tag{1}$$

where $F_{D_i}$ is the $i_{th}$ layer of DinoV2 features and $\mathcal{C}$ is the concatenation operation.

Given mapped point features $F_P$ and extracted image features $F_I$, we utilize $L_{\mathcal{A}}$ layers of stacked cross-attention modules $\mathcal{A} = \{\mathcal{A}_i\}_{i=1}^{L_{\mathcal{A}}}$ from Transformer (Vaswani et al., 2017) for feature extraction from image to point cloud as follows:

$$F_{P_i} = \mathcal{A}_i(F_{P_{i-1}}, F_I), \tag{2}$$

where $F_{P_i}$ is the $i_{th}$ layers of point features in the stacked cross-attention layers.

Following the feature extraction process, each point retains its distinct features. Using these features, we develop decoding headers $\mathcal{D}$, composed of multilayer perceptrons (MLPs), to predict the deformation offset $O_k \in \mathbb{R}^3$ to refine the individual's detailed shape. Gaussian attributes for each point are also regressed for rendering, including color $c_k \in \mathbb{R}^3$, opacity $o_k \in \mathbb{R}$, scale factors $s_k \in \mathbb{R}^3$, and rotation $R_k \in SO(3)$. This decoding process is as follows:

$$\{c_k, o_k, s_k, R_k, O_k\}_{k=1}^{M_C} = \mathcal{D}(F_{P_k}), \tag{3}$$

where $M_C = 5023$ represents the total number of Gaussians in the coarse reconstruction stage, and $F_{P_k}$ denotes the extracted point feature for point $P_k$. Although predicting deformation offsets for each point can enhance the shape, freely moving these points may also negatively impact reconstruction results. We therefore restrict the range of deformation offset to be within $[-\epsilon_{O_C}, \epsilon_{O_C}]$.

**Efficient Point Cloud Densification** After the reconstruction of the coarse Gaussian head, we obtain a sparse point cloud that describes the shape of the target head avatar in the image, each point is also attached with rich features extracted from the image. We then densify these points and features for high-fidelity dense Gaussian head reconstruction. Unlike previous methods (Li et al., 2021; Yu et al., 2018) that require a network trained to upsample sparse points, we propose to leverage the topology in the FLAME mesh for efficient densification. Specifically, the proposed strategy utilizes the barycentric coordinates from mesh subdivision to densify the point cloud and feature. Our key observation is that after the deformation in the coarse reconstruction stage, the topology between each point mostly remains as the original FLAME vertices. Thereafter, we can precompute the barycentric coordinates for point cloud densification utilizing mesh subdivision on FLAME. Specifically, given the original FLAME model, we perform subdivision for faces that are larger than a threshold $\lambda_{area}$, the position of newly added vertices $v'$ can be computed from the three vertices of the original face with barycentric coordinates as: $v' = \sum_{i=1}^{3}(w_i \cdot v_i)$, where $w_i$ and $v_i$ are the barycentric weight and vertex coordinate of the original face, respectively. After the deformation in the coarse reconstruction stage, each vertex moves to a new position w.r.t the predicted offset as $p_i = v_i + O_i$. Then new added point $p'$ with its attached point features $p'_F$ can be interpolated with barycentric weights as: $p' = \sum_{i=1}^{3}(w_i \cdot p_i)$, and $p'_F = \sum_{i=1}^{3}(w_i \cdot p_{F_i})$, where $p_{F_i}$ is the extracted features from the last cross-attention blocks in Equ. (2). In this way, we can efficiently densify the point and features.

**Dense Gaussian Head Generation.** After getting the dense point position with dense point features, we can utilize them for dense Gaussian head regression. Rather than using only unstructured point features for prediction, we also aggregate multi-layer of features from the structured spherical triplane for more effective Gaussian attributes regression, which we will discuss in Sec. 3.3. Denote $F_{P_i}$ the $i_{th}$ point feature and $F_{T_i}$ the aggregated spherical triplane feature for point $P_i$, the dense Gaussian head attributes and a small deformation residual can be regressed similarly to the coarse one as:

$$\{c_i, o_i, s_i, R_i, O_i\}_{i=1}^{M_D} = \mathcal{D}(\mathcal{C}(F_{P_i}, F_{T_i})), \tag{4}$$

where $\mathcal{C}$ denotes the concatenation operation, $M_D$ the number of dense Gaussian points, and $O_i$ the deformation residual restricted within $[-\epsilon_{O_D}, \epsilon_{O_D}]$.

### 3.3 Spherical Triplane Prior Knowledge Distillation

In the previous section, we obtain the dense point cloud with its attached point features extracted from the unposed image. To better leverage the prior knowledge within SphereHead's framework, we introduce a spherical triplane branch to distill the knowledge from a pretrained model.

**Preliminary: The Spherical Triplane representation** from SphereHead Li et al. (2024) combines the three triplanes in the Cartesian coordinate system with another three spherical planes in the spherical coordinate system, where features are queried with projection operations in the respective coordinate systems. This representation alleviates the feature entanglement issue in regular triplanes.

In the spherical triplane branch, we initialize the $H_T \times W_T \times 6 \times C_t$ spherical triplane features with learnable query features, which are then flattened into token features $F_T \in \mathbb{R}^{N_T \times C_t}$, where $N_T = H_T \times W_T \times 6$. These learnable features then interact with the image features $F_I$ with stacked

cross-attention blocks $\mathcal{A}_\mathcal{T} = \{\mathcal{A}_{\mathcal{T}i}\}_{i=1}^{L_{\mathcal{A}_\mathcal{T}}}$ similar to the learnable point features as:

$$F_{T_i} = \mathcal{A}_{\mathcal{T}i}(F_{T_{i-1}}, F_I), \tag{5}$$

where $F_{T_i}$ is the $i_{th}$ layer of spherical triplane features in the stacked cross-attention layers. The final output features are then projected by MLPs to the same channel as the spherical triplane in SphereHead. We then need to query the feature from the spherical triplane to enhance the point features. One vanilla way is to utilize the point-based query strategy that fetches each point's spherical-triplane feature by projection. However, we find that our point clouds are distributed on the shape surface similar to a mesh, and such a single-layer query strategy cannot fully aggregate the corresponding features of each point from the spherical triplane. That's because the triplane designed for NeRF-based rendering requires ray marching to aggregate multiple points(/layers) of features for rendering. Using a single point cloud on the shape surface cannot fully fetch valuable features from the spherical triplane. While ray marching is computationally consuming and inefficient, we propose to sample multiple layers of points from the single-layer point cloud for feature aggregation. As is shown in Fig. 3, we sample 4 virtual cameras that can capture the front, left, right, and back sides of the head. For each camera, we cast a ray from the camera center to each point $P_i$, then $K_m$ number of points $P' = \{P'_i\}_{i=1}^{K_m}$ are sampled along the ray near the point. Each point then queries features from the spherical triplane by projection. We then concatenate and aggregate these queried features with MLPs to obtain the final spherical triplane features as:

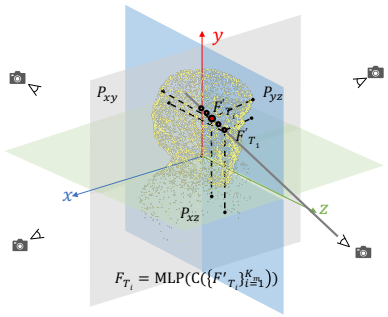


Figure 3: The proposed feature aggregation mechanism samples and aggregates multi-layer features from the spherical triplane for each point.

$$F_{T_i} = MLP(\mathcal{C}(\{F'_{T_i}\}_{i=1}^{K_m})), \tag{6}$$

where $F'_{T_i}$ is the spherical triplane feature queried by sampled point $P'_i$ ; $\mathcal{C}$ is the concatenation operation; and $F_{T_i}$ is the final aggregated feature for point $P_i$. In this way, each point gets its aggregated spherical triplane feature. These aggregated spherical triplane features can be concatenated with the point feature for dense Gaussian attribute regression as in Formula (4). To supervise this branch in distilling knowledge from SphereHead, we add a frozen decoder from SphereHead to decode the feature into the target image, supervised by ground truth images during training, as shown in the bottom right part of Fig 2.

### 3.4 Large-scale Training Data Generation

The scale of the training dataset is a key factor in enabling the generalizability of a trained model. However, obtaining a large dataset with diverse head assets is challenging. Though various datasets (Yin et al., 2006; Zhang et al., 2014b; Hu et al., 2007; Savran et al., 2008; Cao et al., 2014; Cheng et al., 2018; Zhang et al., 2014a; Dai et al., 2020; Kirschstein et al., 2023; Cosker et al., 2011; Zhu et al., 2023; Wang et al., 2022; Pan et al., 2023; He et al., 2024; Martinez et al., 2024) have been proposed for 3D head modeling, as shown in Table 1, they have limited diversity of subjects, which hinders the generalization capability of learning-based models. Inspired by previous methods (Peebles et al., 2022; Cai et al., 2024; He et al., 2022; Wood et al., 2021) in various fields that model trained from synthesis data can generalize to real-world scenarios, we leverage pretrained priors from 3D GANs to generate a large-scale, diverse dataset for training. Specifically, we leverage two prior models for data generation, EG3D for multiview front face images synthesis, and SphereHead for 360° view images generation. Our observation is that EG3D generates a different distribution of images compared to SphereHead (e.g., diverse hairstyles and caps appeared in samples), while SphereHead can generate 360° images. The combination of two model spaces enhances the diversity of the generated images and improves the generalizability of learning-based

Table 1: Comparisons of different 3D avatar datasets. * denotes 3D head with rough captured 3D hair shape.

| Dataset | Sub. | Range | Hair | Dataset | Sub. | Range | Hair |
|---|---|---|---|---|---|---|---|
| BU-3DFE | 100 | front | ✗ | HeadSpace | 1519 | 270° | ✗ |
| BU-4DFE | 101 | front | ✓ | FaceScape | 938 | 360° | ✗ |
| BJUT-3D | 500 | front | ✗ | AVA-256 | 256 | 360° | ✓ |
| Bosphorus | 105 | front | ✗ | FaceVerse* | 128 | 360° | ✓ |
| FaceWarhouse | 150 | front | ✓ | RenderMe360* | 500 | 360° | ✓ |
| 4DFAB | 180 | front | ✗ | SynHead100 | 100 | 360° | ✓ |
| BP4D-S | 41 | front | ✓ | **Ours-Front** | 48,135 | 72° | ✓ |
| Nersemble | 222 | front | ✓ | **Ours-360** | 117,186 | 360° | ✓ |

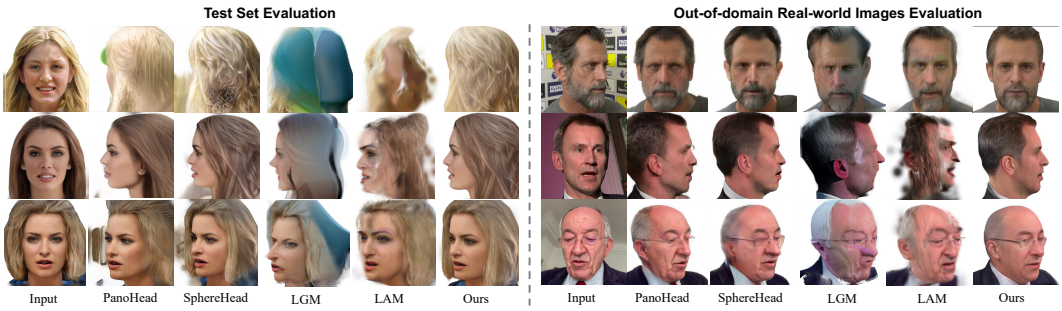


Figure 4: Visualization of reconstruction and novel view synthesis of different methods.

models. However, bad cases occur in the generated dataset, hurting the model training. Therefore, we remove bad cases from the sampled images manually, ending up with 48,135 subjects in Ours-Front and 117,186 subjects in Ours-360 datasets. More details of the dataset construction and labeling pipeline are in Sec E.

### 3.5 Optimization and Regularization

In the training phase, we randomly select $N_v$ different views of images of the same subject. We randomly choose one as the reference image for the Gaussian full-head reconstruction, while the others serve as novel view images for prediction. Note that both branches in our framework render images for supervision. We ensure the accuracy of the rendered RGB images from each branch by comparing them with the ground truth target images, utilizing a combination of $\mathcal{L}_1$ loss and perceptual loss for supervision:

$$\mathcal{L}_{rgb} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{lpips}. \tag{7}$$

Additionally, for the sparse and dense Gaussian Splatting branch, we render the silhouette and supervise it using $\mathcal{L}_1$ loss, referred to as $\mathcal{L}_{mask}$. For the spherical triplane branch, we supervise the extracted feature with those in the pretrained SphereHead by $\mathcal{L}_1$ loss, denoted as $\mathcal{L}_{st}$. The total loss function is a weighted sum as:

$$\mathcal{L} = \lambda_3 \mathcal{L}^G_{rgb} + \lambda_4 \mathcal{L}^{ST}_{rgb} + \lambda_5 \mathcal{L}_{mask} + \lambda_6 \mathcal{L}_{st}, \tag{8}$$

where $\mathcal{L}^G_{rgb}$ denotes the RGB image loss for sparse and dense Gaussian Splatting and $\mathcal{L}^{ST}_{rgb}$ for the auxiliary spherical triplane branch, both denoted in Formula (7).

## 4 Experiments

### 4.1 Experiments Setting

**Implementation Details.** Our framework is implemented in PyTorch, with the DINOv2 image feature extraction backbone frozen. The Transformer comprises 8 layers of basic blocks, featuring 16 attention heads and $C_t = 512$ feature dimensions. The extracted features are translated into Gaussian attributes through three multilayer perceptron (MLP) layers. The network is trained over 100 epochs using the Adam optimizer and a cosine warm-up scheduler. Hyperparameters are empirically set as $N_v = 6$, $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = 1.0$, $\lambda_4 = 0.1$, $\lambda_5 = 1.0$, $\lambda_6 = 0.0001$, $\epsilon_{O_C} = 0.15$, $\epsilon_{O_D} = 0.056$, $M_D = 10K$, $K_m = 32$, $\lambda_{area} = 2.0e-6$, $H_T = W_T = 64$.

**Datasets.** We employ the large-scale dataset that we have generated for training. To assess the 360-degree synthesis capability, we randomly select 100 subjects from our Ours-360 dataset for testing, and the remaining for training.

**Evaluation Metrics.** We evaluate the quality of our synthesized images using three quantitative metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). To evaluate identity similarity (CSIM) of the reconstructed 3D assets, we compute the cosine distance of face recognition features (Deng et al., 2022). For assessing pose fidelity, we employ the Average Pose Distance (APD) (Deng et al., 2019).

### 4.2 Main Results

**Qualitative Results.** Fig. 4 illustrates the novel view synthesis results after reconstruction from various methods on the test set and out-of-domain real-world images, highlighting the superior performance of PanoLAM. Unlike prior approaches, PanoLAM enhances texture detail, maintains identity fidelity, and ensures multiview consistency. Compared to PanoHead and SphreHead, there are no artifacts in the background as well. It is noteworthy that PanoLAM neither employs time-intensive test-time optimization (getting 800X speedup) nor relies on precise camera poses for accurate reconstruction, showcasing the efficacy of our pose-free framework.

**Quantitative Results.** We benchmark our framework and baselines on the test set. For each subject, we evaluate metrics on 4 viewpoints: the front, back, left, and right, except the CSIM metric only uses the front view. We use the default configurations for baselines. As in Table 2, PanoLAM demonstrates exceptional reconstruction quality according to PSNR, SSIM, and LPIPS metrics. The CSIM metrics also indicate strong identity consistency of our methods. Remarkably, these achievements are coupled with fast reconstruction and rendering speeds, underscoring the effectiveness and efficiency of our approach.

Table 2: Quantitative evaluation on the testset.

| Method | PSNR↑ | LPIPS↓ | SSIM↑ | APD↓ | CSIM↑ |
|---|---|---|---|---|---|
| PanoHead-PTI | 16.260 | 0.165 | 0.704 | 0.029 | 0.680 |
| SphereHead-PTI | 16.983 | 0.142 | 0.698 | 0.018 | **0.808** |
| LGM | 11.472 | 0.444 | 0.572 | 0.102 | 0.457 |
| LAM | 15.627 | 0.245 | 0.647 | 0.056 | 0.767 |
| Ours | **23.494** | **0.107** | **0.793** | **0.015** | 0.790 |

Table 3: Running time for 3D avatar generation and rendering.

| Time | PH-PTI | SH-PTI | LAM | Ours |
|---|---|---|---|---|
| Recon.(s) | 89.4 | 96.2 | 1.4 | **0.11** |
| Render(ms) | 10.29 | 19.68 | **3.6** | **3.6** |

Table 4: Effect of Coarse-to-fine generation. OOM: Out of memory (80GB).

| Metrics | D-30K | C2F-30K | D-100K | C2F-100K |
|---|---|---|---|---|
| Infer.(s) ↓ | 0.49 | 0.082 | - | 0.11 |
| Memory(G)↓ | 17.1 | 7.22 | OOM | 12.7 |

Table 5: Module Ablations.

| Method | PSNR↑ | SSIM↑ |
|---|---|---|
| Full | 23.494 | 0.793 |
| w/o C2F | 22.312 | 0.731 |
| w/o ST ray agg. | 22.156 | 0.715 |
| 2K subj. | 19.391 | 0.699 |

**Reconstruction and Render Speed on Various Platforms.** Table 3 shows the comparative running time evaluation of all methods on an A100 GPU. Benefiting from our coarse-to-fine and feedforward framework design, our approach achieves a fast reconstruction speed. The Gaussian Splatting we chose as the 3D representation also enables fast rendering speeds compared to existing techniques.

**Analysis of Feature Sampling and Aggregation Mechanism from Spherical Triplane.** We analyze the feature aggregation mechanism from the spherical triplane of SphereHead to explore a better knowledge distillation strategy. In Fig. 5, we cast rays and visualize the aggregation weights of each sampled feature in the original ray marching. We plot 4 curves showing the weight values of 4 casting rays. Two phenomena occur as in the figure. Firstly, multiple peaks occur near the geometry surface hit by each ray, indicating that we need to sample and aggregate multiple points on a ray to fetch valuable features from the spherical triplane fully. Secondly, each weight curve appears to be a different weight function, indicating that we cannot apply the same manual weight functions, e.g., a Gaussian distribution, for each ray. Thereafter, we propose to leverage learnable MLPs to learn the feature aggregation function from large-scale data. Experimental results in the ablation study validate such a finding and the effectiveness of our design.

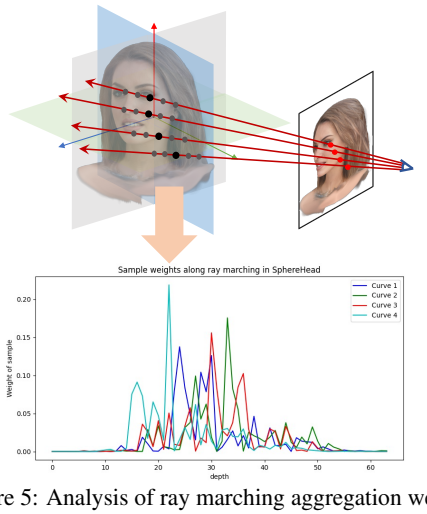


Figure 5: Analysis of ray marching aggregation weights in the original Spherical Triplane in SphereHead.

### 4.3 Ablation Studies

**Effect of Coarse-to-fine Gaussian Point Generation.** Fig. 6 shows the effect of our coarse-to-fine Gaussian generation strategy. The sparse points are easier to optimize and can be easily deformed to carve the shape of the person, but they cannot model texture details like hair strands.

By densifying the points with topology prior in the FLAME mesh, our coarse-to-fine model can learn more detailed textures. In contrast, directly regressing the dense point cloud results in insufficient optimization of Gaussian points. Many points are not well optimized for deformation and remain on the FLAME surface, which are invisible and wasted, causing dual-layer geometry in the hair region as well. The quantitative results in Tab. 5 also support the effectiveness; without our coarse-to-fine training scheme (denoted as w/o C2F in the table), the performance drops. Moreover, as is shown in Table 4, our coarse-to-fine strategy saves computation and GPU memory consumption since fewer points are needed for stacked cross-attention modules. The network forward time and training GPU consumption are also reduced with our coarse-to-fine training strategy.

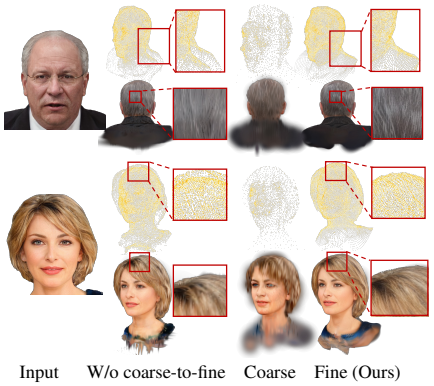


Figure 6: Effect of coarse-to-fine generation.

**Effect of Different Dataset Scale.** Previous 3D head datasets lack subject diversity, as shown in Table 1. We ablate the effect of different numbers of subjects in Fig. 7 and Table 5 (2K subj.). The model trained on a limited number of subjects cannot generalize well to unseen subjects, causing performance drops visually and quantitatively. This shows the effectiveness of our proposed large-scale, diverse dataset for network generalizability.

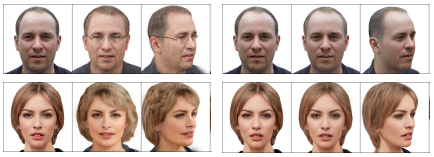


Figure 7: Comparison of our network trained on different numbers of subjects.

**Effect of Different Spherical Triplane Query Strategy.** In Fig. 8, we show the results of different query strategies from the spherical triplane. All strategies use the same feature renderer and frozen decoder from SphereHead for target image decoding. As shown in the figure, directly using the densified point cloud to query a single layer of features from the spherical triplane, as in (Zou et al., 2024), cannot fetch all valuable features, and the rendered images contain artifacts due to missing information. We then add 4 virtual cameras and cast rays from the camera center to each point and sample 32 layers of features near the point surface for feature aggregation. We set the margin between each sampled point to be the same as the fine ray marching stage in SphereHead. We try to assign manual weights sampled from a normal distribution centered at the selected point with standard deviation 1 and 10, and aggregate sampled features with a weighted sum. As is shown in the middle part of Fig. 8, the results look better but are still blurry. Our final solution utilizes learnable MLPs to learn the aggregation function for the NeRF-based rendering. As shown in the figure, ours gets the best results, showing the effectiveness of the proposed aggregation strategy. In Table 5, we also ablate the model without our proposed ray-based feature aggregation from spherical triplane (denoted w/o ST ray agg. in the table). The quantitative results drops and prove the effectiveness of the proposed module.

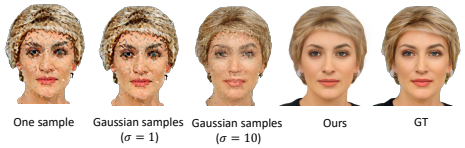


Figure 8: Comparisons on different spherical triplane feature query and aggregation strategies.

## 5 Conclusions

In this work, we present a novel feedforward large avatar model for Gaussian full-head synthesis from a single unposed image. To resolve the problem of lacking 3D assets for network training, we develop a large-scale, diverse 3D avatar dataset from trained 3D GANs. We also propose a coarse-to-fine mechanism with a network-free point cloud densification strategy for efficient reconstruction of a high-fidelity Gaussian full-head from a single image. We also introduce a dual-branch approach to distill knowledge from the spherical triplane prior and improve the reconstruction effectiveness. We analyze the feature sampling mechanism in spherical triplane and propose an efficient aggregation mechanism. Experimental results validate the effectiveness of our approach.

**Ethics Statement.** The generation of realistic avatars from images raises important ethical considerations, especially concerning privacy, consent, and the risks associated with the misuse of technology, such as the creation of deep fakes. It is essential to prioritize responsible use, ensuring adherence to ethical guidelines and standards to mitigate potential negative impacts.

## REFERENCES


Rameen Abdal, Yifan Wang, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9441–9451. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00902. URL https://doi.org/10.1109/CVPR52733.2024.00902.

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Ümit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20950–20959. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02007. URL https://doi.org/10.1109/CVPR52729.2023.02007.

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20332–20341. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01972. URL https://doi.org/10.1109/CVPR52688.2022.01972.

Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. High-fidelity facial avatar reconstruction from monocular video with generative priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 4541–4551. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.00441. URL https://doi.org/10.1109/CVPR52729.2023.00441.

Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, Rohit Pandey, Ping Tan, Thabo Beeler, Sean Fanello, and Yinda Zhang. Learning personalized high quality volumetric head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 16890–16900. IEEE, 2023b. doi: 10.1109/CVPR52729.2023.01620. URL https://doi.org/10.1109/CVPR52729.2023.01620.

Ananta R. Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for EG3D inversion. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 3043–3053. IEEE, 2024. doi: 10.1109/WACV57701.2024.00303. URL https://doi.org/10.1109/WACV57701.2024.00303.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In Warren N. Waggenspack (ed.), *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los Angeles, CA, USA, August 8-13, 1999*, pp. 187–194. ACM, 1999. URL https://dl.acm.org/citation.cfm?id=311556.

Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Open-vocabulary category-level object pose and size estimation. *IEEE Robotics and Automation Letters*, 2024.

Antonio Canela, Pol Caselles, Ibrar Malik, Eduard Ramon, Jaime García, Jordi Sanchez-Riera, Gil Triginer, and Francesc Moreno-Noguer. Instantavatar: Efficient 3d head reconstruction via surface rendering. In *International Conference on 3D Vision, 3DV 2024, Davos, Switzerland, March 18-21, 2024*, pp. 995–1005. IEEE, 2024. doi: 10.1109/3DV62453.2024.00071. URL https://doi.org/10.1109/3DV62453.2024.00071.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.*, 20(3):413–425, 2014. doi: 10.1109/TVCG.2013.249. URL https://doi.org/10.1109/TVCG.2013.249.

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*, pp. 16123–16133, 2022a.

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, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16102–16112. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.01565. URL https://doi.org/10.1109/CVPR52688.2022.01565.

Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2338–2348. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00222. URL https://doi.org/10.1109/ICCV51070.2023.00222.

Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In Andres Burbano, Denis Zorin, and Wojciech Jarosz (eds.), *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, pp. 58. ACM, 2024. doi: 10.1145/3641519.3657499. URL https://doi.org/10.1145/3641519.3657499.

Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5117–5126. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00537. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Cheng_4DFAB_A_Large_CVPR_2018_paper.html.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6a14c7f9fb3f42645cfa6bd5aa446819-Abstract-Conference.html.

Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hgehGq2bDv.

Darren Cosker, Eva Krumhuber, and Adrian Hilton. A FACS valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool (eds.), *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pp. 2296–2303. IEEE Computer Society, 2011. doi: 10.1109/ICCV.2011.6126510. URL https://doi.org/10.1109/ICCV.2011.6126510.

Hang Dai, Nick E. Pears, William A. P. Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *Int. J. Comput. Vis.*, 128(2):547–571, 2020. doi: 10.1007/S11263-019-01260-7. URL https://doi.org/10.1007/s11263-019-01260-7.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. doi: 10.1109/TPAMI.2021.3087709. URL https://doi.org/10.1109/TPAMI.2021.3087709.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 285–295.

Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPRW.2019.00038. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/AMFG/Deng_Accurate_3D_Face_Reconstruction_With_Weakly-Supervised_Learning_From_Single_Image_CVPRW_2019_paper.html.

Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pp. 316–333. Springer, 2024. doi: 10.1007/978-3-031-72643-9_19. URL https://doi.org/10.1007/978-3-031-72643-9_19.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part II*, volume 15060 of *Lecture Notes in Computer Science*, pp. 459–476. Springer, 2024. doi: 10.1007/978-3-031-72627-9_26. URL https://doi.org/10.1007/978-3-031-72627-9_26.

Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 402–411. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00053. URL https://doi.org/10.1109/3DV.2017.00053.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8649–8658. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00854. URL https://openaccess.thecvf.com/content/CVPR2021/html/Gafni_Dynamic_Neural_Radiance_Fields_for_Monocular_4D_Facial_Avatar_Reconstruction_CVPR_2021_paper.html.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Trans. Graph.*, 41(6): 200:1–200:12, 2022. doi: 10.1145/3550454.3555501. URL https://doi.org/10.1145/3550454.3555501.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5764–5774. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00573. URL https://doi.org/10.1109/ICCV48922.2021.00573.

Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. Fs6d: Few-shot 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6814–6824, 2022.

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. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–13, 2025.

Yuxiao He, Yiyu Zhuang, Yanwen Wang, Yao Yao, Siyu Zhu, Xiaoyu Li, Qi Zhang, Xun Cao, and Hao Zhu. Head360: Learning a parametric 3d full-head for free-view synthesis in 360$ˆ\circ $. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, volume 15114 of *Lecture Notes in Computer Science*, pp. 254–272. Springer, 2024. doi: 10.1007/978-3-031-72992-8_15. URL https://doi.org/10.1007/978-3-031-72992-8_15.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *CoRR*, abs/2112.05637, 2021. URL https://arxiv.org/abs/2112.05637.

Yuxiao Hu, ZhenQiu Zhang, Xun Xu, Yun Fu, and Thomas S. Huang. Building large scale 3d face database for face analysis. In Nicu Sebe, Yuncai Liu, Yueting Zhuang, and Thomas S. Huang (eds.), *Multimedia Content Analysis and Mining, International Workshop, MCAM 2007, Weihai, China, June 30 - July 1, 2007, Proceedings*, volume 4577 of *Lecture Notes in Computer Science*, pp. 343–350. Springer, 2007. doi: 10.1007/978-3-540-73417-8_42. URL https://doi.org/10.1007/978-3-540-73417-8_42.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 1867–1874. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.241. URL https://doi.org/10.1109/CVPR.2014.241.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part II*, volume 13662 of *Lecture Notes in Computer Science*, pp. 345–362. Springer, 2022. doi: 10.1007/978-3-031-20086-1_20. URL https://doi.org/10.1007/978-3-031-20086-1_20.

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Learning to generate conditional tri-plane for 3d-aware expression controllable portrait animation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I*, volume 15059 of *Lecture Notes in Computer Science*, pp. 476–493. Springer, 2024. doi: 10.1007/978-3-031-73232-4_27. URL https://doi.org/10.1007/978-3-031-73232-4_27.

Taewoo Kim, Chaeyeon Chung, Sunghyun Park, Gyojung Gu, Keonmin Nam, Wonzo Choe, Jaesung Lee, and Jaegul Choo. K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, pp. 1299–1303. IEEE, 2021. doi: 10.1109/ICIP42928.2021.9506557. URL https://doi.org/10.1109/ICIP42928.2021.9506557.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), jul 2023. ISSN 0730-0301. doi: 10.1145/3592455. URL https://doi.org/10.1145/3592455.

Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In Takeo Igarashi, Ariel Shamir, and Hao (Richard) Zhang (eds.), *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pp. 126:1–126:11. ACM, 2024. doi: 10.1145/3680528.3687686. URL https://doi.org/10.1145/3680528.3687686.

Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *CoRR*,

abs/2502.20220, 2025. doi: 10.48550/ARXIV.2502.20220. URL https://doi.org/10.48550/arXiv.2502.20220.

Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d GAN inversion with pose optimization. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 2966–2975. IEEE, 2023. doi: 10.1109/WACV56688.2023.00298. URL https://doi.org/10.1109/WACV56688.2023.00298.

Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXV*, volume 15133 of *Lecture Notes in Computer Science*, pp. 324–341. Springer, 2024. doi: 10.1007/978-3-031-73226-3_19. URL https://doi.org/10.1007/978-3-031-73226-3_19.

Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Point cloud upsampling via disentangled refinement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 344–353. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00041. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Point_Cloud_Upsampling_via_Disentangled_Refinement_CVPR_2021_paper.html.

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):194:1–194:17, 2017. doi: 10.1145/3130800.3130813. URL https://doi.org/10.1145/3130800.3130813.

Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 17969–17978. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.01723. URL https://doi.org/10.1109/CVPR52729.2023.01723.

Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Tingting Liao, Yujian Zheng, Adilbek Karmanov, Liwen Hu, Leyang Jin, Yuliang Xiu, and Hao Li. Soap: Style-omniscient animatable portraits. In *ACM SIGGRAPH 2025 Conference Proceedings*, 2025.

Yiyi Liao, Katja Schwarz, Lars M. Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 5870–5879. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00591. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Liao_Towards_Unsupervised_Learning_of_Generative_Models_for_3D_Controllable_Image_CVPR_2020_paper.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425. URL https://doi.org/10.1109/ICCV.2015.425.

Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In Andres Burbano, Denis Zorin, and Wojciech Jarosz (eds.), *ACM SIGGRAPH*

*2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, pp. 60. ACM, 2024. doi: 10.1145/3641519.3657462. URL https://doi.org/10.1145/3641519.3657462.

Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 16901–16910. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01621. URL https://doi.org/10.1109/CVPR52729.2023.01621.

Julieta Martinez, Emily Kim, Javier Romero, Timur M. Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason M. Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Maeta, Andrew Jewett, Simion Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matthew Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Timothy Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9712b78386cebdc3db7f1a48c2d20edb-Abstract-Datasets_and_Benchmarks_Track.html.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://doi.org/10.1007/978-3-030-58452-8_24.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 7587–7596. IEEE, 2019. doi: 10.1109/ICCV.2019.00768. URL https://doi.org/10.1109/ICCV.2019.00768.

Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy J. Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4b29fa4efe4fb7bc667c7b301b74d52d-Abstract.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1909ac72220bf5016b6c93f08b66cf36-Abstract-Datasets_and_Benchmarks.html.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5845–5854. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00581. URL https://doi.org/10.1109/ICCV48922.2021.00581.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021b. doi: 10.1145/3478513.3480487. URL https://doi.org/10.1145/3478513.3480487.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In Stefano Tubaro and Jean-Luc Dugelay (eds.), *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, 2-4 September 2009, Genova, Italy*, pp. 296–301. IEEE Computer Society, 2009. doi: 10.1109/AVSS.2009.58. URL https://doi.org/10.1109/AVSS.2009.58.

William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13470–13481, 2022.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 20299–20309. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01919. URL https://doi.org/10.1109/CVPR52733.2024.01919.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 12159–12168. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01196. URL https://doi.org/10.1109/ICCV48922.2021.01196.

Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 42(1):6:1–6:13, 2023. doi: 10.1145/3544777. URL https://doi.org/10.1145/3544777.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 130–141. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00021. URL https://doi.org/10.1109/CVPR52733.2024.00021.

Arman Savran, Nese Alyüz, Hamdi Dibeklioglu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In Ben A. M. Schouten, Niels Christian Juul, Andrzej Drygajlo, and Massimo Tistarelli (eds.), *Biometrics and Identity Management, First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers*, volume 5372 of *Lecture Notes in Computer Science*, pp. 47–56. Springer, 2008. doi: 10.1007/978-3-540-89991-4_6. URL https://doi.org/10.1007/978-3-540-89991-4_6.

Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2d stylegan for 3d-aware face generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6258–6266. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00619. URL https:

//openaccess.thecvf.com/content/CVPR2021/html/Shi_Lifting_2D_StyleGAN_for_3D-Aware_Face_Generation_CVPR_2021_paper.html.

Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Trans. Graph.*, 41(6): 270:1–270:10, 2022. doi: 10.1145/3550454.3555506. URL https://doi.org/10.1145/3550454.3555506.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20991–21002. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02011. URL https://doi.org/10.1109/CVPR52729.2023.02011.

Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *CoRR*, abs/1910.00287, 2019. URL http://arxiv.org/abs/1910.00287.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: large multi-view gaussian model for high-resolution 3d content creation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IV*, volume 15062 of *Lecture Notes in Computer Science*, pp. 1–18. Springer, 2024. doi: 10.1007/978-3-031-73235-5_1. URL https://doi.org/10.1007/978-3-031-73235-5_1.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 12939–12950. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01272. URL https://doi.org/10.1109/ICCV48922.2021.01272.

Alex Trevithick, Matthew A. Chan, Michael Stengel, Eric R. 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 Trans. Graph.*, 42(4):135:1–135:15, 2023. doi: 10.1145/3592460. URL https://doi.org/10.1145/3592460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20301–20310. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01969. URL https://doi.org/10.1109/CVPR52688.2022.01969.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. RODIN: A generative model for sculpting 3d digital avatars using diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 4563–4573. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.00443. URL https://doi.org/10.1109/CVPR52729.2023.00443.

Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Yiqian Wu, Jing Zhang, Hongbo Fu, and Xiaogang Jin. LPFF: A portrait dataset for face generators across large poses. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 20270–20280. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01859. URL https://doi.org/10.1109/ICCV51070.2023.01859.

Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. GRAM-HD: 3d-consistent image generation at high resolution with generative radiance manifolds. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2195–2205. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00209. URL https://doi.org/10.1109/ICCV51070.2023.00209.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Yuelang Xu, Bengwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1931–1941. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00189. URL https://doi.org/10.1109/CVPR52733.2024.00189.

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK*, pp. 211–216. IEEE Computer Society, 2006. doi: 10.1109/FGR.2006.6. URL https://doi.org/10.1109/FGR.2006.6.

Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 2790–2799. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00295. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Yu_PU-Net_Point_Cloud_CVPR_2018_paper.html.

Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and Baoyuan Wu. NOFA: nerf-based one-shot facial avatar reconstruction. In Erik Brunvand, Alla Sheffer, and Michael Wimmer (eds.), *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 85:1–85:12. ACM, 2023. doi: 10.1145/3588432.3591555. URL https://doi.org/10.1145/3588432.3591555.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2437–2447. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00231. URL https://doi.org/10.1109/ICCV51070.2023.00231.

Mihai Zanfir, Thiemo Alldieck, and Cristian Sminchisescu. Phomoh: Implicit photorealistic 3d models of human heads. In *International Conference on 3D Vision, 3DV 2024, Davos, Switzerland, March 18-21, 2024*, pp. 1229–1239. IEEE, 2024. doi: 10.1109/3DV62453.2024.00107. URL https://doi.org/10.1109/3DV62453.2024.00107.

Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIV*, volume 15072 of *Lecture Notes in Computer Science*, pp. 465–483. Springer, 2024a. doi: 10.1007/978-3-031-72630-9_27. URL https://doi.org/10.1007/978-3-031-72630-9_27.

Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial

expression database. *Image Vis. Comput.*, 32(10):692–706, 2014a. doi: 10.1016/J.IMAVIS.2014.06.002. URL https://doi.org/10.1016/j.imavis.2014.06.002.

Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.*, 32(10):692–706, 2014b. doi: 10.1016/J.IMAVIS.2014.06.002. URL https://doi.org/10.1016/j.imavis.2014.06.002.

Zicheng Zhang, Ruobing Zheng, Bonan Li, Congying Han, Tianqi Li, Meng Wang, Tiande Guo, Jingdong Chen, Ziwen Liu, and Ming Yang. Learning dynamic tetrahedra for high-quality talking head synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 5209–5219. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.00498. URL https://doi.org/10.1109/CVPR52733.2024.00498.

Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Trans. Graph.*, 43 (1):6:1–6:16, 2024. doi: 10.1145/3626316. URL https://doi.org/10.1145/3626316.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20311–20320. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01970. URL https://doi.org/10.1109/CVPR52688.2022.01970.

Mingwu Zheng, Haiyu Zhang, Hongyu Yang, Liming Chen, and Di Huang. Imface++: A sophisticated nonlinear 3d morphable face model with implicit neural representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2):994–1012, 2025. doi: 10.1109/TPAMI.2024.3480151. URL https://doi.org/10.1109/TPAMI.2024.3480151.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 21057–21067. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02017. URL https://doi.org/10.1109/CVPR52729.2023.02017.

Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):14528–14545, 2023. doi: 10.1109/TPAMI.2023.3307338. URL https://doi.org/10.1109/TPAMI.2023.3307338.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 4574–4584. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00444. URL https://doi.org/10.1109/CVPR52729.2023.00444.

Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 10324–10335. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00983. URL https://doi.org/10.1109/CVPR52733.2024.00983.

# A Appendix Overview

In this supplement, we present more results in Sec. B, more ablation studies in Sec. C, introduce our dataset generation and labeling process in Sec. E, discuss limitations of our work in Sec. D. Sec. F illustrates the usage of LLM.

# B More Results

In this section, we show more results as well as more comparisons with previous work.

**Quantitative results on out-of-domain real-world datasets.** We conduct experiments on in-the-wild real-world data to evaluate the generalizability of the proposed methods. We sample 9 identities from the VFHQ Xie et al. (2022) test set and 9 identities from the Nersemble Kirschstein et al. (2023) dataset as GaussianAvatar Qian et al. (2024). We evaluate reconstruction quality using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), following the protocol of GRAM-HD Xiang et al. (2023). Note that these data are not seen during training, which is suitable to evaluate the out-of-domain reconstruction capability of different methods. For each method, we synthesize 30 views of the same subject, train the surface reconstruction method NeuS2 Wang et al. (2023b) on these images, and compute PSNR and SSIM on the resulting reconstructions. As shown in Table A1, PanoLam gets the best results on the reconstruction quality. While LAM He et al. (2025) is trained on videos with a limited viewing angle, it cannot reconstruct the full head well, especially in the side and back views, and the results degrade largely in the 360° evaluation. We also visualize some results in the right part of Fig. 4 and Fig. A4.

Table A1: Quantitative evaluation on real-world datasets.

| Method | PSNR↑ | SSIM↑ |
|---|---|---|
| PanoHead-PTI | 24.37 | 0.806 |
| SphereHead-PTI | 24.65 | 0.811 |
| LAM | 18.72 | 0.713 |
| Ours | **27.45** | **0.882** |

**360° synthesis results from our framework** In Fig. A2, we render 360° images from our reconstructed Gaussian full-head given unposed images as input. As shown in the figure, after training on the proposed large-scale dataset, our framework can reconstruct Gaussian full-head and synthesize consistent novel view images.

**More comparison results with previous methods on the testsets.** In Fig. A3, we show more results of different methods on the testset. Four views of images are rendered from the reconstructed 3D representation of different methods, showing the full-head quality. As shown in the figure, our framework gets much higher fidelity results and has fewer artifacts in unseen regions compared to previous works.

**More comparison results with previous methods on real-world images.** In Fig. A4, we show more results of different methods on real-world images sampled from the VFHQ dataset. As shown in the figure, though trained on our synthesis only large-scale dataset, PanoLAM is able to generalize to the real-world images. Compared with previous methods, our framework reconstructs more texture details and maintains better identity fidelity.

# C More Ablation

**Effect of Different Number of Gaussian Points.** We ablate the effect of different numbers of Gaussian points with our coarse-to-fine strategy in Fig. A1. As shown in the figure, a small number of Gaussian points cannot model texture details like hair strands and teeth, leading to blurry results in these regions. In contrast, more high-fidelity Gaussian avatars are reconstructed with an increasing number of Gaussian points. However, directly increasing the number of Gaussian points causes the problem of

Figure A1: Comparison of different numbers of Gaussian points. Zoom in to see more details.

insufficient optimization of Gaussians and out-of-memory issues. This further validate the effectiveness of our proposed coarse-to-fine training strategy.

## D LIMITATIONS

Our dataset and model are built upon 3D GANs trained on different 2D datasets, e.g., FFHQ (Kazemi & Sullivan, 2014) for EG3D and the WildHead (Li et al., 2024), CelebA (Liu et al., 2015), FFHQ, LPFF (Wu et al., 2023), and K-Hairstyle (Kim et al., 2021) datasets for SphereHead. The proposed dataset is sampled from these trained 3D GANs, and the network trained on it has similar biases to these datasets. The sampled datasets contain less Asian data and no cartoon heads, leading to worse reconstruction results on Asian faces and bad results on cartoon faces from the proposed framework. A possible solution to this problem is to train the 3D GANs on a more diverse 2D dataset for more diverse 3D head avatar sampling and 3D avatar reconstruction network training. Our model only handles static full-head reconstruction, and the generated head is not animatable. We leave this as future work.

## E GENERATION AND LABELING PROCESS FOR OUR LARGE-SCALE SYNTHESIS DATASET.

In this section, we introduce our dataset generation process and labeling strategy.

**Dataset generation from trained 3D GANs.** Given that EG3D is limited to rendering near-frontal images using its generated triplane NeRF, we sample images within a view range of approximately $72^\circ$ from this model. Specifically, cameras are sampled from a pitch range of $\pm 26^\circ$ degrees and a yaw range of $\pm 36^\circ$ degrees relative to the front of the human face. To enhance the network's adaptability to real-world images captured by various cameras, we vary the camera radius and focal length. The camera radius is drawn from a normal distribution centered at 2.7 with a standard deviation of 0.1, while the focal length is sampled from a normal distribution centered at 18.83 with a standard deviation of 1. Each triplane generates 32 random images. For SphereHead (Li et al., 2024), we generate $360^\circ$ dataset. Specifically, we initially sample 8 views evenly spaced along the equatorial plane to ensure full-head coverage, dividing the 360-degree range. Additionally, we sample 24 images at random angles and apply the same randomization of camera radius and focal length as used in the EG3D sampling process. Visualization of some cases is shown in Fig. A5.

**Dataset Labeling.** Since the sampled data from 3D GANs may contain bad cases, this may hurt the training of the learning-based network. We also develop a labeling tool and remove bad cases manually to improve the dataset quality. Some examples of bad cases we removed are shown in Fig. A6.

## F STATE OF LLM USAGE

We use LLM to assist with paper polishing on grammar. Each sentence polished by LLM is checked to express our original meaning. There is no further use of LLM for the idea formulation, experiments, coding, and main paper writing.

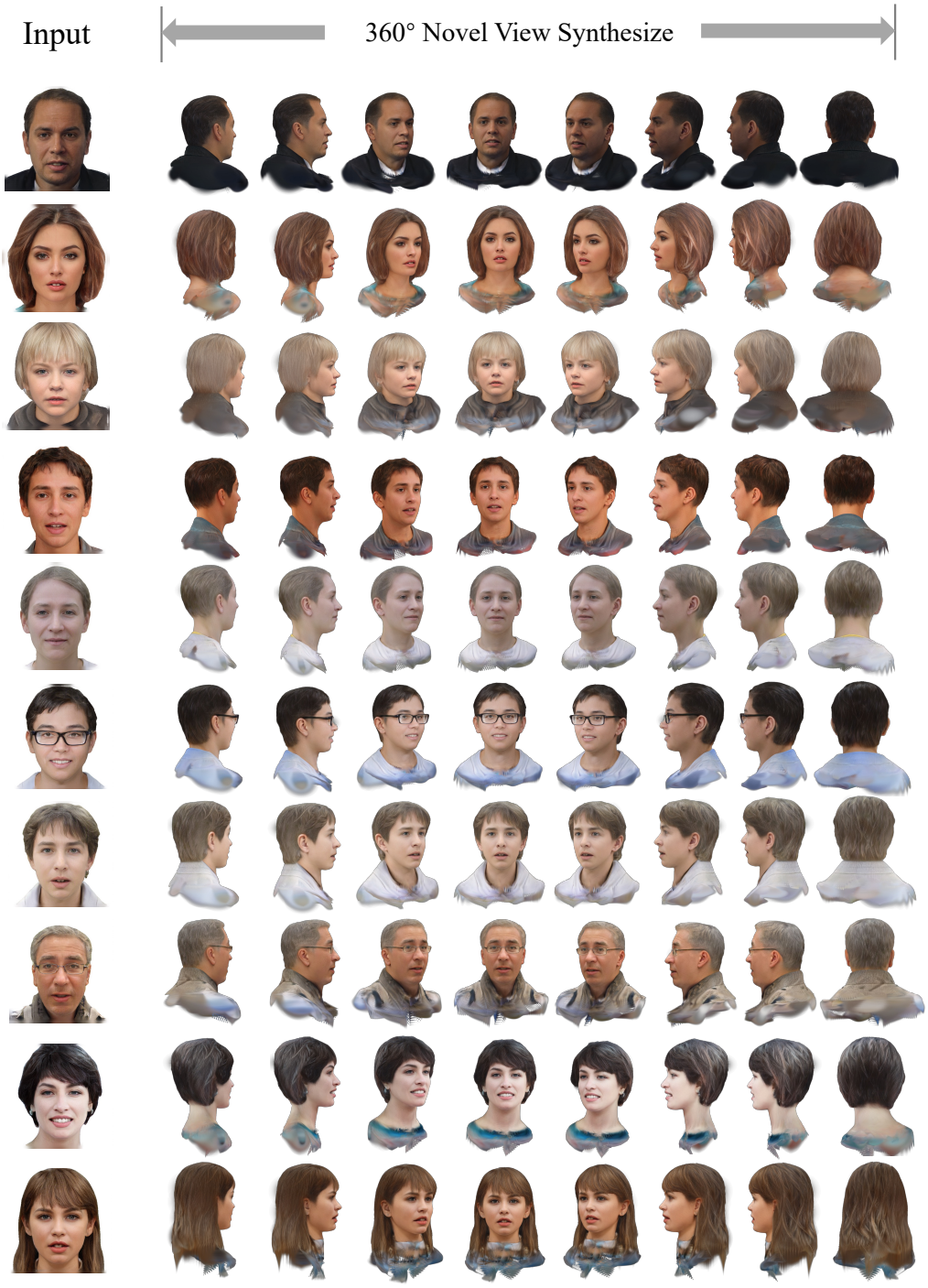


Figure A2: Visualization of our Gaussian full-head synthesis from one-shot unposed image.


Figure A3: Comparison of reconstruction and novel view synthesis of different methods on the testset.

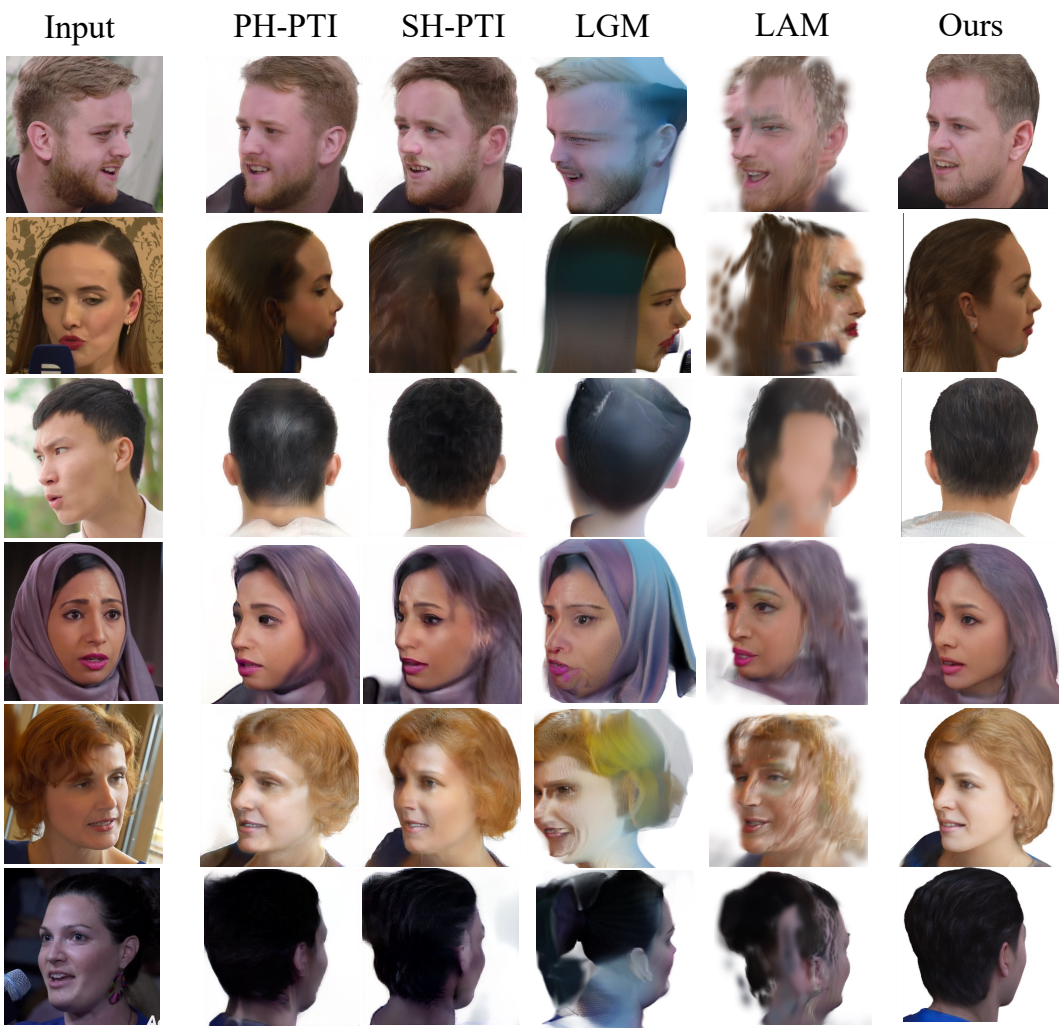


Figure A4: Comparison of reconstruction and novel view synthesis of different methods on in-the-wild real-world images.

Ours-Front from EG3D

Ours-360 from SphereHead

Figure A5: Visualization of example cases in our sampled and cleaned datasets.

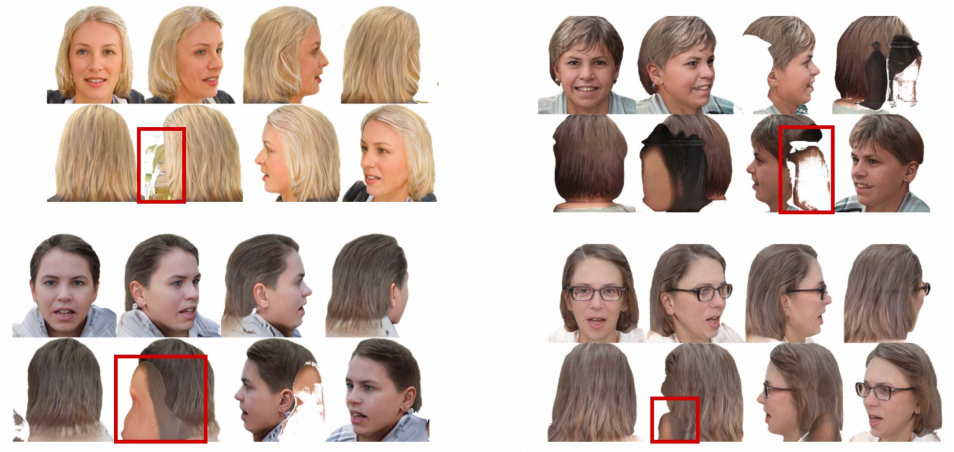

Figure A6: Examples of bad cases sampled from the 3D GANs prior, which are removed manually using our labeling tools.