HQ3DAvatar: High-quality Implicit 3D Head Avatar

KARTIK TEOTIA, Max Planck Institute for Informatics and Saarland University, Saarbrucken, Germany
MALLIKARJUN B R, Max Planck Institute for Informatics and Saarland University, Saarbrucken, Germany
XINGANG PAN, Max Planck Institute for Informatics, Saarbrucken, Germany and Nanyang Technological University, Singapore, Singapore
HYEONGWOO KIM, Imperial College London, London, United Kingdom
PABLO GARRIDO, Flawless AI, Los Angeles, United States of America
MOHAMED ELGHARIB, Max Planck Institute for Informatics, Saarbrucken, Germany
CHRISTIAN THEOBALT, Max Planck Institute for Informatics and Saarland University, Saarbrucken, Germany

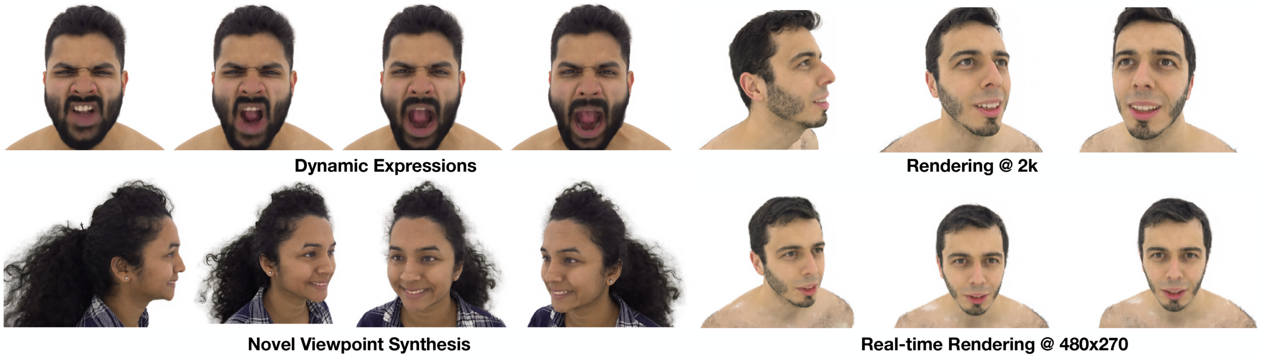


Fig. 1. Our method generates a high-quality 3D head avatar that can be rendered with unseen expressions and camera viewpoints (left). It is trained using multi-view data and multiresolution hash encoding and at test it is driven by a monocular RGB video. Our approach generates 2K full-head renderings (top right) for the first time in literature. It can also run in real time at 480 × 270 (bottom right).

Multi-view volumetric rendering techniques have recently shown great potential in modeling and synthesizing high-quality head avatars. A common approach to capture full head dynamic performances is to track the underlying geometry using a mesh-based template or 3D cube-based graphics primitives. While these model-based approaches achieve promising results, they often fail to learn complex geometric details such as the mouth interior, hair, and topological changes over time. This article presents a novel approach to building highly photorealistic digital head avatars. Our method learns a canonical space via an implicit function parameterized by a neural network. It leverages multiresolution hash encoding in the learned feature space, allowing for high quality, faster training, and high-resolution rendering. At test time, our method is driven by a monocular RGB video. Here, an image encoder extracts face-specific features that also condition the learnable canonical space. This encourages deformation-dependent texture variations during training. We also propose a novel optical flow-based loss that ensures correspondences in the learned canonical space, thus encouraging artifact-free and temporally consistent renderings. We show results on challenging facial expressions and show free-viewpoint renderings at interactive real-time rates for a resolution of $480x270$. Our method outperforms related approaches both visually and numerically. We will release our multiple-identity dataset to encourage further research.

CCS Concepts: • **Computing methodologies** → *Rendering*; *Volumetric models;*

Additional Key Words and Phrases: Volumetric rendering, implicit representations, Neural Radiance Fields, neural avatars, free-viewpoint rendering

ACM Reference Format:
Kartik Teotia, Mallikarjun B R, Xingang Pan, Hyeongwoo Kim, Pablo Garrido, Mohamed Elgharib, and Christian Theobalt. 2024. HQ3DAvatar: High-quality Implicit 3D Head Avatar. *ACM Trans. Graph.* 43, 3, Article 27 (April 2024), 24 pages. https://doi.org/10.1145/3649889

1 INTRODUCTION

The human face is at the center of our visual communications, and hence its digitization is of utmost importance for applications

Authors' addresses: K. Teotia, Mallikarjun B R, and C. Theobalt, Max Planck Institute for Informatics and Saarland University, Saarbrucken, Germany; e-mails: ktoetia@mpi-inf.mpg.de, mallik.jeevan@gmail.com, theobalt@mpi-inf.mpg.de; X. Pan, Max Planck Institute for Informatics, Saarbrucken, Germany and Nanyang Technological University, Singapore, Singapore; e-mail: xingang.pan@ntu.edu.sg; H. Kim, Imperial College London, London, United Kingdom; e-mail: hyeongwoo.kim@imperial.ac.uk; P. Garrido, Flawless AI, Los Angeles, CA; e-mail: pablo.garrido@flawlessai.com; M. Elgharib, Max Planck Institute for Informatics, Saarbrucken, Germany; e-mail: elgharib@mpi-inf.mpg.de.

ACM 0730-0301/2024/04-ART27
https://doi.org/10.1145/3649889

such as Virtual Telepresence. Learning a high-quality controllable 3D digital head is a long-standing research problem with several applications in VR/AR, VFX, and media production, among others. Solutions to this task progressed significantly over the past few years, including early works that create a static textured face model from a monocular RGB camera [Thies et al. 2016], all the way to recent multi-view methods that learn a highly photorealistic model, which can be rendered from an arbitrary camera viewpoint [Lombardi et al. 2021].

Early methods for facial avatar creation are based on explicit scene representations, such as meshes [Kim et al. 2018; Thies et al. 2019a; Zollhöfer et al. 2018]. While these methods produce photorealistic results, they cannot guarantee 3D-consistent reconstructions, as these approaches use 2*D* image-to-image translation models to generate the output RGB reconstructions. Recently, implicit scene representations have significantly attracted the attention of the research community [Tewari et al. 2022]. Due to their inherent characteristics, such as in the case of **Neural Radiance Fields (NeRFs)** [Mildenhall et al. 2020], these models exhibit resilience to alterations in topology (e.g., hairstyles) and are capable of accommodating transparent objects. Moreover, they are inherently designed to maintain 3D consistency. Furthermore, implicit scene representations such as NeRF can be learned from multiple 2D images and produce multi-view consistent renderings. These features make implicit representations suitable for the general task of 3D scene reconstruction and rendering, including human face digitization.

Neural implicit representations [Mildenhall et al. 2020; Park et al. 2019] and, in particular, NeRF, have been used for face digitization due to its high level of photorealism [Athar et al. 2022; Gafni et al. 2021; Zheng et al. 2022]. Here, one of the main challenges is how to model complex facial motions. Faces are dynamic objects and are often influenced by the activation of facial expressions and head poses. An early adaptation of NeRFs, applied to the human face, represents such motion by simply conditioning the implicit function, represented as an MLP, on 3DMM parameters [Gafni et al. 2021]. While this produces interesting results, it has a few limitations, primarily the inability of such 3DMMs to reconstruct high-frequency skin deformations and model the mouth interior. In follow-up methods, a common approach is to model motion by learning a canonical space via template-based deformation supervision [Athar et al. 2022; Zheng et al. 2022]. However, this kind of supervision limits the ability of these methods to accurately model regions not represented by the underlying parametric model, e.g., the mouth interior.

Mixture of Volumetric Primitives (MVPs) [Lombardi et al. 2021] combines the advantage of mesh-based approaches with a voxel-based volumetric representation that allows for efficient rendering. Specifically, it utilizes a template-based mesh tracker to initialize voxels and prune empty spaces. Here, a primitive motion decoder modifies the initialized positions of the primitives. This method produces state-of-the-art results with the highest level of photorealism, mainly due to its hybrid voxel-NeRF representation as well as its capability to train on multi-view video data. However, finding the optimal orientation of the primitives solely based on a photometric reconstruction loss is highly challenging. As a result, this method produces inaccurate reconstructions and artifacts in regions exhibiting fine-scale details such as the hair. It is also expensive to train, requiring around 2.5 days when trained on an NVIDIA A40 GPU.

In this article, we present a novel approach for producing high-quality personalized facial avatars at the state-of-the-art level of photorealism. Our approach uses a voxelized feature grid and leverages multiresolution hash encoding. It is trained using a multi-view video camera setup and, at test time, drives the avatar via a monocular RGB camera. Unlike related methods [Gao et al. 2022; Lombardi et al. 2021], our approach does not require a template to aid in modeling scene dynamics or pruning of empty space. Instead, we learn a fully implicit canonical space that is conditioned on features extracted from the driving monocular video. We regularize the canonical space using a novel optical flow-based loss that encourages artifact-free reconstructions. Our model can be rendered under novel camera viewpoints and facial expressions during inference (see Figure 1, left). It produces highly photorealistic results and outperforms state-of-the-art approaches [Gao et al. 2022; Lombardi et al. 2021; Park et al. 2021b], even on challenging regions such as the scalp hair.

Our contributions are summarized as follows:

- We present a method that leverages a multiresolution hash table to generate volumetric head avatars with state-of-the-art photorealism. The avatar is trained using multi-view data and is driven by a monocular video sequence at test time. The core of our method is an implicitly learned canonical space conditioned on features extracted from the driving video.
- We propose a novel optical flow-based loss to enforce temporally coherent correspondences in the learnable canonical space, thus encouraging artifact-free reconstructions. We also show that our proposed optical flow-based loss helps with novel view synthesis in our sparse camera setup.
- Our model training time is 4–5 times faster than the state-of-the-art [Lombardi et al. 2021]. We show a result with 2K resolution for a volumetric head avatar for the first time in literature. We also show a setting for rendering our results in real time (see Figure 1, bottom right).
- We have collected a novel dataset of 16 identities performing a variety of expressions. The identities are captured using a multi-view video camera setup with 24 cameras. Our multi-view video dataset is the first full-head dataset to be publicly released at 4K resolution, and we will release it to encourage further research.
- We show that the high level of photorealism of our model can even generate synthetic training data at high fidelity, opening the door to generalizing the image encoder to arbitrary input views for driving the avatar.

We evaluate our approach visually and numerically against ground truth data. Here, we ablate our method with different design choices to illustrate their importance in the overall performance. Our approach outperforms the related approaches [Gao et al. 2022; Lombardi et al. 2021; Park et al. 2021b] visually and numerically, including a multi-view implementation of Gao et al. [2022] and Park et al. [2021b].

2 RELATED WORK

This section reviews prior work on photorealistic human head avatar generation, including approaches using monocular or multi-view RGB data. Early methods are based on explicit 3D scene representations, while recent ones leverage implicit representations.

2.1 Monocular Head Avatar Generation

Several monocular avatar generation methods rely on explicit 3D models to estimate or regress a 3D face [Gecer et al. 2019; Lattas et al. 2022; Lin et al. 2020; Ren et al. 2022; Shamai et al. 2019; Tewari et al. 2018; Thies et al. 2019b; Tran et al. 2019; Yamaguchi et al. 2018] or a 3D head containing the face, ears, neck, and hair [Cao et al. 2016; Ichim et al. 2015; Nagano et al. 2018] with photorealistic appearance from 2D images. These methods employ a statistical deformable shape model (a.k.a. 3DMM) of human faces [Cao et al. 2014; Gerig et al. 2018; Li et al. 2017], which provides parametric information to represent the global shape and the dynamics of the face. However, explicit model-based approaches often generate avatars with coarse expressions or facial dynamics and usually lack a detailed representation of the scalp hair, eyes, and/or mouth interior, e.g., tongue. Other approaches attempt to synthesize dynamic full head avatars in a video via generative 2D neural rendering, driven via sparse keypoints [Meshry et al. 2021; Wang et al. 2021b] or dense parametric mesh priors [Chandran et al. 2021; Kim et al. 2018; Tewari et al. 2020; Thies et al. 2019a; Wang et al. 2023]. These methods usually utilize GANs to translate parametric models into photorealistic 2D face portraits with pose-dependent appearance. Still, these methods struggle with fine-scale facial details, and they fail to generate 3D-consistent views.

Recent advances in neural implicit models for personalized head avatar creation from monocular video data have shown great promise. Most approaches learn deformation fields in a canonical space using dense mesh priors [Athar et al. 2022; Bharadwaj et al. 2023; Gao et al. 2022; Zheng et al. 2022, 2023; Zielonka et al. 2023]. Here, Gao et al. [2022], Xu et al. [2023], and Zielonka et al. [2023] leverage multi-level hash tables to encode expression-specific voxel fields efficiently. BakedAvatar [Duan et al. 2023] proposes a hybrid radiance field and rasterization-based framework to produce detailed human head renderings at interactive run-time rates. Similarly, HAvatar [Zhao et al. 2023] proposes a hybrid implicit-explicit rendering framework to produce high-fidelity head renderings. However, these approaches still need to regress to an intermediate expression space defined via 3DMM, thus limiting the representation power.

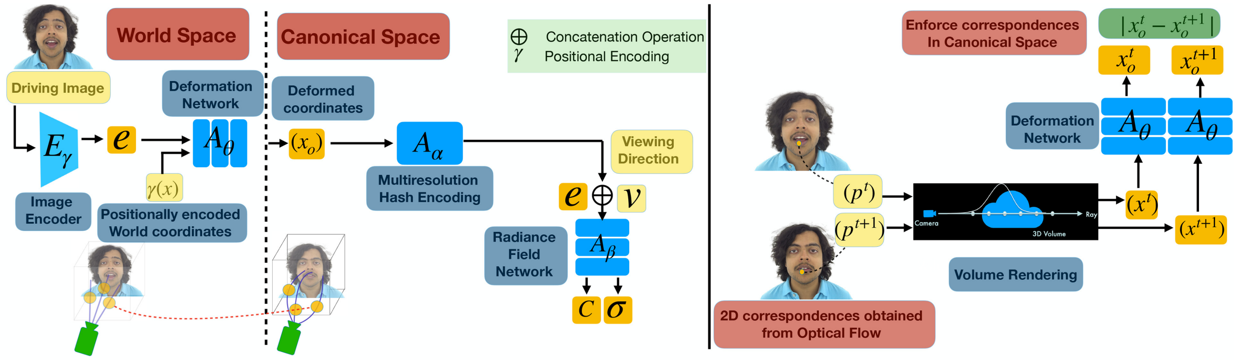
While the above methods generate photorealistic 3D heads with full parametric control, reconstructions can lack dynamics and fine-scale geometrical details, and they cannot handle extreme expressions. However, our approach is not 3DMM-based and thus can model complex geometry and appearance under novel views. This is attributed to our learnable fully implicit canonical space conditioned on the driving video, as well as a novel scene flow constraint.

2.2 Multi-view Head Avatar Reconstruction

A number of approaches leverage multi-view video data to create view-consistent and photorealistic human head avatars with a high level of fidelity. In the literature, we identify approaches that can reconstruct avatars from sparse views (<= 10 high-resolution cameras) or require dense multi-camera systems with dozens of high-resolution views to achieve high-quality results. Due to the large volume of high-resolution video data, recent approaches have also focused on reducing computational and memory costs. Strategies such as efficient sampling [Wang et al. 2021a] and empty space pruning [Lombardi et al. 2021] have been proposed. We also adopt these strategies for efficient and highly detailed rendering at high resolutions.

Sparse multi-view methods. A line of research investigates lightweight volumetric approaches that aim at reducing the number of input views while attempting to preserve the reconstruction fidelity of dense camera approaches. Sparse methods often resort to a canonical space representation [Park et al. 2021a], which serves as a scene template for learning complex non-linear deformations. **Pixel aligned volumetric avatars (PAVA)** [Raj et al. 2021] is a multi-identity avatar model that employs local, pixel-aligned neural feature maps extracted from 3D scene locations. KeypointNeRF [Mihajlovic et al. 2022] is another generalized volumetric avatar morphable model that encodes relative spatial 3D information via sparse 3D keypoints. At inference, both PAVA and KeypointNeRF can robustly reconstruct unseen identities performing new expressions from 2 or 3 input views. TAVA [Li et al. 2022b] encodes non-linear deformations around a canonical pose using a linear blend skinning formulation. TAVA requires 4–10 input views to train a personalized model. While these approaches can generate photorealistic avatars with plausible dynamic deformations from sparse input views, they cannot generate fine-scale details and are sensitive to occlusions, producing rendering artifacts. We demonstrate that regions that undergo sparse sampling can still be reconstructed at high fidelity by imposing temporal coherency via optical flow.

Dense multi-view methods. Early work with dense setups, called **Deep Appearance Models (DAM)** learn vertex locations and view-specific textures of personalized face models via Variational Autoencoders [Lombardi et al. 2018]. **Pixel Codec Avatars (PiCA)** [Ma et al. 2021] improve upon DAM by decoding per-pixel renderings of the face model via an implicit neural function (SIREN) with learnable facial expression and surface positional encodings. The work of Cao et al. [2021] and Chen et al. [2021] demonstrate high-quality textured mesh avatars, especially for the skin surface, driven from commodity hardware. To allow for photorealistic representation of fine details like hair, most recent dense approaches adopt volumetric representations, such as discrete voxel grids [Lombardi et al. 2019], hybrid volumetric models [Lombardi et al. 2021; Wang et al. 2021a], or NeRFs [Wang et al. 2022a]. Here, hybrid approaches combine coarse 3D structure-aware grids and implicit radiance functions, locally conditioned on voxel grids [Wang et al. 2021a] or template-based head tracking with differentiable volumetric raymarching [Lombardi et al. 2021]. In Wang et al. [2022a], a morphable radiance fields framework for 3D head modeling, called MoRF, is proposed. This framework learns statistical face shape and appearance variations from a small-scale database, though it demonstrates good generalization capabilities. The work of Cao et al. [2022] extends MVP



(a) Method Pipeline **(b) Proposed Optical Flow based Loss**

Fig. 2. Left: To extract a robust encoding that parameterizes the dynamics of the head, we pass a driving image through a CNN encoder to obtain a low dimensional vector e. A deformation network A_θ conditioned on e deforms the input coordinates $\gamma(x)$, where $\gamma(.)$ denotes positional encoding. We then use multiresolution hash encoder A_α to encode the deformed points in the canonical space, and feed the features from the hash grid, and encoding e as input to a radiance field network A_β, which outputs density and color values. By combining these values through volume rendering, we are able to render the avatar under unseen input and camera viewpoints. Right: We impose a novel scene flow-based constraint by utilizing the optical flow at frame t and $t+1$ (see Equation (5)). Such constraints enforce good correspondences in the canonical space, thus reducing rendering artifacts.

[Lombardi et al. 2021] to work in a generalized driving setting with non-subject-specific photorealistic avatars. While dense methods produce photo-realistic avatars, renderings tend to exhibit inaccuracies and blur artifacts, especially for complex structures and in infrequently observed areas, such as the scalp hair and mouth interior. Besides, most dense approaches rely on head priors, either mesh tracking or coarse voxel grids, and thus, they are prone to reconstruction errors and have limited representation power, e.g., handling details, mouth interior, and hair. Our approach overcomes existing limitations by solely relying on a well-constrained canonical representation that preserves expression semantics and scene flow correspondences.

2.3 Generalized 3D Consistent Neural Representations

Modeling 3D-aware scenes with implicit models has been active research in recent years. Popular methods are NeRFs [Mildenhall et al. 2020] and neural **Signed Distance Functions (SDFs)** [Park et al. 2019]; both parameterize the 3D space using **multi-layer perceptrons (MLPs)**. Since such methods are often computationally expensive, efficient feature and/or scene space encodings, such as hash grids [Fridovich-Keil et al. 2022; Müller et al. 2022] or trees [Takikawa et al. 2021; Yu et al. 2021], have been proposed to boost performance.

In the literature, generalized implicit models for head avatar reconstruction are learned from a large corpus of 2D face images with varying pose and facial shape using neural SDFs [Or-El et al. 2022; Ramon et al. 2021], GAN-based NeRFs [Bergman et al. 2022; Chan et al. 2021; Deng et al. 2022; Gu et al. 2022], or hybrid volumetric approaches with tensor representations [Chan et al. 2022; Wang et al. 2021a]. Generalized models often lack personalized details. However, they have proven themselves to be robust priors for downstream tasks, such as landmark detection [Zhang et al. 2022], personalized face reenactment [Bai et al. 2022], and 3D face modeling [Abdal et al. 2023].

We remark that NeRFs have stood out as superior implicit representations for head avatar creation, as they excel at reconstructing complex scene structures. Some recent prior-free NeRF-based methods focus on generating detailed avatars from very sparse 2D imagery, e.g., using local pixel-aligned encodings [Mihajlovic et al. 2022; Raj et al. 2021], while others model dynamic deformations when working with unstructured 2D videos by warping observed points into a canonical frame configuration [Park et al. 2021a, b] or modeling time-dependent latent codes [Li et al. 2022a, 2021]. We remark that dynamic approaches, while achieving impressive results, are designed to memorize the scene representations and cannot control the model beyond interpolations. In addition, some approaches build upon dynamic NeRF approaches by incorporating parametric models, e.g., 3DMMs [Egger et al. 2020; Li et al. 2017], as input priors to enable full facial control [Hong et al. 2022; Sun et al. 2022].

3 METHOD

Let $\{I_j^i\}$ $(j = 1 \ldots N, i = 1 \ldots M)$ be multi-view frames of a person's head performing diverse expressions, where N is the number of frames and M is the total number of cameras. Our goal is to create a high-quality volumetric avatar of the person's head, which can be built in a reasonable time and rendered under novel views and expressions at unprecedented photorealism and accuracy.

Humans are capable of performing extremely diverse and extreme expressions. Our model should be able to capture these in a multi-view consistent manner with a high degree of photorealism. As shown in Figure 2(a), we have four components. Our model drives the avatar from a monocular image encoded via a CNN-based image network E_γ. We then have an MLP-based deformation network A_θ, which can map a point in the world coordinate system to a canonical space conditioned on the image encoding. We learn features in the canonical space using a multiresolution hash grid A_α. The features in the grid are interpreted to infer color and density values using an MLP-based network A_β. Given any camera parameters, we use volumetric integration to render the avatar. In the following, we provide details about the capture setup and data pre-processing step (Section 3.1), describe the scene representation of our model (Section 3.2), and formulate various objective functions used for model training (Section 3.4).

3.1 Data Capture

Capture Setting. Our approach is trained using multi-view images captured from a 360-degree camera rig. The rig is equipped with 24 Sony RXO II cameras, which are hardware-synced and record 4K resolution videos at 25 frames per second. The cameras are positioned in such a way that they capture the entire human head, including the scalp hair. The rig is covered by LED strips to ensure uniform illumination. In our setup, we recorded a total of 16 identities performing a wide variety of facial expressions and head movements. Please see Figure 3 for a sample identity captured from multiple viewpoints. For a more detailed description of our dataset, please refer to Section 4.1.

Preprocessing. Cameras are calibrated using a static structure with a large number of distinctive features. Here, we use Metashape [2020] to estimate the extrinsic and intrinsic parameters. We also perform background subtraction using the matting approach of Lin et al. [2021] to remove any static elements from the scene, e.g., wires, cameras. To simplify background subtraction, a diffused white sheet was placed inside the rig, with holes for each of the camera lenses.

3.2 Scene Representation

We parameterize our model using Neural Radiance Fields inspired by the state-of-the-art novel view synthesis method NeRF [Mildenhall et al. 2020]. Since the original method is slow to train and render, we utilize a multiresolution hash grid-based representation to make our model efficient, akin to instant NGP [Müller et al. 2022]. As both original NeRF and instant NGP were proposed for static scene reconstruction, we seek to model the dynamic performance of the head, including facial expressions. To this end, we represent our model, A as

$$A : (\mathbf{x}, \mathbf{v}, \mathbf{e}) \rightarrow (\mathbf{c}, \sigma)\ , \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a point in $3D$, $\mathbf{v} \in \mathbb{S}^2$ is the viewing direction, $e \in \mathbb{R}^{256}$ represents the latent vector obtained from the image encoding network E_γ. This latent vector parameterizes deformations due to expressions and head movements. Furthermore, $\mathbf{c}$ and σ are the color and density values, respectively. Mathematically, instant NGP parameterizes A with two modules. The first module is based on a multiresolution hash grid, denoted A_α, and the second module is parameterized by an MLP, denoted A_β. The latter takes features looked up from A_α and decodes a given point $\mathbf{x}$ and view direction $\mathbf{v}$ into $\mathbf{c}$ and σ. To model dynamic variations of the input driving performance, we introduce another module, denoted A_θ, which takes as input a point in world space and expression latent vector, and regresses a deformation field that converts the world point x to a canonical space, as follows:

$$\mathbf{x}_o = A_\theta(\mathbf{x}, \mathbf{e}) + \mathbf{x}\ . \tag{2}$$

Table 1. Different Parameters Used for Defining the Hash Grid

Parameter	Values
Number of levels	16
Max. entries per level (hash table size)	2^{14}
Number of feature dimensions per entry	2
Coarsest resolution	16
Finest resolution	2,048

We learn the radiance field in this canonical space using A_α and A_β, and we parameterize the operator A_θ using a linear MLP. One could also naively provide the driving image latent code directly to A_β instead of modeling a deformation field to canonical space. However, we show in our experiments (see Section 4.4) that such a naive parameterization creates artifacts. Thus, learning a deformation field is critical in reducing the artifacts. Once we have the radiance field representation of the scene, we use standard volumetric integration to synthesize color $\mathbf{C}$ for each ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, with near and far bounds t_n and t_f, as follows:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))dt\ ,$$
$$\text{where} \quad T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right). \tag{3}$$

Efficient ray marching. As in instant NGP, we improve efficiency by skipping regions that do not contribute to the final color based on the coarse occupancy grid. The occupancy grid typically spans 64^3 resolution, with each cell represented by a single bit. The occupancy grid is updated at regular intervals by evaluating the density of the model in the corresponding region in space. The high in each bit represents the corresponding $3D$ region that has density above a certain threshold. Note that only these regions contribute to the final rendering. As our scene is dynamic, we make certain changes to suit this setting. We initialize G separated occupancy grids corresponding to G uniformly sampled frames. We update each of these grids independently for $200,000$ iterations. Then, we take the union of all the grids to create a single occupancy grid that we utilize for the rest of the training and novel view synthesis. By employing the union operation, we enhance the inference speed; this ensures that only those points exceeding a specified density threshold are evaluated during inference.

3.3 Encoder

Our model is conditioned on a latent vector $\mathbf{e}$ to drive the avatar. In the literature, some methods use expression parameters obtained from face tracking using an existing morphable model [Athar et al. 2022; Gafni et al. 2021]. Other methods parameterize the latent vector obtained from an image encoder [Raj et al. 2021]. While an image encoder might constrain the range of settings for driving the avatar, it has certain advantages over 3DMM-based representations. This includes capturing diverse detailed expressions instead of coarse expression parameters obtained from a 3DMM. Typically, tracking pipelines utilize linear morphable models that have limited expressivity and are prone to tracking errors [B.R. et al. 2021]. In this article, we rely on image encoder E_γ to parameterize the

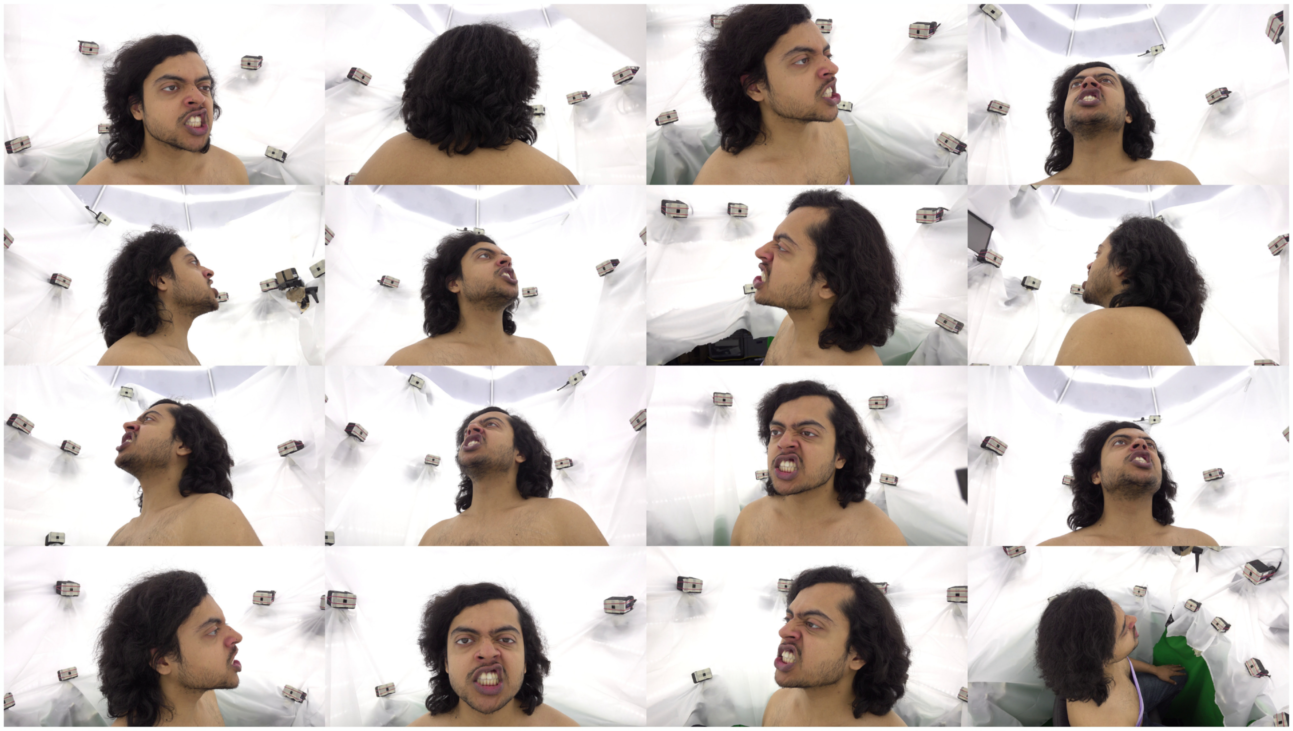


Fig. 3. An example of our camera rig capturing the same expression from 16 different viewpoints.

Table 2. Ablation Study: Image Quality and Perceptual Metrics for Different Design Choices

Metrics	Without canonical space	Without image feature conditioning	Without optical flow-based loss	Ours
PSNR ↑	29.24	29.64	29.38	**31.23**
L1 ↓	3.61	3.64	3.32	**2.79**
SSIM ↑	0.8698	0.8744	0.8517	**0.8837**
LPIPS ↓	0.1408	0.1191	0.1200	**0.1130**

L1 measures the absolute error of unnormalized RGB images. Our full method produces the best results (see bold text).

dynamics of the human head, because it allows us to capture diverse and extreme expressions faithfully, which is the main focus of our article. We parameterize E_γ using a CNN-based network, which takes as input an image **I** of the training subject from a fixed camera viewpoint, and outputs the encoding vector **e**. Specifically, we adopt a pre-trained VGG-Face model [Parkhi et al. 2015] as our encoder and add a custom linear layer at the end. During training, we fine-tune all the VGG layers as well as the custom layer.

3.4 Objective Function

Given the above representation of our model, we learn the parameters of $E_\gamma, A_\theta, A_\alpha$, and A_β modules in a supervised manner using multi-view image and perceptual constraints as well as dense temporal correspondences:

$$\mathcal{L} = \mathcal{L}_{L2} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{of}\mathcal{L}_{of} \ . \tag{4}$$

Reconstruction Losses. Given camera extrinsic and model representation, we render images and employ image reconstruction loss, $\mathcal{L}_{L2}$ using L2 loss between ground truth and rendered images. This term introduces multi-view constraints to train our model. However, L2 loss alone could result in missing some high-frequency details, which are perceptually very important. As a result, we introduce a widely used patch-based perceptual loss $\mathcal{L}_{perc}$, based on a pre-trained VGG Face network [Parkhi et al. 2015]. We use the output of the first 6 layers obtained from an input patch size of 64×64 to compute this loss term.

Optical flow-based Loss. As our dataset consists of sparse views and hash grid-based representation has localized features, a model trained only with $\mathcal{L}_{L2}$ and $\mathcal{L}_{perc}$ losses tend to overfit training views, resulting in artifacts when rendering novel views. To mitigate it, we propose a novel loss term $\mathcal{L}_{of}$ based on pre-computed 2D optical flow between concurrent frames. The motivation behind this loss term is to propagate pixel correspondences to the 3D canonical space with the aim to regularize the dynamic scene and mitigate the model's artifacts when trained with sparser views. We achieve this by enforcing the canonical points of neighboring temporal frames to be close to each other for the points near the surface of the avatar. Mathematically, let p^t, p^{t+1} be the corresponding pixels between consecutive frames obtained using 2D optical

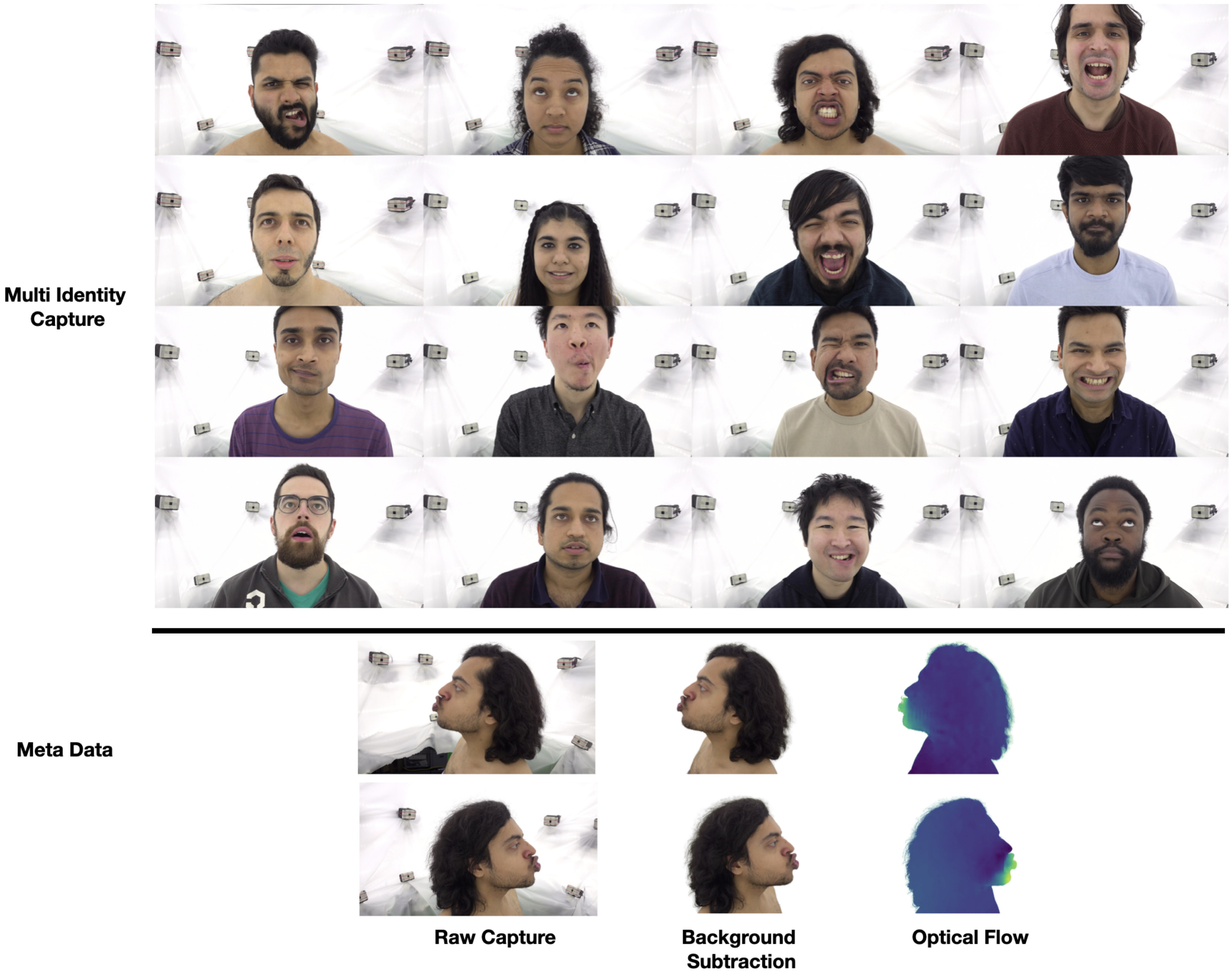


Fig. 4. Top: Visualization of all identities captured in our multi-view camera setup. Our dataset captures a variety of facial hair, hairstyles, expressions, and ethnicities, among others. Bottom: Example of meta data released with our dataset.

flow. For these pixels, we first obtain their corresponding expected depth values through volume rendering. The corresponding $3D$ points x^t, x^{t+1} associated with expected depth can be considered to be close to the surface. We find the corresponding points in the canonical space using A_θ, as defined in Equation (2). Let x_o^t and x_o^{t+1} be the corresponding points in the canonical space. We enforce all such points to be close between them by employing an L1 loss, similar to Kasten et al. [2021]:

$$\mathcal{L}_{of} = \|x_o^t - x_o^{t+1}\|_1 . \tag{5}$$

While multi-view optical flow has previously been used in HVH [Wang et al. 2022b], our formulation does not require explicit tracking of the subject's primitives to utilize the optical flow information. Our formulation instead leverages the fast density updates of the underlying hash-grid-based representation, thus ensuring access to coarse depth information even in earlier stages of the training. Please refer to Figure 2(b) for an illustration of the proposed loss term.

3.5 Implementation Details

We use a 3-layer MLP with 128 neurons as our deformation network A_θ. To encode the coordinates in the world space, we use positional encoding as introduced in Mildenhall et al. [2020], with 10 frequency bands. We provide hash encoding parameters used in our experiments in Table 1. Our radiance field network A_β is parameterized by a 5-layer-deep MLP. It comprises a 2-layer network with 64 neurons that outputs the density feature values $\sigma \in \mathbb{R}^{16}$ and a 3-layer MLP with 64 neurons for regressing the RGB color values. The RGB color values are conditioned on the density features σ and the viewing direction v. The viewing direction v is encoded using spherical harmonics projection on the first four basis functions [Müller et al. 2022]. We set $\lambda_{perc} = 0.1$ and $\lambda_{of} = 0.2$ in

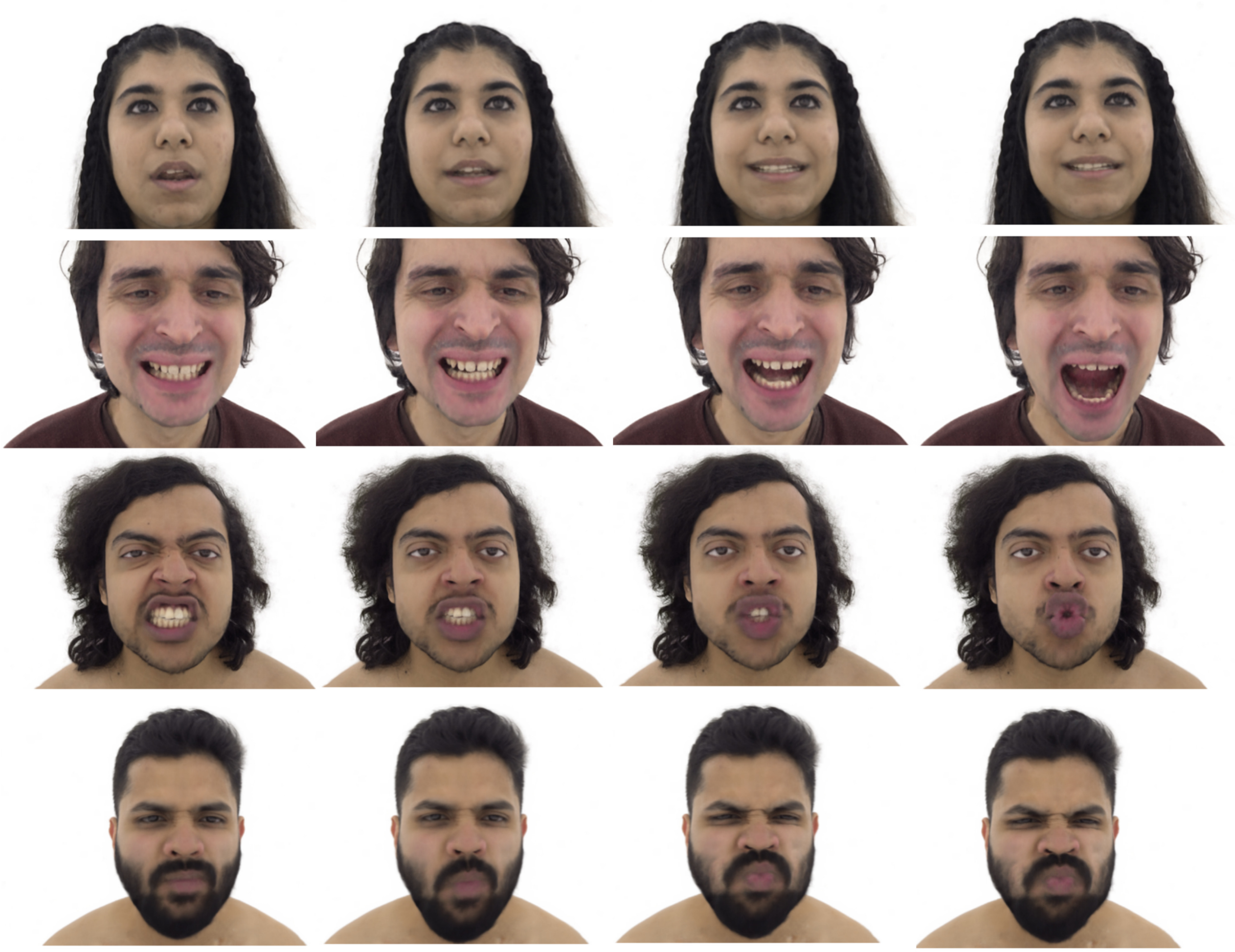


Fig. 5. Qualitative results: Dynamic expression changes. *Top to bottom*: *Subject* 1, 2, 3, and 4.

our experiments. We also follow a PyTorch implementation [Tang 2022] of instant NGP [Müller et al. 2022] to employ error map-based pixel sampling while training, for better convergence. Specifically, we maintain a 128×128 resolution error map for each training image, which is updated in every iteration to reflect the pixel-wise L_2 error. This is then used to sample rays where errors are the highest at each iteration. Finally, we update our encoder E_γ, deformation network A_θ, hash grid A_α, and radiance field A_β with learning rates $1e{-}5$, $1e{-}3$, $1e{-}2$, and $1e{-}3$, respectively. Our model is trained for 500, 000 iterations. We have observed that model convergence is faster than in MVP [Lombardi et al. 2021]. It takes about 12 hours to converge, as opposed to the 50 hours required by MVP with the same GPU resources.

4 EXPERIMENTS

In this section, we show the effectiveness of our high-quality volumetric head avatar reconstruction method in synthesizing novel dynamic expressions and views at high fidelity and resolution. We show two main applications our approach enables, namely, dynamic free-view synthesis from arbitrary monocular viewpoints as well as renderings at different image resolutions, including FHD. We also perform a thorough analysis of our modeling choices and conduct quantitative and qualitative evaluations with state-of-the-art baselines. We refer the reader to the supplemental for video results.

4.1 Datasets

Our multi-view video dataset consists of 16 subjects, including 14 males and 2 females, and most of them are in their 20s or 30s. The subjects have short- to long-length hairstyles. Male subjects either are shaved or have stubble or hairy beards. A collage of the recorded subjects is shown in Figure 4, top. To build our dynamic dataset, we instructed subjects to perform random expressive faces during 2 minutes and/or recite 47 phonetically balanced sentences. Among the 16 subjects, 4 have only performed expressions, 1 has only performed reciting, while 11 have performed both. We will

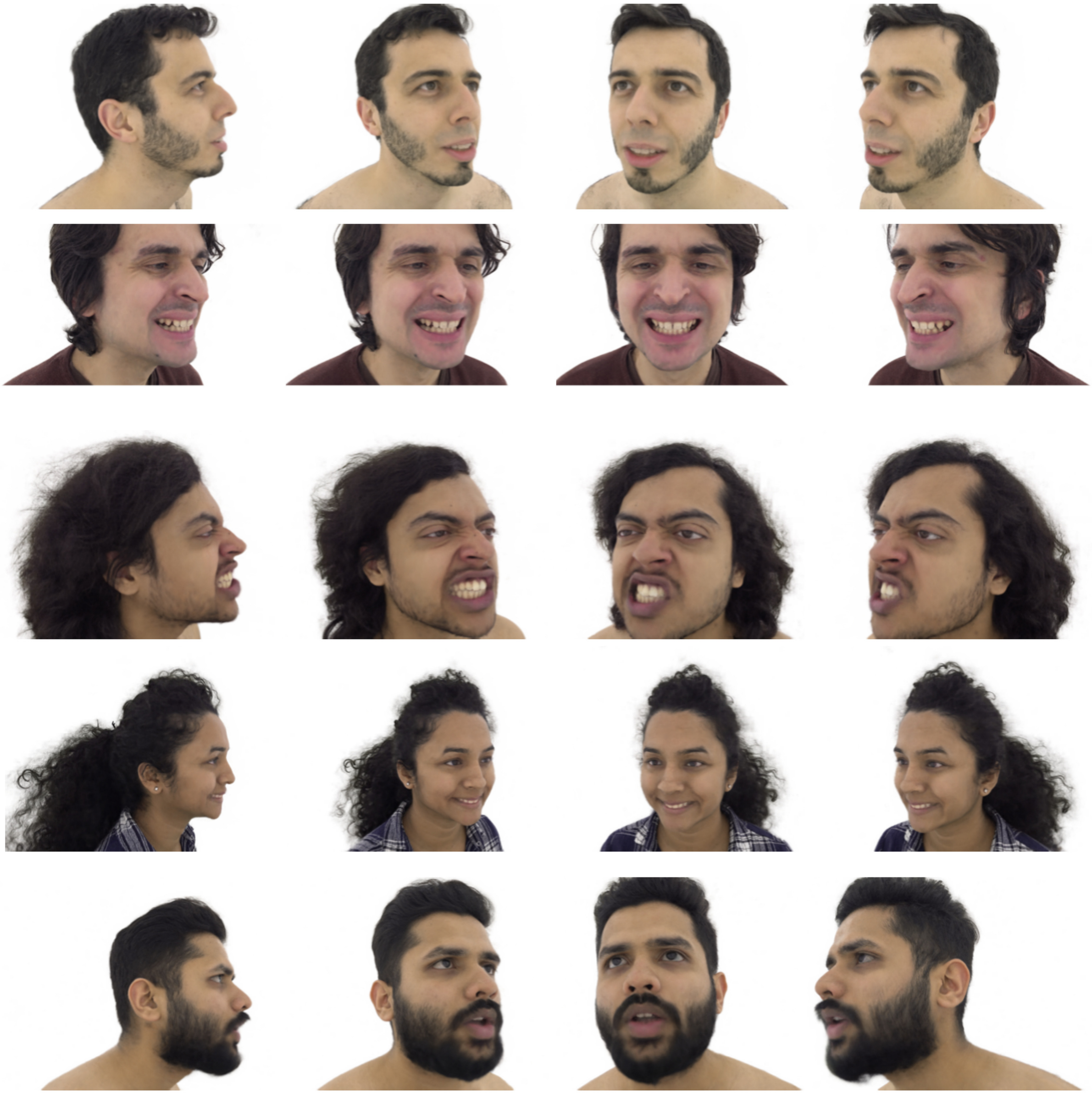


Fig. 6. Qualitative results: Dynamic novel view synthesis for different subjects. *Top to bottom*: *Subject* 5, 2, 3, 6, and 4.

release our full multi-view video dataset to foster future research on head avatar generation. For all of our experiments reported next, we utilize 18 views, each containing 1700 consecutive frames at 960 × 540 resolution. To train our personalized models, we train on the first 1,500 frames from the dataset and evaluate on the last 200 frames. Additionally, we hold out 2 views for quantitative evaluation, while 16 views are used for training. For qualitative results, we use all 18 views. We processed 9 subjects covering a wide variety of our dataset, e.g., gender, expressions, facial hair, hairstyles, ethnicity.

4.2 Qualitative and Quantitative Results

Our experiments involve two types of sequences: extreme expressions and speaking sequences.

(1) Extreme Expressions: Subjects cycle through a predetermined set of expressions.
  - Training: Initial 1,500 frames.
  - Evaluation: Subsequent 200 frames.

(2) Speaking Sequences: This is further divided into:
  (a) Panagram-speaking: Subjects speak 10 predetermined sentences.

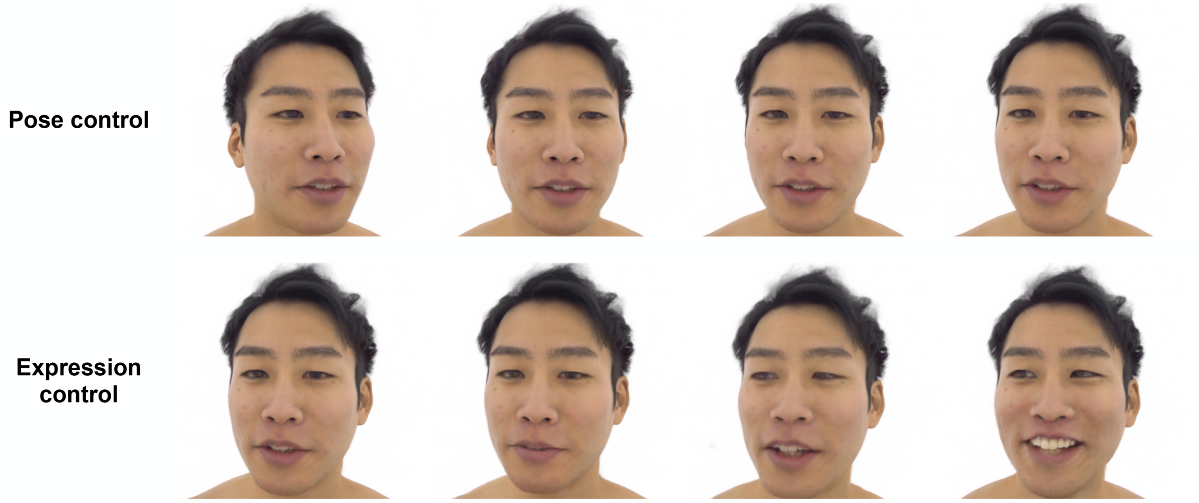


Fig. 7. Pose and expression control of our avatar. *Top:* We fix the image input and change the rigid head-pose. *Bottom:* We change expression while fixing the rigid head-pose.

— Training: 10 sentences.
— Evaluation: 3 held-out sentences.

(b) Free Speaking: Subjects speak about a topic for about 2 minutes.
— Training: First 1,500 frames.
— Evaluation: Following 400 frames.

Figure 5 shows dynamic expression synthesis of 4 personalized avatar models on test sequences, while Figure 6 illustrates free viewpoint synthesis of 5 personalized models. Note that the generated views represent interpolations from training views. Figure 7 demonstrates that given rigid head-pose information, we can control the head-pose independently from the expressions by applying the rigid transformation to the camera parameters. In these figures, the avatars are driven by a frontal-looking monocular RGB video. Our approach achieves high-quality renderings of head avatars under novel camera viewpoints and for challenging novel expressions. Table 2 shows that our approach on average obtains high PSNR (over 31 dB) and low reconstruction errors on test sequences based on different image quality and perceptual metrics. Please see the supplemental for video results.

4.3 Applications

Avatar Synthesis from an Arbitrary Monocular Viewpoint. In previous experiments, we have shown that we can drive our head avatar using a monocular video captured from a frontal view. Here, we further show an application where we can drive our head avatar from an arbitrary viewpoint. To achieve this, we define a fine-tuning scheme described as follows: First, we synthesize a training dataset from a novel viewpoint, say, $\hat{v}$, with the personalized avatar model described in Section 3. This synthetic data generation at the holdout viewpoint for 1,500 frames takes about 3 minutes. This dataset contains the same dynamic expressions used for training. Then, we fine-tune the image encoder with this synthetic video stream for 100k iterations, which takes about 2 hours. Note that the deformation and radiance field networks as well as the multiresolution hash encoding remain unchanged. Once the image encoder has been fine-tuned, we can drive the personalized avatar model with the real data stream coming from the viewpoint $\hat{v}$. In our experiments, $\hat{v}$ is a held-out viewpoint not used when training the avatar model.

Figure 8 compares frontal renderings of Subject 3's avatar model, driven from two video streams with unseen expressions: one driven from a frontal view camera and another driven from a held-out bottom view. Our method produces high-fidelity renderings regardless of the driving video viewpoint, and the rendered expressions faithfully reproduce those shown in the driving video. This demonstrates that our personalized avatar model can generate photo-realistic renderings from arbitrary viewpoints at high fidelity. These renderings can be used as a good approximation of real images to fine-tune the image encoder from arbitrary driving viewpoints. Note that this experiment paves the way for learning high-fidelity personalized avatars that can be driven from video captured in the wild.

FHD Image Synthesis. Our multiresolution hash grid encoding allows for training a personalized avatar model at full HD resolutions, which surpasses the capabilities of state-of-the-art approaches. Our method can render HD images (960 × 540) at about 10 fps and FHD (1,920 × 1,080) images a bit below 3 fps. Figure 9 compares renderings of personalized models trained at HD and FHD resolutions. Both models generate visually similar facial features and details, though the FHD model produces crisper results, as expected. Overall, our approach scales well, and the decrease in runtime is near linear. Figure 10 shows that our approach can also

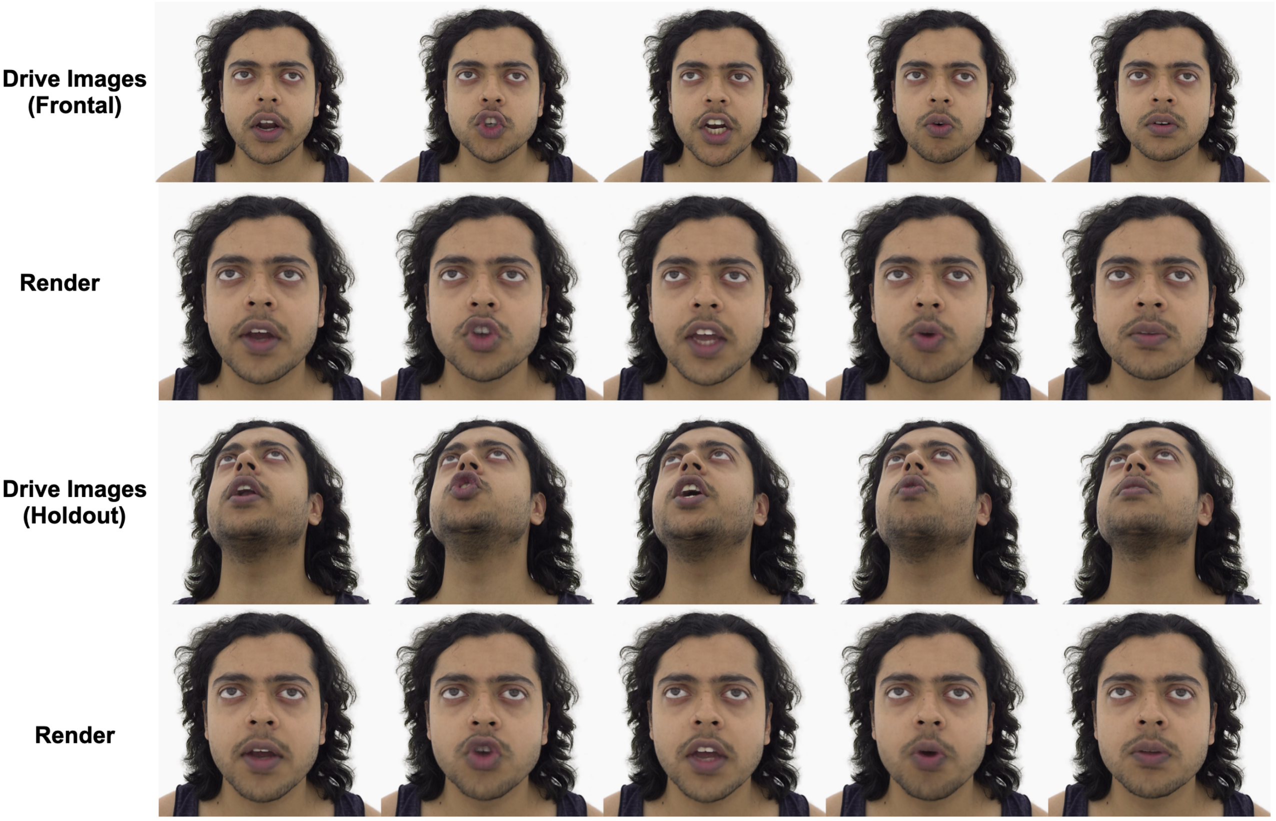


Fig. 8. Avatar synthesis from different driving viewpoints. *Top*: Frontal view driving video and frontal rendering. *Bottom*: Bottom view driving video and frontal rendering.

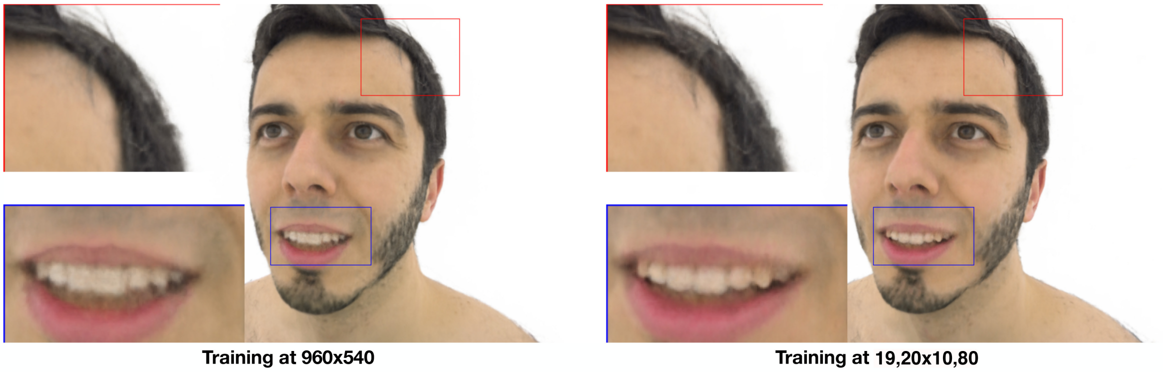


Fig. 9. Avatar synthesis at different resolutions. *Left to right*: Model trained at HD and FHD resolutions, respectively.

run on a resolution of 480 × 270 in real time (25 fps) while still maintaining high fidelity in the reconstructions. Note that the reported runtimes are based on a single NVIDIA A100 GPU. Please see the supplemental video for more results.

Driving using a parametric head model. In our experiments, we have driven avatars using RGB image inputs. Nonetheless, our image encoder can be fine-tuned to accommodate the expression and pose parameters of a parametric head model or 3DMMs. By

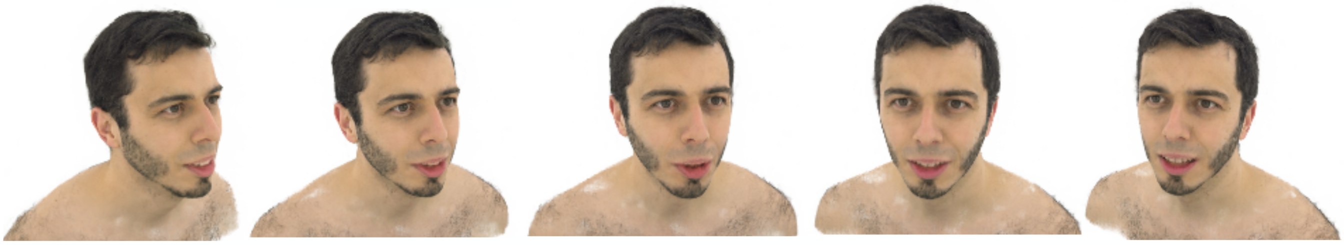


Fig. 10. Real-time rendering (25 fps) at 480 × 270 resolution.

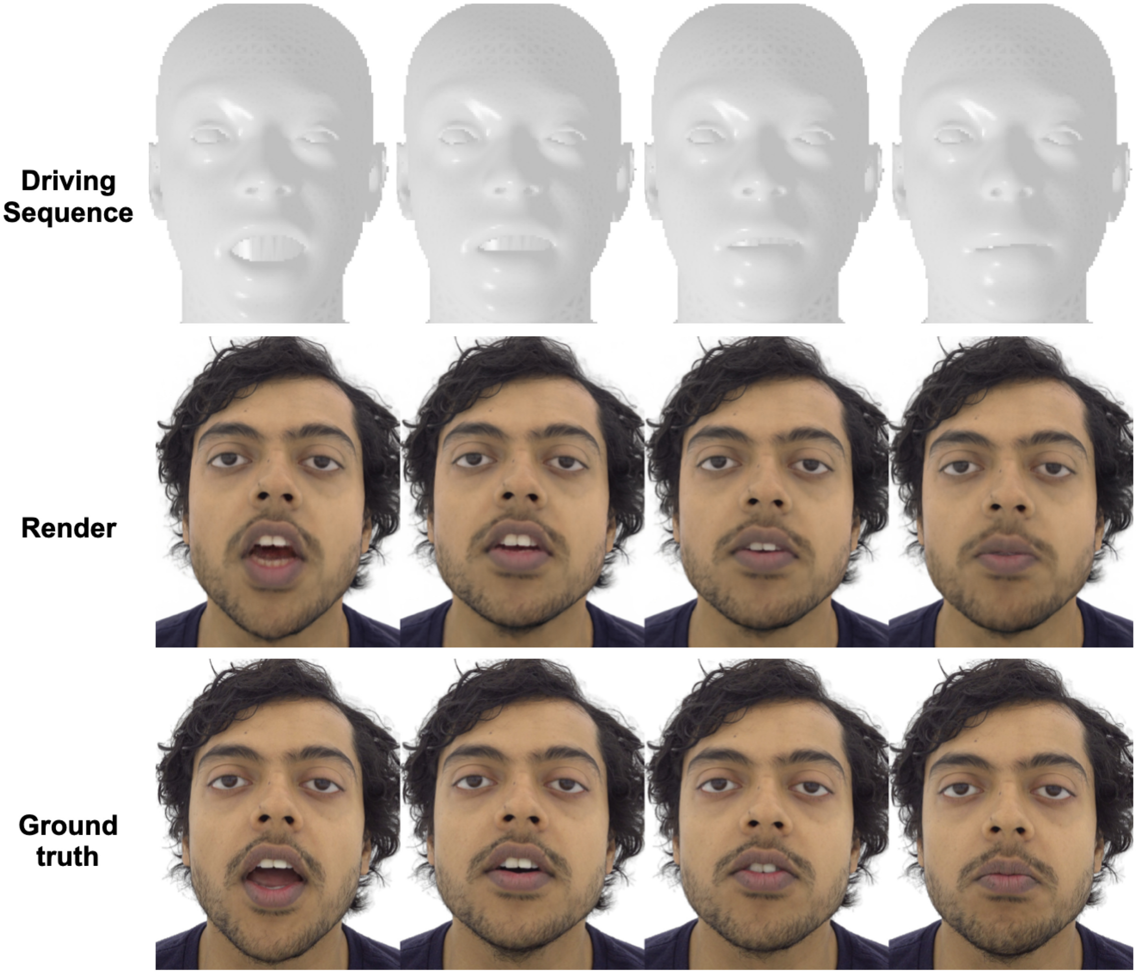


Fig. 11. Avatar synthesis using the rendering of a parametric 3D face model. *Top:* Input sequence of mesh renderings. *Middle*: Output renderings of our personalized head model. *Bottom*: Ground truth.

rasterizing the tracked mesh for a given pose under fixed illumination in screen space, we can drive our personalized head model using rendered mesh images, termed as I_{MD}. To fine-tune our model for I_{MD}, we initially train our personalized head model following the procedure in Section 3. Thereafter, we employ I_{MD} as the driving images for the same training frames. We train our model by freezing all the components except the image encoder. This fine-tuning is done with I_{MD} as driving images for 100k iterations. In adopting this fine-tuning approach, renderings of the parametric model can drive our personalized head model at test time, as

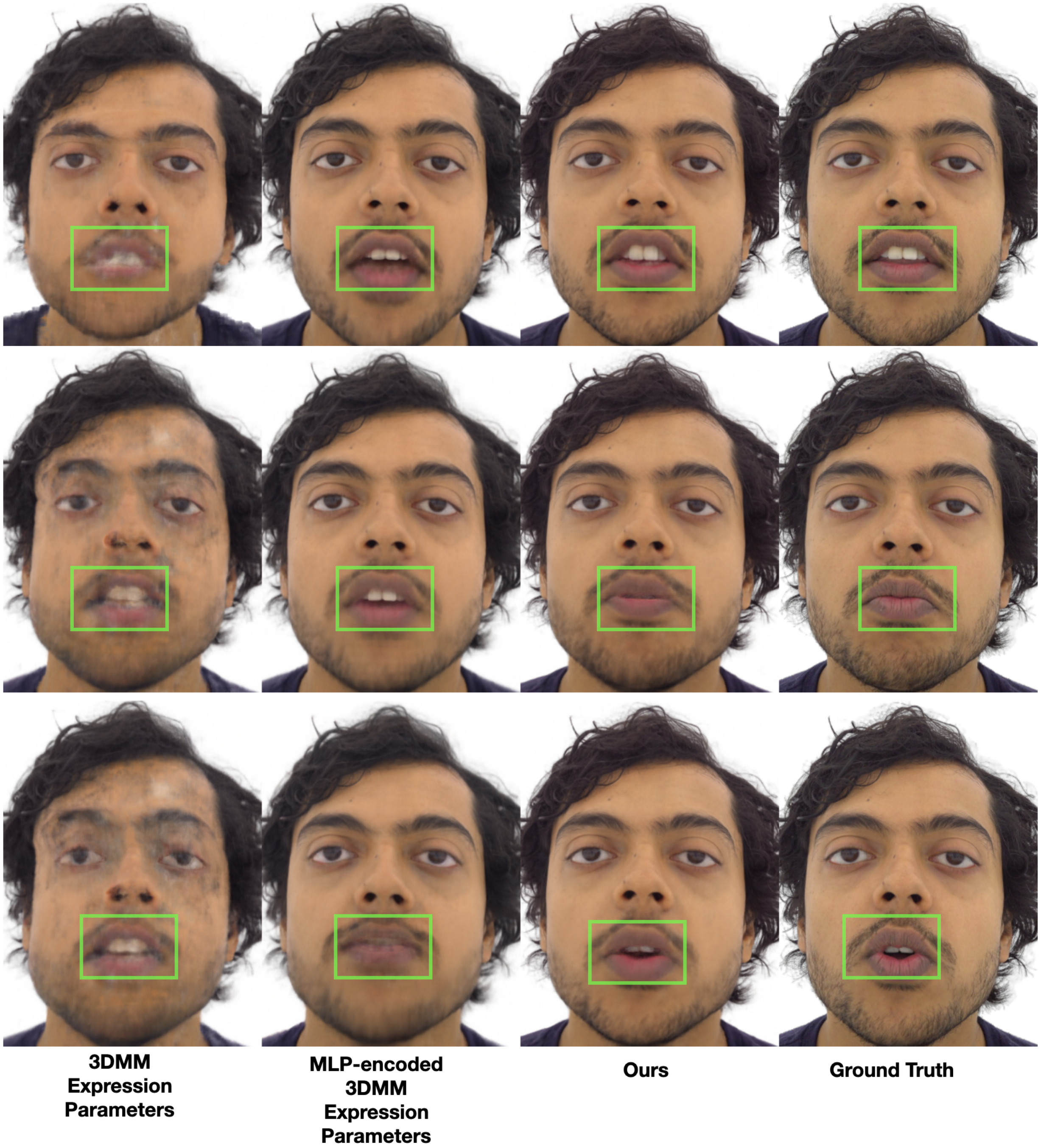

3DMM Expression Parameters
MLP-encoded 3DMM Expression Parameters
Ours
Ground Truth

Fig. 12. Qualitative comparison of different input representations. *Left to right*: driving the trained subject using the 3DMM expression parameters directly, with further MLP encoding, ours, and the ground truth. The highlighted areas in *green* show the level of details in the mouth region.

shown in Figure 11. Such a driving approach facilitates our model's use beyond our training setup, as illustrated in our supplementary video. For this application, we use MICA [Zielonka et al. 2022] to estimate the head-pose and expression parameters. Figure 12 shows a qualitative comparison of driving the trained subject directly with expression parameters, MLP-encoded expression parameters, and our approach. Training our model directly using the MICA [Zielonka et al. 2022] expression parameters as encoding

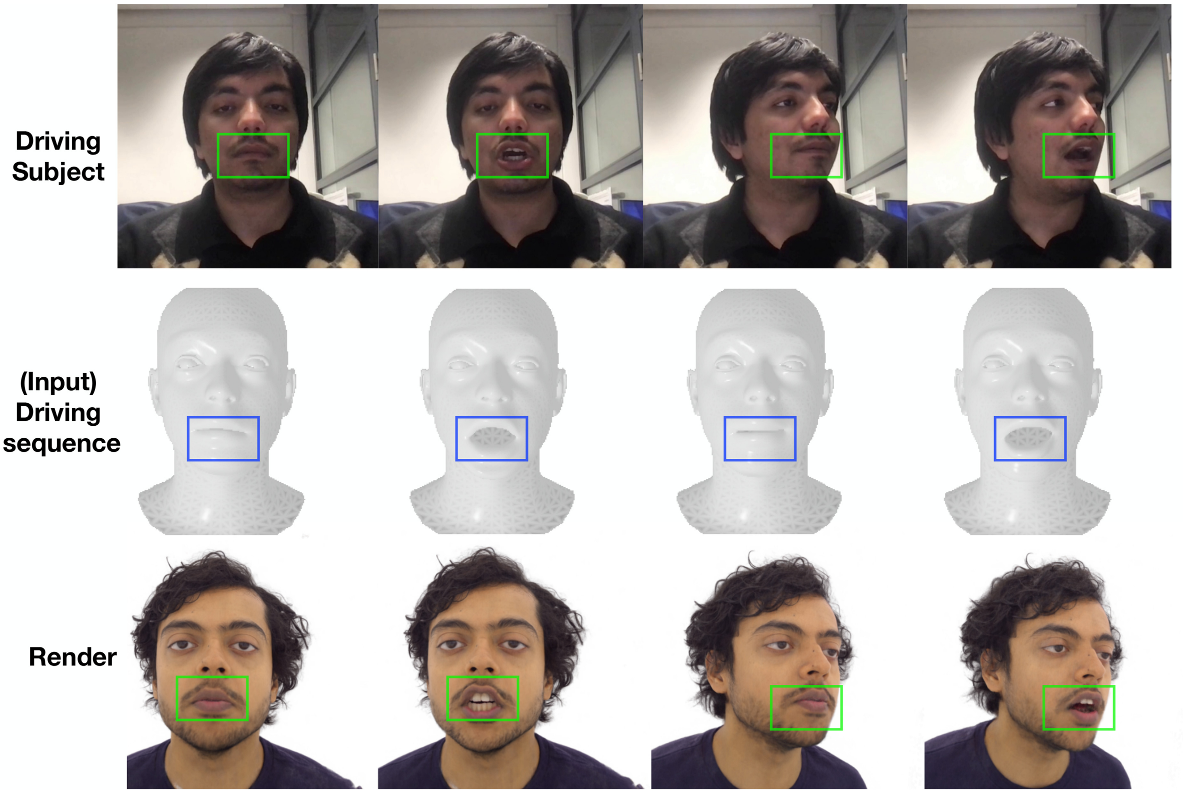


Fig. 13. Cross-identity expression transfer results. We transfer the facial expressions of the driving subject (top row) to the trained subject (bottom row) using rasterized 3DMMs (middle row). The regions highlighted in *blue* show the mouth region expression tracking of the driving subject. The regions highlighted in *green* demonstrate the expression alignment between the driving subject and our model's rendered output.

Table 3. Comparison of Runtimes for Various Rendering Components

Component	E_γ	A_θ	A_α	A_β	ray-marching
Time	0.0125	0.0087	0.01386	0.0087	0.0828

The table lists five distinct components and their associated computational times (in seconds).

Table 4. FPS vs. Quality Comparison for Different Training Design Choices

Configuration	PSNR ↑	SSIM ↑	FPS ↑
Ours	32.72	86.60	8.17
$N_{\text{views}} = 12$	32.90	86.46	7.48
$N_{\text{views}} = 8$	31.97	84.67	7.11
$A_{\theta(5)}$	32.25	84.18	6.96
$A_{\theta(9)}$	31.87	82.91	6.17
$N_{\text{train}} = 350$	32.05	83.07	8.05
$N_{\text{train}} = 750$	32.32	84.73	8.14
$N_{\text{steps}} = 512$	32.64	85.86	14.17
$N_{\text{steps}} = 768$	32.72	86.46	10.79

We show the impact on rendering speed and quality for different settings: Number of training views N_{views}, deformation network size $A_{\theta(.)}$, number of training frames N_{frames}, and maximum number of ray-marching steps N_{steps}.

fails to produce coherent renderings of the head. This is because our model requires a translation of the input via an encoder to produce coherent renderings, as demonstrated in the second and third columns of Figure 12. Here, the encoded versions of the input perform better than naively passing the expression parameters directly as encoding. We also observe that passing the expression parameters through an MLP-based encoder results in clear artifacts in the mouth region. For the MLP-based encoder baseline, we employ a 5 layer deep neural network, each with 128 neurons. Overall, our approach of encoding the 3DMM-rasterized images via our Image Encoder produces the best results. For video results of this qualitative evaluation, please refer to the supplementary video. The 3DMM-based driving application also enables cross-identity expression transfer as shown in Figure 13. This is specifically the case for some key expressions. However, we observe that the expressions can be transferred incorrectly if the 3DMM-based tracking fails to track the driving subject's expressions accurately as shown in Figure 14. For video results of cross-identity expression transfer application, please refer to the supplemental video.

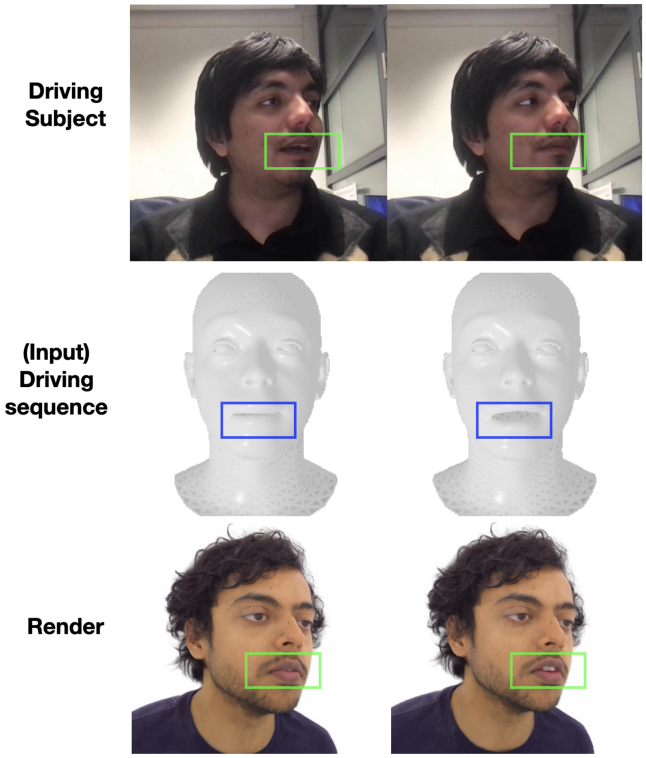


Fig. 14. Expression misalignments in the cross-identity performance transfer. Incorrect 3DMM expression tracking (highlighted in *blue*) of the driving subjects results in the expression mismatch between the driving subject and the training subject (highlighted in *green*).

Table 5. Quantitative Comparison with State-of-the-art Approaches

Metrics	HyperNeRF++	MVP	NeRFBlendshape++	Ours
PSNR ↑	26.42	28.72	29.66	**31.23**
L1 ↓	5.61	3.64	3.23	**2.79**
SSIM ↑	0.8509	0.8283	0.8745	**0.8837**
LPIPS ↓	0.1721	0.1432	0.1326	**0.1130**

L1 measures the absolute error of unnormalized RGB images. Our approach outperforms related methods (see bold text).

4.4 Ablative Analysis

We demonstrate our main contributions and the influence of design choices via a number of ablation studies. Specifically, we study our novel optical flow-based loss, learned image-based feature conditioning of the canonical radiance field network, and canonical space representation. We also analyze the influence of perceptual loss and error map-based pixel sampling in the reconstruction quality. Note that for these experiments, we train our personalized avatar models on 18 views, while we keep out 2 views for our quantitative evaluations.

Figure 15 shows the reconstruction quality of our method and different modeling choices for a fixed unseen expression and a novel camera viewpoint rendering (a held-out view). Here, the error map (bottom row) represents a pixel-wise **mean square error (MSE)** of head renderings in RGB color space. Figure 16 further compares our approach with the same design choices, for a fixed expression but under dynamic novel viewpoint synthesis. Note that dynamic viewpoints are interpolated from different camera viewpoints. From these results, we can observe that without conditioning the canonical space on the driving image features the reconstruction has blurry artifacts all over the mouth. Without the optical flow-based loss, blocky artifacts and/or inconsistent fine-scale details appear in sparsely sampled regions, such as hair, eyelids, and teeth. Figure 17 shows that our proposed optical flow-based loss effectively mitigates these artifacts in a static scene reconstruction setting as well. Overall, our optical flow-based formulation achieves better novel-view synthesis in our sparse camera setup. For this particular comparison, we compare our method's reconstruction result on an unseen static frame vs Instant-NGP [Müller et al. 2022] trained on the same static frame.

Figure 18 shows that while our method with and without optical flow starts at similar perceptual error, using optical flow quickly improves in perceptual similarity as iterations increase. Thus, our optical flow formulation effectively acts as a robust regularizer on the learned volume. Figure 19 shows the state of visual quality for both with and without optical flow-based loss settings with iterations. We notice that optical flow-based loss leads to visual improvement faster than its without optical flow-based counterpart. Note that a canonical space representation is required for proper encoding of facial dynamics; otherwise, artifacts emerge. Table 2 confirms that using the canonical space representation results in a lower reconstruction error. Please refer to the supplementary video for RGB rendering results and depth and surface normal visualizations obtained by our approach.

The error heatmap visualization in Figure 15 (bottom row) provides a quantitative measurement of the error distribution, showing that our approach with all design choices achieves the best rendering quality. Table 2 shows the average reconstruction error over the entire test set (200 frames) for different well-established image-based quality metrics. We adopt similar metrics to that of MVP [Lombardi et al. 2021]. We measure the Manhattan distance $L1$ in the RGB color space, PSNR, SSIM [Wang et al. 2004], and LPIPS [Zhang et al. 2018]. Overall, our approach attains the best numerical results. This study confirms that our key modeling choices optimize the rendering quality. We also show in Figure 20 that the perceptual loss and error map-based sampling improve the rendering results. While we have noticed that these components help in improving rendering quality, we do not emphasize them as a contribution.

Table 3 summarizes the computational runtimes (in seconds) of different rendering components. The computational runtime numbers are aggregated from 400 test frames, with each frame evaluated at a holdout viewpoint and a resolution of 960 × 540. In Table 4, we analyze various rendering configurations to assess their impact on image quality and processing speed for a free speaking sequence at a resolution of 960 × 540. The rendering configurations include: Number of views N_{views}, number of layers in the deformation network $A_{\theta(.)}$, number of training frames N_{frames}, and maximum number of ray-marching steps N_{steps}. For the presented image and speed metrics, a higher value indicates better performance. Note that FPS indicates rendering speed

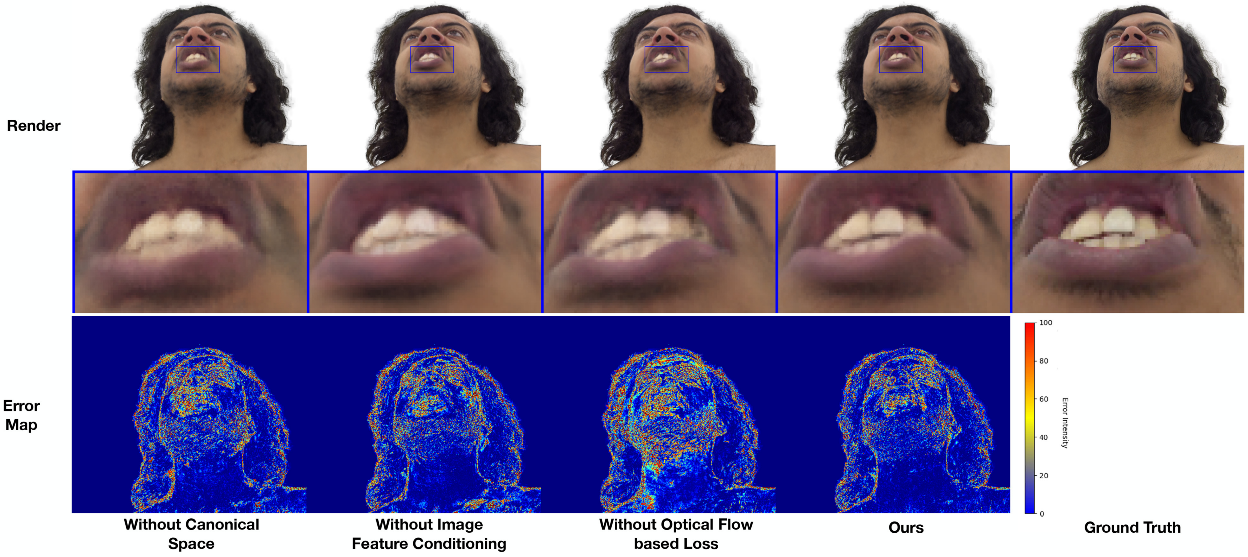


Fig. 15. Ablation study: Fixed view image synthesis for different design choices. *Left to right*: Without canonical space, without feature conditioning, without optical flow-based loss, and ours. The top row shows a rendering of Subject 3 (and ground truth), while the bottom row shows the error map. The error is computed as the per-pixel mean squared error (MSE), encoded in RGB color space. Here, blue denotes 0 MSE, yellow is 60 MSE, and reddish colors mean over 100 MSE. Our full method achieves the best results.

as frames per second. We evaluate on 2 holdout views for each experiment. The configuration labeled "Ours" represents our default setting, with a 5-layer-deep deformation network, A_θ, 1,500 frames for training, 1,024 maximum ray-marching steps, and 16 views for training.

4.5 Comparisons with the State-of-the-art

In this section, we compare our approach with a recent multi-view state-of-the-art method, called MVP [Lombardi et al. 2021], which produces detailed avatars with high fidelity under a similar setup to ours. We disregard direct comparisons with state-of-the-art sparse multi-view approaches, since they tend to lack fine-scale details or are prone to artifacts for novel viewpoint synthesis (see Section 2). In addition, we provide baseline comparisons with an adaptation of a template-free dynamic representation, called HyperNeRF [Park et al. 2021b], and a multi-level hash table-based approach for expression encoding, called NeRFBlendShape [Gao et al. 2022]. We will call our multi-view and image-driven adaptation of these approaches HyperNeRF++ and NeRFBlendShape++.

To train NeRFBlendShape++, we pass each entry of the expression latent vector to a learnable multi-level hash table. We linearly combine the output of these hash tables and condition the NeRF network on it. To train HyperNeRF++, we feed the neural features passed on by the image encoder to an ambient and deformation network and then as appearance conditioning to the NeRF network. To run MVP, we use 4k primitives. We employ an in-house FLAME-based tracking to obtain a non-detailed dense reconstruction of the subject's head to guide the initialization of the primitives at each frame.

Figure 21 shows the reconstruction quality of our method and baseline approaches for a fixed unseen expression and a novel camera viewpoint rendering (a held-out view), while Figure 22 compares them in a free-viewpoint synthesis setup. HyperNeRF++ over-smooths regions. Both NeRFBlendShape++ and HyperNeRF++ exhibit artifacts in regions that undergo recurrent topological changes, e.g., the mouth interior, or that have complex structures, e.g., scalp hair. The latter not only produces stronger artifacts in the form of grid patterns but also removes facial details. Overall, these methods generalize poorly due to over-parameterized representations.

MVP [Lombardi et al. 2021] can sometimes produce wrong facial expressions in extreme cases or even show unusual block artifacts for the same regions mentioned above (see Figure 21 and Figure 22). One of the main reasons is that MVP relies on very dense multi-view imagery to supervise volume rendering. However, in a sparser camera setup, undersampled areas, especially those undergoing disocclusions, become ambiguous without explicit dense volume deformation constraints. The error heatmap visualization of Figure 21 (last row), shows that our method reduces reconstruction errors. Overall, our approach produces sharper, more accurate, and more photorealistic rendering results. Please refer to the supplementary video for further comparisons in dynamic viewpoint synthesis.

We perform quantitative evaluations on the 2 held-out views, with 200 frames each. Quantitative comparisons are reported in Table 5. Our approach clearly outperforms other baseline approaches, especially when comparing perceptual metrics, such as SSIM and LPIPS. L1 reconstruction error is also significantly

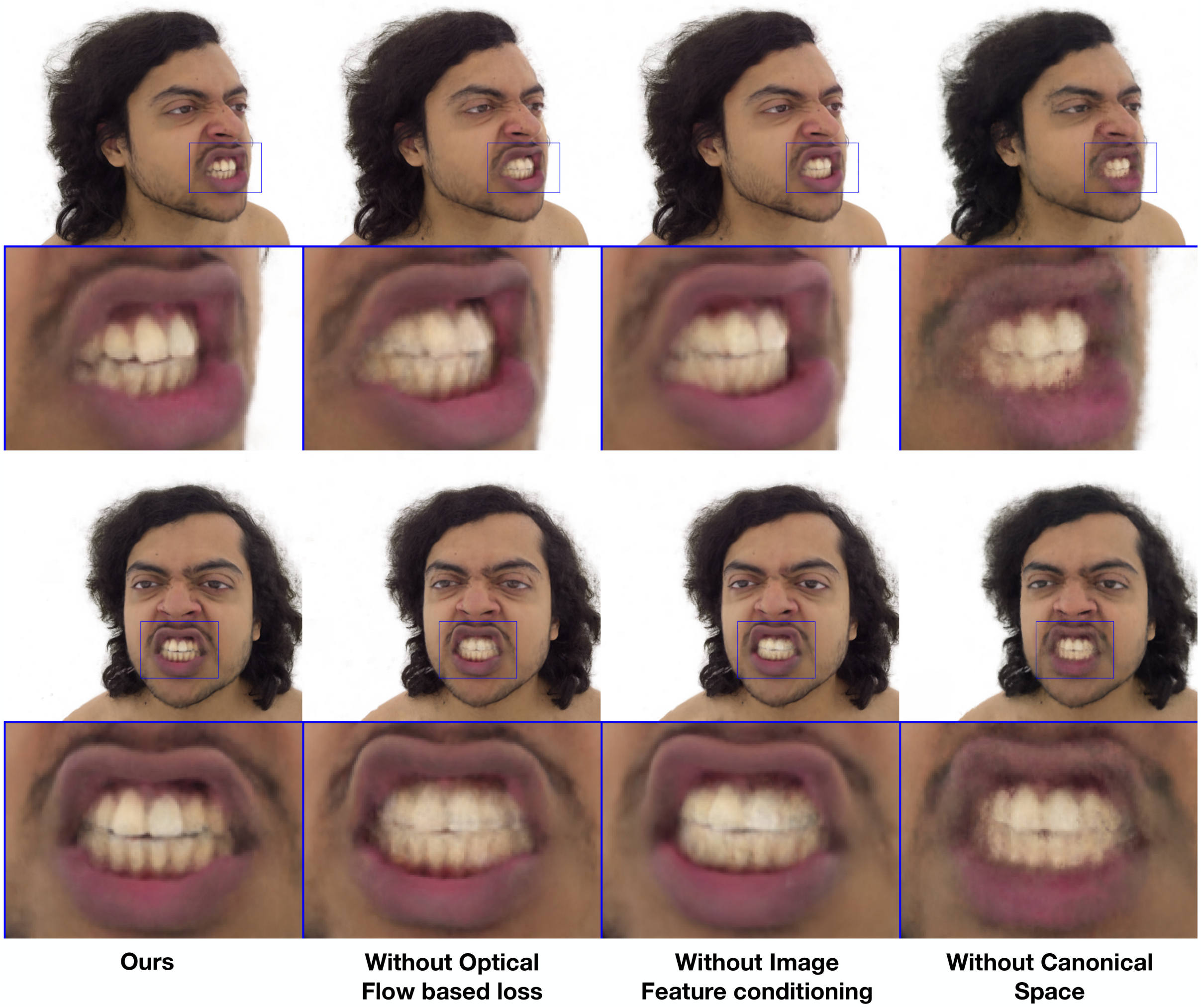

Ours
Without Optical Flow based loss
Without Image Feature conditioning
Without Canonical Space

Fig. 16. Ablation study: Novel view synthesis quality. *Left to right*: Ours, without optical flow-based loss, without image feature conditioning, and without canonical space. Our full method achieves the best results.

reduced. We remark that our approach attains sharper reconstructions with faster convergence and efficiency, the latter thanks to hash-encoding and empty-space pruning techniques.

5 LIMITATIONS AND FUTURE WORK

Our method produces highly photorealistic renderings with novel viewpoints and expressions. However, it suffers from a number of limitations. First, we noticed that it can generate artifacts in motions undergoing strong disocclusions (uncovering occlusions). For instance, in the case of the tongue, artifacts could occur around the mouth boundaries as the tongue starts to stick out (see Figure 23, Frame 1, blue region). The rendering quality, however, stabilizes with good quality as soon as the tongue becomes fully visible (see Figure 23, Frame 2). In the same figure, we also notice that the beard might be blurry. This could be a result of optical flow being unable to track this region to a necessary level of granularity to produce sharp enough results. Future work could address these limitations, e.g., by including occlusion-aware priors and designing a beard-specific synthesis approach. Second, our solution is currently person-specific. Future work could examine building a model that generalizes to unseen identities. For this, our dataset of 16 identities is a good starting point, though it might require more identities. Here, we could also investigate refining the model using in-the-wild data. Third, while we have shown real-time renderings at a resolution of 480 × 270, future avenues could enable real-time rendering at higher resolutions, e.g., FHD synthesis. Here, we could investigate for instance super-resolution techniques, akin to Chan et al. [2022] and Xiang et al. [2022].

Our encoding framework has certain limitations in avatar controllability due to its sensitivity to RGB input, a challenge also observed in Lombardi et al. [2021], Lombardi et al. [2018], Lombardi et al. [2019], and Elgharib et al. [2020]. These methods and ours leverage appearance-based encodings. While it facilitates

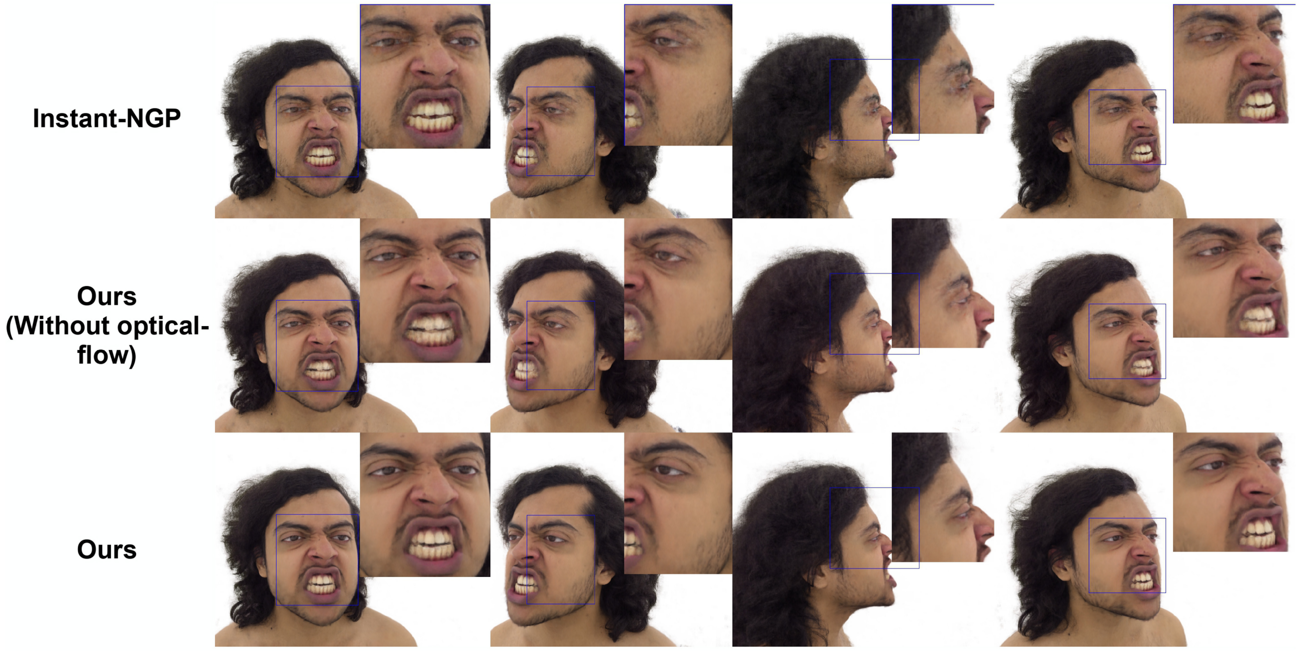


Fig. 17. Comparison of our approach with Instant NGP [Müller et al. 2022] and ours (without optical-flow). We observe that our proposed optical flow-based loss helps remove artifacts for static reconstruction under our sparse camera-setup.

high-quality renderings, it also makes them vulnerable to variations in lighting, subjects, or clothing. The auto-encoder style learning framework of our approach also limits the ability to render extrapolated expressions, as it does not generally perform well for data that is significantly different from what is seen during training [Amodio et al. 2019]. Moreover, our current solution does not account for controllable neck articulation. This particular limitation could be addressed by solutions that integrate neck tracking in the deformation module. Finally, we have mostly shown results driven by monocular RGB videos so far. Theoretically, our image encoder could be replaced with other pre-trained encoders of different input modalities, such as audio signals. This would increase the spectrum of applications of our work.

6 CONCLUSION

We presented a novel approach for building high-quality digital head avatars using multiresolution hash encoding. Our approach models a full head avatar as a deformation of a canonical space conditioned on the input image. Our approach utilizes a novel optical flow-based loss that enforces correspondences in the learnable canonical space. This encourages artifact-free and temporally smooth results. Our technique is trained in a supervised manner using multi-view RGB data and at inference is driven using monocular input. We have shown results rendered with novel camera viewpoints and expressions. We have also shown different applications including driving the model from novel viewpoints. Our approach also shows the first 2K renderings in literature and can run in real-time at a 480 × 270 resolution. Overall, our approach outperforms related methods, both visually and numerically. We will release a novel dataset of 16 identities captured by 24 camera viewpoints and performing a variety of expressions. We hope our work brings human digitization closer to reality so we all can stay in touch with our friends, family, and loved ones, over a distance.

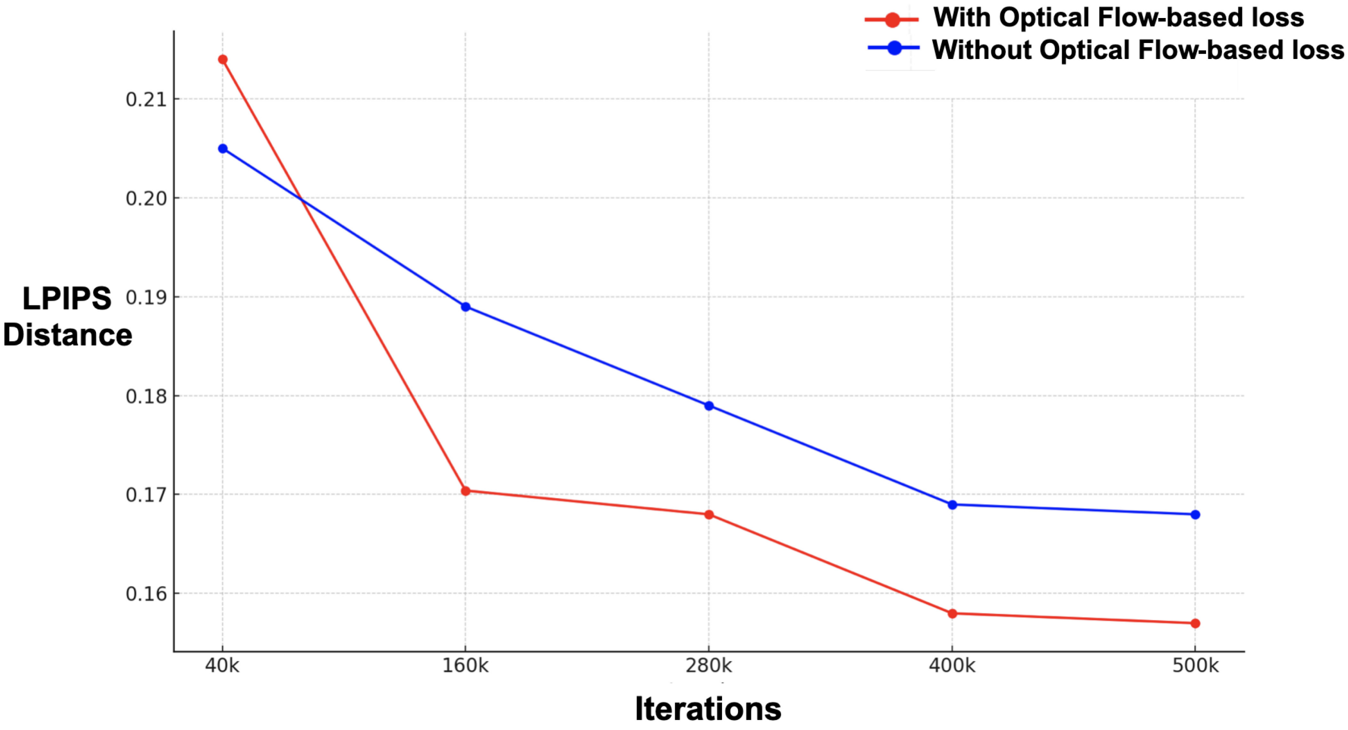


Fig. 18. LPIPS distances across iterations: The red curve depicts training with optical flow-based loss, while the blue shows without. The flow-based approach achieves improved perceptual similarity at earlier iterations.

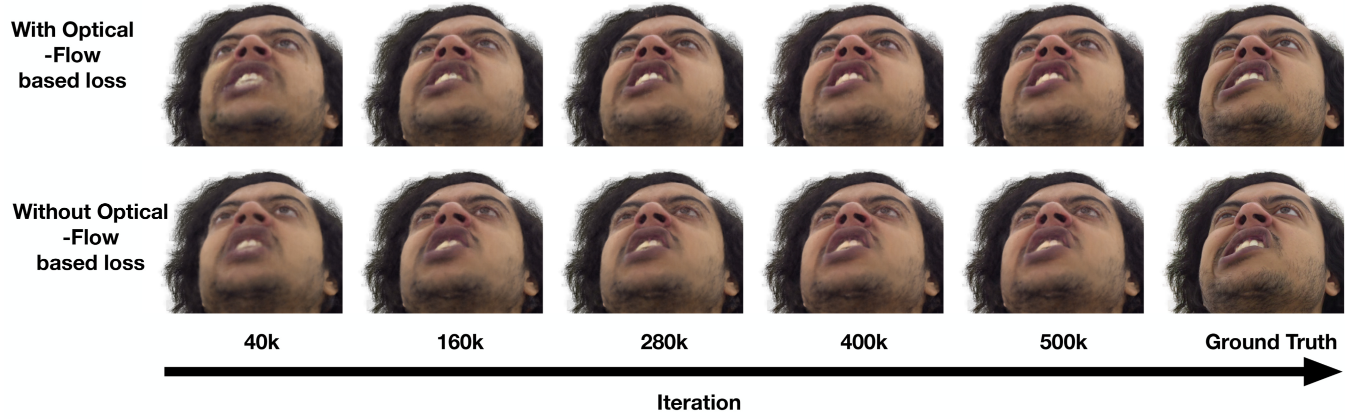

With Optical -Flow based loss
Without Optical -Flow based loss
40k
160k
280k
400k
500k
Ground Truth
Iteration

Fig. 19. Visual comparison of image quality over iterations: Each column represents a distinct iteration, highlighting the evolution of the training process. *Top:* Images trained with optical flow-based loss. *Bottom:* Images trained without optical flow-based loss. As training evolves, differences in detail and structure become more pronounced. Notably, the use of optical flow-based loss results in enhanced details, e.g., teeth, which emerge at earlier stages.

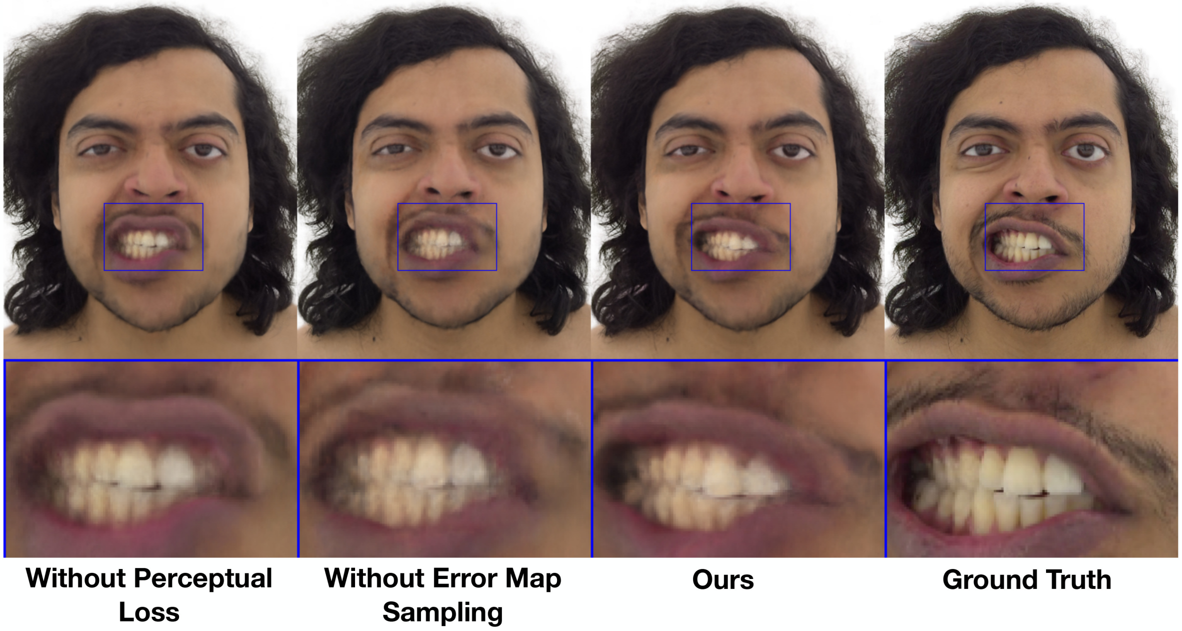


Fig. 20. Ablation study: Structural consistency and detail quality. *Left to right*: No perceptual loss, no error map sampling, ours, and ground truth.

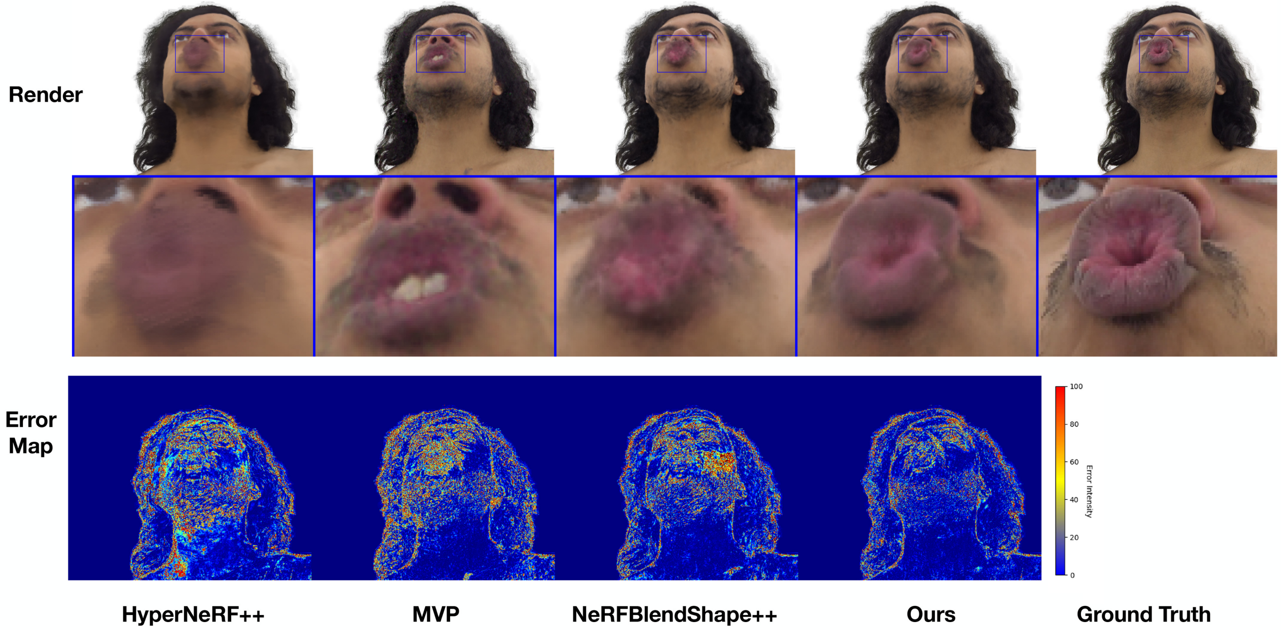


Fig. 21. Quantitative comparison with the state-of-the-art: *Left to right*: Results of HyperNeRF++ [Park et al. 2021b], MVP [Lombardi et al. 2021], NeRFBlendShape++ [Gao et al. 2022], ours, and ground truth. The top row shows visual results, while error maps are shown in the bottom row. The error is computed as the per-pixel mean squared error (MSE), encoded in RGB color space. Here, blue denotes 0 MSE, yellow is 60 MSE, and reddish colors mean over 100 MSE. Our method clearly outperforms the state-of-the-art.

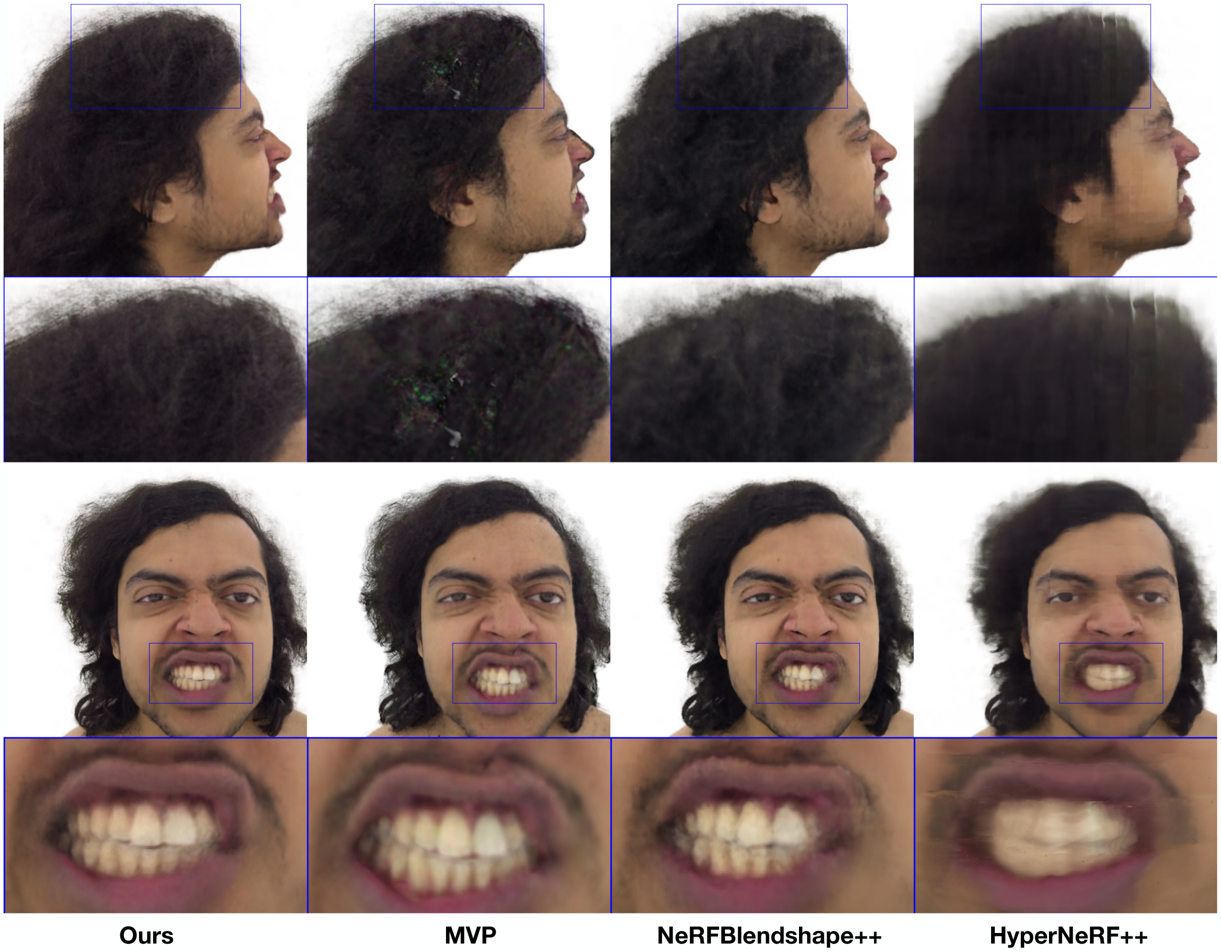


Fig. 22. Qualitative comparisons with the state-of-the-art in a novel view synthesis setting. *Left to right*: Ours, MVP [Lombardi et al. 2021], NerFBlendshape++ [Gao et al. 2022], and HyperNeRF++ [Park et al. 2021b]. Unlike other baseline implementations, our approach produces crisper details and more accurate results.

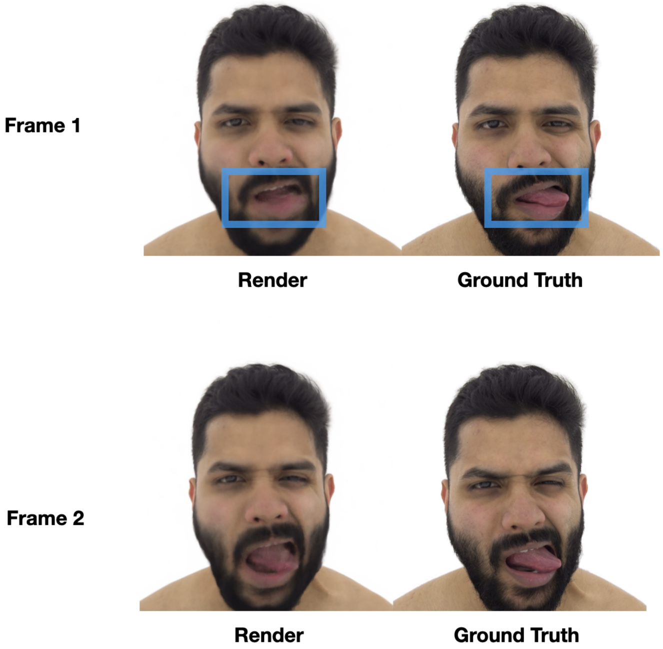


Fig. 23. Our method can struggle to faithfully reconstruct the tongue while transitioning from the mouth interior to the outside of the mouth (Frame 1, see blue region). Once the tongue is out, our method captures the tongue with good quality (Frame 2).

REFERENCES

Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 2023. 3DAvatarGAN: Bridging domains for personalized editable avatars. *CoRR* abs/2301.02700 (2023).

Matthew Amodio, David van Dijk, Ruth Montgomery, Guy Wolf, and Smita Krishnaswamy. 2019. Out-of-sample Extrapolation with Neuron Editing. arXiv:q-bio.QM/1805.12198 (2019).

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully controllable neural 3D portraits. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 20332–20341.

Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, and Ying Shan. 2022. High-fidelity facial avatar reconstruction from monocular video with generative priors. *CoRR* abs/2211.15064 (2022).

Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. In *Conference on Advances in Neural Information Processing Systems.*

Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. 2023. FLARE: Fast learning of animatable and relightable mesh avatars. *ACM Trans. Graph.* 42 (Dec. 2023), 15. DOI: https://doi.org/10.1145/3618401

Mallikarjun B. R., Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. 2021. Learning complete 3D morphable face models from images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 3361–3371.

Chen Cao, Vasu Agrawal, Fernando De La Torre, Lele Chen, Jason M. Saragih, Tomas Simon, and Yaser Sheikh. 2021. Real-time 3D neural facial animation from binocular video. *ACM Trans. Graph.* 40, 4 (2021), 87:1–87:17.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason M. Saragih. 2022. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.* 41, 4 (2022), 163:1–163:19.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.* 20, 3 (2014), 413–425.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.* 35, 4 (2016), 126:1–126:12.

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 16102–16112.

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 *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 5799–5809.

Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus H. Gross, Paulo F. U. Gotardo, and Derek Bradley. 2021. Rendering with style: Combining traditional and neural approaches for high-quality face rendering. *ACM Trans. Graph.* 40, 6 (2021), 223:1–223:14.

Lele Chen, Chen Cao, Fernando De la Torre, Jason M. Saragih, Chenliang Xu, and Yaser Sheikh. 2021. High-fidelity face tracking for AR/VR via deep lighting adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 13059–13069.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. GRAM: Generative radiance manifolds for 3D-aware image generation. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 10663–10673.

Hao-Bin Duan, Miao Wang, Jin-Chuan Shi, Xu-Chuan Chen, and Yan-Pei Cao. 2023. BakedAvatar: Baking neural fields for real-time head avatar synthesis. *ACM Trans. Graph.* 42, 6, Article 225 (Sep. 2023), 14 pages. DOI: https://doi.org/10.1145/3618399

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D morphable face models—Past, present, and future. *ACM Trans. Graph.* 39, 5 (2020), 157:1–157:38.

Mohamed Elgharib, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt. 2020. Egocentric videoconferencing. *ACM Trans. Graph.* 39, 6 (2020), 268:1–268:16.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance fields without neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 5491–5500.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 8649–8658.

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

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 1155–1164.

Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. 2018. Morphable face models—An open framework. In *Conference on Automatic Face & Gesture Recognition.* IEEE Computer Society, 75–82.

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.* OpenReview.net.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A realtime NeRF-based parametric head model. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 20342–20352.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D avatar creation from hand-held video input. *ACM Trans. Graph.* 34, 4 (2015), 45:1–45:14.

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Trans. Graph.* 40, 6 (2021), 210:1–210:12.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Trans. Graph.* 37, 4 (2018), 163.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. 2022. AvatarMe^{++}: Facial shape and BRDF inference with photorealistic rendering-aware GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 12 (2022), 9269–9284.

Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. 2022b. TAVA: Template-free animatable volumetric actors. In *European Conference on Computer Vision.* (Lecture Notes in Computer Science), Vol. 13692. Springer, 419–436.

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), 194:1–194:17.

Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. 2022a. Neural 3D video synthesis from multi-view video. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 5511–5521.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 6498–6508.

Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 5890–5899.

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 8762–8771.

Stephen Lombardi, Jason M. Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Trans. Graph.* 37, 4 (2018), 68.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.* 38, 4 (2019), 65:1–65:14.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.* 40, 4 (2021), 59:1–59:13.

Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 64–73.

Moustafa Meshry, Saksham Suri, Larry S. Davis, and Abhinav Shrivastava. 2021. Learned spatial representations for few-shot talking-head synthesis. In *International Conference on Computer Vision.* IEEE, 13809–13818.

Metashape. 2020. Agisoft Metashape (Version 1.8.4) (Software). (2020). Retrieved from https://www.agisoft.com/downloads/installer/

Marko Mihajlovic, Aayush Bansal, Michael Zollhöfer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European Conference on Computer Vision.* (Lecture Notes in Computer Science), Vol. 13675. Springer, 179–197.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2020), 99–106.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* 41, 4 (2022), 102:1–102:15.

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. paGAN: Real-time avatars using dynamic textures. *ACM Trans. Graph.* 37, 6 (2018), 258.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-resolution 3D-consistent image and geometry generation. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 13493–13503.

Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 165–174.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable neural radiance fields. In *International Conference on Computer Vision.* IEEE, 5845–5854.

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

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *British Machine Vision Conference.*

Amit Raj, Michael Zollhöfer, Tomas Simon, Jason M. Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. Pixel-aligned volumetric avatars. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 11733–11742.

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia Giraldez, Xavier Giró-i-Nieto, and Francesc Moreno-Noguer. 2021. H3D-Net: Few-shot high-fidelity 3D head reconstruction. In *International Conference on Computer Vision.* IEEE, 5600–5609.

Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, Xiaokang Yang, and Stefanos Zafeiriou. 2022. Facial geometric detail recovery via implicit representation. *CoRR* abs/2203.09692 (2022).

Gil Shamai, Ron Slossberg, and Ron Kimmel. 2019. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *CoRR* abs/1901.06551 (2019).

Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. 2022. Controllable 3D face synthesis with conditional generative occupancy fields. *CoRR* abs/2206.08361 (2022).

Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles T. Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. 2021. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 11358–11367.

Jiaxiang Tang. 2022. Torch-ngp: A PyTorch Implementation of Instant-ngp. (2022). Retrieved from https://github.com/ashawkey/torch-ngp

Ayush Tewari, Mohamed Elgharib, Mallikarjun B. R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait image embedding for semantic control. *ACM Trans. Graph.* 39, 6 (2020), 223:1–223:14.

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. 2022. Advances in neural rendering. *Comput. Graph. Forum* 41, 2 (2022), 703–735.

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 2549–2559.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019a. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* 38, 4 (2019), 66:1–66:12.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, 2387–2395.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2019b. Face2Face: Real-time face capture and reenactment of RGB videos. *Commun. ACM* 62, 1 (2019), 96–104.

Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE Computer Society, 1126–1135.

Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo F. U. Gotardo. 2022a. MoRF: Morphable radiance fields for multiview neural head modeling. In *SIGGRAPH.* ACM, 55:1–55:9.

Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. 2023. StyleAvatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH Conference (SIGGRAPH'23)*. Association for Computing Machinery, New York, NY, Article 67, 10 pages. DOI: https://doi.org/10.1145/3588432.3591517

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021b. One-shot free-view neural talking-head synthesis for video conferencing. In *IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE Computer Society, 10039–10049.

Ziyan Wang, Timur M. Bagautdinov, Stephen Lombardi, Tomas Simon, Jason M. Saragih, Jessica K. Hodgins, and Michael Zollhöfer. 2021a. Learning compositional radiance fields of dynamic human heads. In *IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE Computer Society, 5704–5713.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (2004), 600–612.

Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhöfer, Jessica K. Hodgins, and Christoph Lassner. 2022b. HVH: Learning a hybrid neural volumetric representation for dynamic hair performance capture. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6133–6144.

Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. 2022. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds. *CoRR* abs/2206.07255 (2022).

Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023. AvatarMAV: Fast 3D head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH Conference (SIGGRAPH'23)*.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.* 37, 4 (2018), 162.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for real-time rendering of neural radiance fields. In *International Conference on Computer Vision*. IEEE, 5732–5741.

Hao Zhang, Tianyuan Dai, Yu-Wing Tai, and Chi-Keung Tang. 2022. FLNeRF: 3D facial landmarks estimation in neural radiance fields. *CoRR* abs/2211.11202 (2022).

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE Computer Society, 586–595.

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 Trans. Graph.* 43, 1, Article 6 (Nov 2023), 16 pages. DOI: https://doi.org/10.1145/3626316

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M avatar: Implicit morphable head avatars from videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 13535–13545.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023. PointAvatar: Deformable point-based head avatars from videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 21057–21067.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4574–4584.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. *Comput. Graph. Forum* 37, 2 (2018), 523–550.

Received 15 March 2023; revised 9 February 2024; accepted 13 February 2024