HAvatar: High-fidelity Head Avatar via Facial Model Conditioned Neural Radiance Field

XIAOCHEN ZHAO, Tsinghua University, China
LIZHEN WANG, Tsinghua University & NNKosmos Technology, China
JINGXIANG SUN, Tsinghua University, China
HONGWEN ZHANG, Tsinghua University, China
JINLI SUO, Tsinghua University, China
YEBIN LIU, Tsinghua University, China

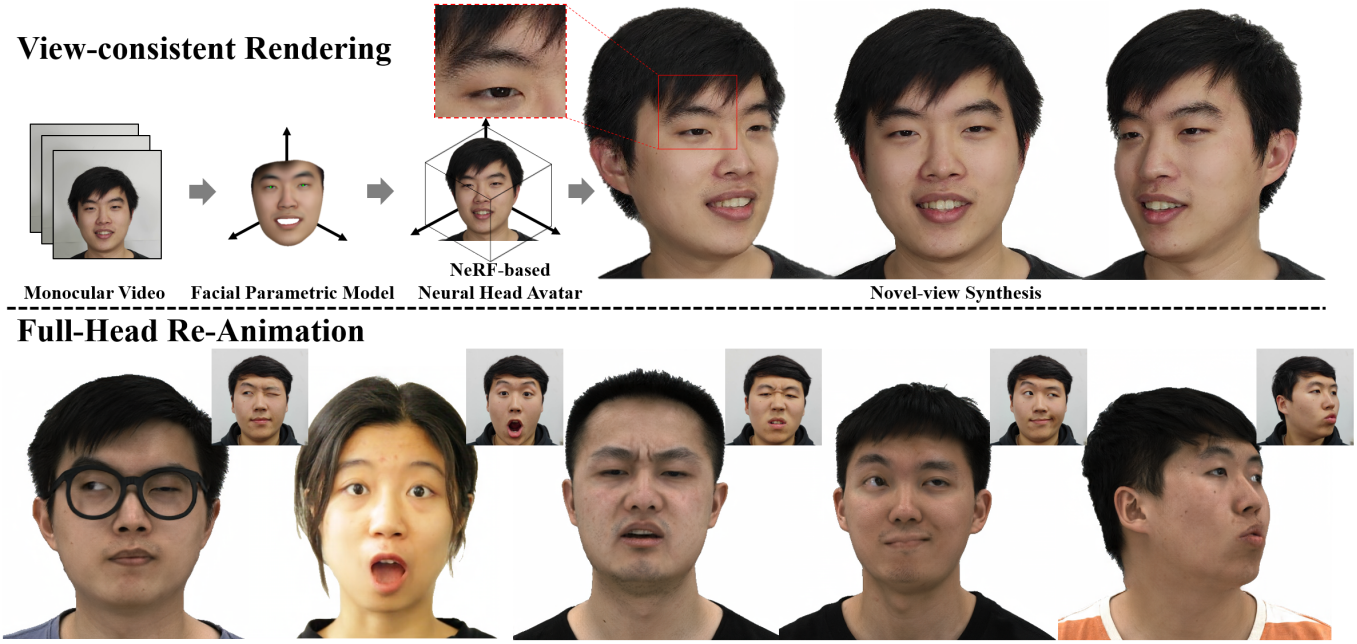

Fig. 1. Our method is able to synthesize high-resolution, photo-realistic and view-consistent head images, achieving fine-grained control over head poses and facial expressions.

The problem of modeling an animatable 3D human head avatar under light-weight setups is of significant importance but has not been well solved. Existing 3D representations either perform well in the realism of portrait images synthesis or the accuracy of expression control, but not both. To address the problem, we introduce a novel hybrid explicit-implicit 3D representation, Facial Model Conditioned Neural Radiance Field, which integrates the expressiveness of NeRF and the prior information from the parametric template. At the core of our representation, a synthetic-renderings-based condition method is proposed to fuse the prior information from the parametric model into the implicit field without constraining its topological flexibility. Besides, based on the hybrid representation, we properly overcome the inconsistent shape issue presented in existing methods and improve the animation stability. Moreover, by adopting an overall GAN-based architecture using an image-to-image translation network, we achieve high-resolution, realistic and view-consistent synthesis of dynamic head appearance. Experiments demonstrate that our method can achieve state-of-the-art performance for 3D head avatar animation compared with previous methods.

CCS Concepts: • **Computing methodologies** → **Image-based rendering**.

Additional Key Words and Phrases: head avatar, image synthesis, parametric facial model, neural radiance field, image-to-image translation

ACM Reference Format:
XIAOCHEN ZHAO, LIZHEN WANG, JINGXIANG SUN, HONGWEN ZHANG, JINLI SUO, and YEBIN LIU. 2018. HAvatar: High-fidelity Head Avatar via

Authors' addresses: XIAOCHEN ZHAO, zhaoxc19@mails.tsinghua.edu.cn, Tsinghua University, Beijing, China; LIZHEN WANG, wang-lz@mail.tsinghua.edu.cn, Tsinghua University & NNKosmos Technology, Beijing & Hangzhou, China; JINGXIANG SUN, starkjxsun@gmail.com, Tsinghua University, Beijing, China; HONGWEN ZHANG, zhanghongwen@mail.tsinghua.edu.cn, Tsinghua University, Beijing, China; JINLI SUO, jlsuo@tsinghua.edu.cn, Tsinghua University, Beijing, China; YEBIN LIU, liuyebin@mail.tsinghua.edu.cn, Tsinghua University, Beijing, China.

0730-0301/2018/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

Facial Model Conditioned Neural Radiance Field. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Animatable 3D human head avatar modeling is of great significance in many applications such as VR/AR games and telepresence. There are two key factors for a lifelike virtual personal character: the accuracy of facial expression control and the realism of portrait images synthesis. Though exiting solutions [Lombardi et al. 2018, 2021; Ma et al. 2021] are able to reconstruct high-quality dynamic human heads, they typically depend on complicated dense-view capture systems and even rely on hundreds of cameras. By leveraging the learning-based techniques, researchers have shifted interest to explore the possibility of automatically modeling human head avatar, with accurate controllability and high-fidelity appearance, under light-weight setup.

Firstly, to establish a controllable personalized head character, the most straightforward way is to directly learn a global parameter conditioned neural head avatar from image sequences, but such method [Gafni et al. 2021] limits the generalization ability in expression control. To improve control robustness, other works [Grassal et al. 2022; Zheng et al. 2022] attempt to leverage parametric templates [Li et al. 2017] to help regulate the avatar modeling during the training stage. However, the explicit surface prior from the parametric model constrains the expressive power for complex-topology parts (i.e. glasses).

Secondly, for high-fidelity human head avatar modeling, recent implicit-surface-based methods [Grassal et al. 2022; Yenamandra et al. 2021; Zheng et al. 2022] recover more texture details compared with conventional methods [Cao et al. 2014; Li et al. 2017; Wang et al. 2022b; Yang et al. 2020] with limited-resolution texture representation. Nevertheless, the quality of the recovered appearance is still far from satisfactory. Built on the expressive neural radiance field (NeRF) [Mildenhall et al. 2020], Nerface [Gafni et al. 2021] is able to generate more promising dynamic appearance results. However, based on the MLP backbone, it is trained in an auto-decoding fashion and tends to overfit training sequences, leading to the obvious inconsistent shape across different frames and unnatural head shaking in the test phase.

Combining the expressiveness of NeRF and the prior information from the parametric template is a promising way for achieving fine-grained expression control and realistic portrait synthesis. Recent work [Athar et al. 2022] establishes a deformable-mesh-guided dynamic NeRF for head avatar modeling. However, the prominent challenge for the coupling of geometry models and NeRF comes from the difficulty in establishing reliable dense correspondences between the real-world subject and the fitted parametric template. Due to the limited expressiveness of the morphable model, it is hard for the deformed mesh to perfectly align with the real-world head with high diversity in terms of geometry and topology. Resulted from the obvious misalignment, the spatial sampling points in the neural radiance field tend to establish ambiguous correspondences with the mesh surface, leading to blurry or unstable rendering performance.

In this paper, we introduce a novel Parametric Model-Conditioned Neural Radiance Field for Human Head Avatar. Inspired by the effective rendering-to-video translation architecture adopted by [Kim et al. 2018], we extend the synthetic-rendering-based condition for 3D head control, by integrating it with triplane-based neural volumetric representation [Chan et al. 2022]. The dynamic head character is conditioned by the axis-aligned feature planes generated by the orthogonal renderings of the textured fitted parametric model in the canonical space. We leverage a powerful convolutional network to learn the reasonable correspondences between the canonical synthetic renderings and the observed head appearance, hence avoiding the ambiguous correspondences determined by the Euclidean distance. On the one hand, such a synthetic-rendering-based condition introduces the prior of the fully-controllable 3D facial model into the neural representation to achieve fine-grained and consistent expression control. On the other hand, orthogonal renderings can supply rough 3D descriptions and avoid excessive restriction from the coarse geometry of the model mesh, so that our head avatar is capable of describing complex topology structure. Considering that the dynamic content mainly comes from facial expressions, we utilize a facial parametric model rather than a full-head model in practice, leaving only the facial region benefiting from the model's prior.

Moreover, while retaining the powerful appearance expressiveness of NeRF [Mildenhall et al. 2020], our method also overcomes the inconsistent shape issue that commonly occurs in NeRF-based modeling methods [Gafni et al. 2021]. Based on our synthetic-renderings-based orthogonal-plane representation, we utilize learnable embeddings to modulate the plane feature generators rather than condition the MLP decoder in an auto-decoding fashion like Nerface [Gafni et al. 2021]. By modulating the convolutional kernels and normalizing the feature generation, the embeddings are able to regulate the whole feature volume to avoid overfitting, leading to the consistent head shape in the animation. Our experiments prove that, with per-frame embeddings modulating on the convolutional generators, the shape consistency and the animation stability of our head avatar are significantly improved.

Finally, our method inherits the advantage of NeRF [Mildenhall et al. 2020], which intrinsically supports differentiable rendering and maintains multiview consistency. Thanks to this strength, we further integrate the NeRF-based volume rendering with the neural rendering, and optimize the whole architecture end-to-end with image observations to recover facial details. Specifically, by leveraging the effective image-to-image translation network commonly used in researches of portrait video synthesis [Chen et al. 2020; Kim et al. 2018; Thies et al. 2019; Xu et al. 2020; Zakharov et al. 2019], we translate the rendered 3D-aware feature map into RGB images. Training the overall network in an adversarial manner, our solution firstly achieves high-resolution and view-consistent photo-realistic synthesis for 3D head avatar.

Given monocular or sparse-views videos, after fitting per frame 3D facial models with an off-the-shelf tracker, our approach is able to learn a high-fidelity and view-consistent personalized 3D head avatar, including hair, accessories and torso, under full control of head poses and facial expressions. Meanwhile, we optimize a linear blend skinning (LBS) weight field as well, that decouples the motions of the head and the torso via a backward warping. During test time, given a single-view driving video, pose and expression parameters

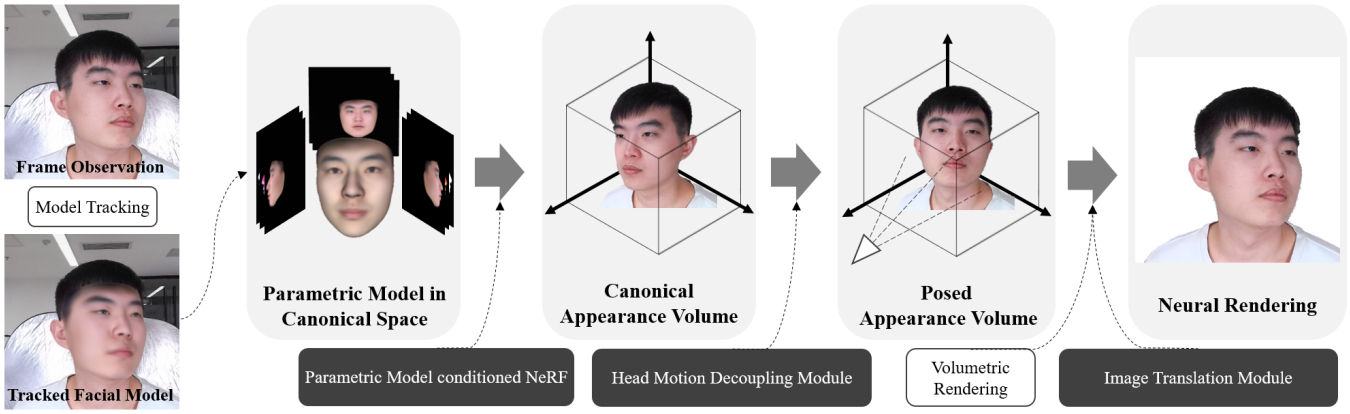

Fig. 2. The overview of our parametric-model-based Neural Head Avatar.

are extracted to deform the facial model, and our method can faithfully recover the entire head appearance under novel expressions, poses, and viewpoints.

In summary, the main contributions of this paper include:

- We propose a novel facial model conditioned NeRF for personalized 3D head avatar, which is built on an orthogonal synthetic-renderings based feature volume. Our representation enables flexible topology and accurate control over the head motion and facial expressions.
- Benefiting from our hybrid representation, we develop a new strategy of modulating generators with conditional embeddings to handle the inconsistent shape issue presented in existing NeRF-based avatar modeling methods and significantly improve the animation stability.
- We firstly achieve high-resolution realistic and view-consistent synthesis of dynamic head appearance, by adopting an overall GAN-based architecture combining our efficient avatar representation with an image-to-image translation module.
- Besides the learning head avatar from monocular videos, we also present head avatar modeling from multiview videos (using 6 cameras), and experiments demonstrate the superior performance of our approach compared with other modified SOTA methods.

2 RELATED WORKS

Our method draws inspirations from explicit parametrical facial model, synthetic-renderings-based 2D facial avatar and implicit 3D head avatar. So we divide this section into three parts.

2.1 Explicit Parametrical Facial Model

Parametric modeling of 3D face has been intensively studied in the past two decades. In the form of explicit meshes, parametric face models are compact, controllable, and easy to be animated. The pioneer work [Blanz and Vetter 1999] builds 3D morphable model to represent facial shape, expression, and appearances. Recently, the parametric face models become more expressive by exploiting more powerful modeling techniques, including multi-linear or nonlinear models [Brunton et al. 2014; Li et al. 2010, 2020; Neumann et al. 2013; Tewari et al. 2018; Tran et al. 2019; Tran and Liu 2018; Vlasic et al. 2006] and the articulated control of expression [Li et al. 2017]. To model detailed deformations of expression, recent state-of-the-art methods [Danecek et al. 2022; Feng et al. 2021a] further learn additional displacement maps with the conditions of image inputs. Moreover, learning-based generative models such as GAN [Karras et al. 2017] or styleGAN [Karras et al. 2020; Karras et al. 2020a] are also used in existing models [Cheng et al. 2019; Gecer et al. 2021; Lattas et al. 2021; Luo et al. 2021; Nagano et al. 2019, 2018; Wang et al. 2022b] to enhance the accuracy of facial texture or geometry modeling. Despite the remarkable progress, all these parametric models can only capture the relatively coarse geometry and appearance of the facial region with the explicit mesh representations, which limits the realism of those reconstruction and animation approaches [cao 2016; Grassal et al. 2022; Hu et al. 2017] built upon them. Instead of solely relying on explicit face models, our approach proposes a controllable hybrid explicit-implicit representation for photo-realistic rendering of 3D face.

2.2 Synthetic-Renderings-based 2D Facial Avatar

To utilize the explicit facial model to represent the entire dynamic human head, some methods [Doukas et al. 2021; Kim et al. 2018; Koujan et al. 2020; Thies et al. 2020, 2019] combine classical rendering and learned image synthesis to establish 2D avatar based on the monocular video. Deep Video Portraits [Kim et al. 2018] presented impressive full head reenactment and photo-realistic image results based on an image2image translation framework. Head2Head [Doukas et al. 2021; Koujan et al. 2020] further improved the temporal coherency with a sequential, video-based rendering network. Instead of using the raw texture of the fitted coarse facial model, Deferred Neural Render [Thies et al. 2020, 2019] extended the idea by rendering the local feature embedded on the mesh surface. Though the rendering-to-video architecture shows an impressive performance in video portrait synthesis, it does not establish 3D representation for the full-head appearance.

2.3 Implicit 3D Head Avatar

In the past three years, it has been an emerging trend to model 3D scenes and objects in an implicit fashion with the success of implicit representations [Mildenhall et al. 2020; Yariv et al. 2020], based on which some works [Chan et al. 2022; Kellnhofer et al. 2021; Mihajlovic et al. 2022; Wang et al. 2022a] have explored to reconstruct high-fidelity view-consistent 3D appearance for static portraits or [Park et al. 2021a,b] model the dynamic scene with head movements. As for animatable personalized head character, many methods [Athar et al. 2022; Gafni et al. 2021; Grassal et al. 2022; Lombardi et al. 2019, 2021; Wang et al. 2021; Zheng et al. 2022] attempted to build implicit representation-based personalized full-head avatar. Based on dense multiview capture systems, some researches [Cao et al. 2021; Lombardi et al. 2019, 2021; Wang et al. 2021] are able to generate facial avatars with impressive subtle details and highly flexible controllability for immersive metric-telepresence. Though the recent work [Cao et al. 2022] supports creating authentic avatars from a phone scan, it relies on a prior model that is pretrained in a large-scale multiview-videos dataset captured in a complicated systems. High cost in data acquisition limited the broad applications. Under light-weight camera settings, based on implicit surface representation, IMAvatar [Zheng et al. 2022] improved generalization to novel expressions by incorporating skinning fields within an implicit morphing-based model, but showed blurry unsatisfying appearance performance. Nerface [Gafni et al. 2021] showed state-of-the-art reenactment results with a parameter-controlled neural radiance field, but struggled to extrapolate to unseen expressions. Recently, RigNeRF [Athar et al. 2022] proposed to maintain a canonical neural radiance field with a backward deformation field guided by parametric model mesh, but suffers from the ambiguous correspondences determined by the Euclidean distance. Besides, for NeRF-based head avatar modeling methods [Gafni et al. 2021; Guo et al. 2021; Hong et al. 2022], there is a tendency to generate frame-wise inconsistent shape. The problem is originated from the unavoidable noise in the estimation of expressions and head poses, thus the similar input expressions may correspond to slightly-different observed appearances, causing unstable canonical shape recovery. Skillfully incorporating the synthetic renderings of parametric model into neural radiance field, our approach achieves both expressive appearance and robust full-head control, and further addresses the inconsistent shape by modulating feature generation with learnable embeddings.

3 OVERVIEW

The overview of our proposed method is illustrated in Fig. 2. Given the monocular or sparse-view videos, we estimate per-frame facial parametric model $\mathbf{M_t}$ from image sequences $\mathbf{I}_t, t = 1, \ldots, T$. Our method conditions the NeRF on the orthogonal synthetic renderings of the model to describe the expression-related head appearance in the canonical space H_C, which supports arbitrary topology and precise expressions control. Besides, per-frame learnable embeddings are utilized to modulate plane feature generation to address expression-shape coupling issue (Sec. 4.1). Based on the learned LBS weight field, the canonical appearance volume H_C is warped into the observed space H using the estimated head pose, resulting in

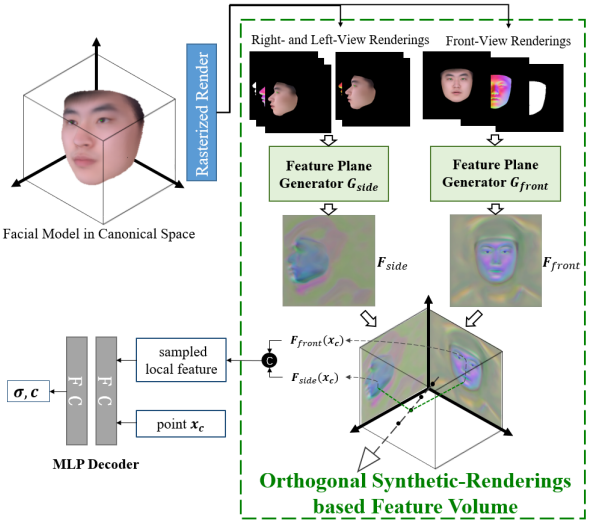

Fig. 3. Parametric model conditioned Volumetric Representation for canonical head appearance.

the decoupled motions of the head and the body (Sec. 4.2). With an image-to-image translation network to transferring the volumetrically rendered 2D feature maps to final RGB images, our method achieves high-resolution, photo-realistic and view-consistent portrait image synthesis (Sec. 4.3). The overall framework is trained in an adversarial manner with image observations and the established head avatar can be applied for training sequence 4D reconstruction or novel full head reenactment (Sec. 4.4).

3.1 Recap: Nerface

Nerface [Gafni et al. 2021] firstly extends NeRF [Mildenhall et al. 2020] to describe expression-related dynamic head appearance. Based on the classical backbone of 8 fully-connected layers, Nerface additionally inputs low dimensional expressions of the morphable model to condition the neural scene representation network for dynamically changing content. By employing the estimated head pose to transform the rays into the canonical space shared by all frames, the head canonical appearance volume H_C can be formulated as:

$$H_C(\mathbf{x_c}, \gamma_\mathbf{t}, \delta_\mathbf{t}) = (\mathbf{c}, \sigma) \tag{1}$$

where the implicit function maps the position in canonical space $\mathbf{x_c}$ to density σ and color feature $\mathbf{c}$, under the control of facial expression parameters δ_t, as well as per-frame embeddings γ_t to compensate for missing tracking information.

Nerface relies on global expression blendshape parameters to represent diverse expression-related appearances. However, by simply learning the mapping from the global conditional vectors to appearances with only a short video sequence, it is easy to be overfit. Hence, though Nerface is good at faithfully reconstructing the training sequences, without the awareness of the underlying 3D structure of human face, it struggles to generalize to unseen expressions.

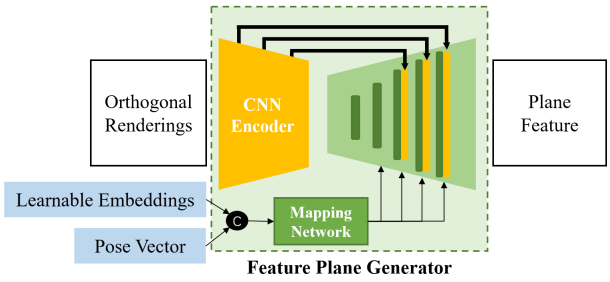

Fig. 4. The architecture of orthogonal plane feature generator network.

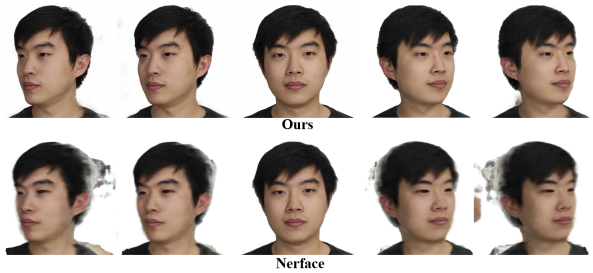

Fig. 5. We show the novel viewpoint synthesis results of monocular-video-based avatars. By introducing 3D prior into NeRF, our approach improves the robustness of image synthesis under large rotations

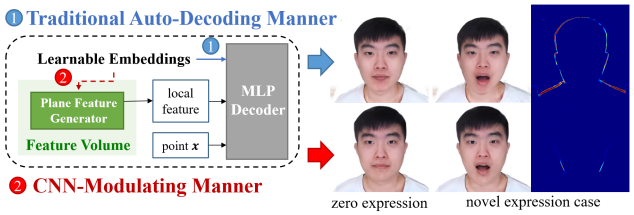

Fig. 6. Comparison of the two different conditional embeddings. Different from conditioning the learnable embeddings in an auto-decoding fashion (marked as ①), we utilize them to modulate the generators of the orthogonal plane features (marked as ②) and prevent embeddings from overfitting to the training dataset. The middle column shows the canonical appearance of the avatar. By only changing the expression (the rightmost column), we illustrate the corresponding rendered appearance and the error map of the generated mask between target cases and the base case.

4 METHOD

4.1 Parametric Model-conditioned NeRF

To introduce the facial structure prior into NeRF [Mildenhall et al. 2020], we propose Parametric Face Model-Conditioned Neural Radiance Field. Our definition of H_C is reformulated as:

$$H_C(\mathbf{x_c}, M_t, \gamma_t, \mathbf{p_t}) = (\mathbf{c}, \sigma) \tag{2}$$

where we utilize the tracked deformed mesh model M_t in zero pose to condition the implicit function, as well as the head pose $\mathbf{p_t}$ to describe the pose-related non-rigid deformation.

4.1.1 Synthetic-Renderings based Feature Volume. Fig. 3 illustrates the architecture of our NeRF-based representation. The head avatar, embedded with a neural network, is conditioned by model-related local features rather than a global vector for better generalization and precision. Specifically, the orthogonal synthetic renderings of the facial model are leveraged to generate the feature volume for the canonical head appearance.

We orthogonally render the 3D facial model in zero pose and integrate the renderings similar to tri-plane-based neural representation [Chan et al. 2022]. Considering the special structure of the human head, we abandon the horizontal plane and utilize the front-view and two side-views planes to characterize the head avatar in the canonical space. Instead of sharing one StyleGAN-based backbone to generate all the feature planes, our method utilizes two separate 2D generators to output feature maps individually [1]. It's also feasible to condition one StyleGAN-based backbone with synthetic renderings to generate all feature planes, but we empirically found that utilizing two separate 2D generators individually contributes to accelerate convergence. As shown in Fig. 4, the synthetic renderings are introduced to the generators to condition the plane feature generation in an explicit manner, for achieving fine-grained controllability. With convolutional encoders extracting image features from the renderings, the extracted multi-resolution features are injected into the generators for spatial-wise feature fusion. In practice, we generate the front-view plane feature F_{front} based on front-view orthogonal renderings and leverage both left- and right-view renderings to get the side-view plane feature F_{side}. Practically, the deformed mesh M_t is rendered as a normal map, a texture map, and a mask map in each view. For the experiments reported in this work, each generator produces a $128 \times 128 \times 64$ feature image.

Based on the generated plane feature images, F_{front} and F_{side}, for any 3D point in the canonical space, we retrieve its feature vectors via orthogonal projection and bilinear interpolation. All the sampled feature vectors, as well as the positional encoding vector of the coordinate, are concatenated into the point feature $\mathbf{f}$ which is fed into an additional lightweight MLP module with two hidden layers of 128 units. Finally, a scalar density σ and a 64-channel color feature $\mathbf{c}$ are predicted for the query point. Indeed, the combination of orthogonal plane features and light-weight MLP makes the burden of scene representation learning fall on the plane feature generation. Hence, we can rely on the powerful and efficient 2D convolutional network, rather than the large MLP backbone, to extract condition information from synthetic renderings and characterize the dynamic head appearance.

As shown in Fig. 5, with the 3D hint from the facial model, our representation improves the quality of view-consistent image synthesis. The usage of both front, left and right elevation is a succinct but efficient description for 3D human head, containing the full observation of the primary part of the head, as well as getting rid of the constrain from the coarse geometry of the mesh model. Setting more planes will lead to information redundancy and unnecessary memory consumption.

[1] Removing texture map is also feasible for person-specific avatar modeling, but we empirically find that adding texture renderings can accelerate convergence.

4.1.2 Conditional Learnable Embeddings. Though our proposed representation is competent for the generation of expression-related canonical head appearance, there is still an unsolved problem: the misalignment between the tracked facial model and the ground-truth observation, which may lead to the frame-wise inconsistent shape. We tackle it by setting additional conditional embeddings for our representation to distinguish similar expressions at different frames.

To account for the misalignment challenges, the previous method [Gafni et al. 2021] also provides per-frame learnable embedding to the neural head avatar, which contribute to better training sequences reconstruction but cannot eliminate the unnatural head shaking while being driven by time-varying expressions in the test phase. This is because it conditions the MLP backbone with embeddings in an auto-decoding fashion [Park et al. 2019], causing the embeddings to overfit the training dataset. Thanks to our representation that conditions the scene with orthogonal synthetic renderings, we condition the plane feature generators with the learnable embeddings, which are fed into a mapping network to modulate the convolutional kernels of the networks, in the manner of StyleGAN2 [Karras et al. 2020b]. The embeddings essentially serve as the normalization of the overall feature and concentrate on maximizing global similarity instead of overfitting per-frame local details during the training. Hence our condition manner contributes to producing a latent space with better interpolation performance and learning a consistent expression-independent head shape. As shown in the Fig. 6, apart from the reasonable expression-related deformation around the cheek, our animation results hardly present shape shaking, proving that our conditional embeddings are able to improve the animation stability.

Specifically, the per-frame embedding is firstly input to a shared mapping network to yield an intermediate latent code which then modulates the convolutional layers of all the separate generators. By constraining the variance of the learnable embeddings, there is a preference to let the generator mainly rely on the synthetic renderings for prediction, and per-frame embedding is utilized to account for the variability resulting from the tracking error.

4.1.3 Pose-Related Non-Rigid Deformation. Though our solution is able to tackle the skeleton motion of head that will be introduced in next section, there still exists pose-related non-rigid deformation caused by head movements in the canonical space, especially in the neck region. In order to describe this, similar to the tackling of per-frame learnable embedding, the estimated head poses are also fed to the mapping network to condition the avatar generation.

4.2 Head Motion Decoupling Module

In this section, we will explain how to handle the rigid skeleton deformation driven by head poses. The straightforward treatment in Nerface [Gafni et al. 2021], that the estimated head poses serves as camera poses, leads to the identical motion of both head and body, which is unrealistic. In order to render images agreeing with the ground-truth observation, the relative movement between the head and torso needs to be considered. As shown in Fig. 7, the canonical appearance volume H_C should be warped to an observed posed

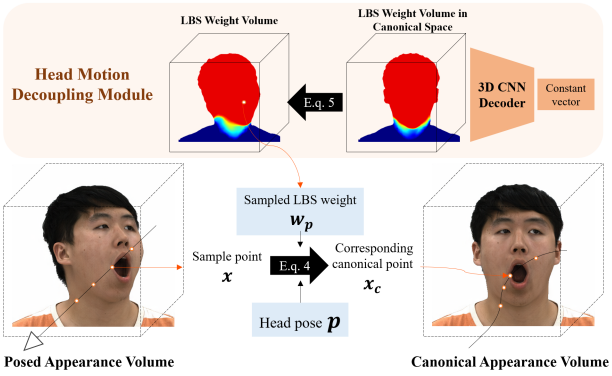

Fig. 7. Decomposition of head movement. The heatmap in LBS weight volume illustrates that the head (red region) moves according to the pose vector p and the torso (blue region) is hardly affected by the head pose.

appearance volume H with the rigid deformation T :

$$H(\mathbf{x}, M_t, \mathbf{p_t}, \gamma_t) = H_C(T(\mathbf{x}, \mathbf{p_t}), M_t, \gamma_t, \mathbf{p_t}) \tag{3}$$

Specifically, we compute the head rigid deformation T as inverse linear blend skinning that maps points from the posed space to the shared canonical space:

$$T(\mathbf{x}, \mathbf{p_t}) = w_p(\mathbf{x})(R_{head}\mathbf{x}+t_{head})+(1-w_p(\mathbf{x}))(R_{torso}\mathbf{x}+t_{torso}) \tag{4}$$

where w_p represents the blend weight, R_{head} and t_{head} the head rotation and translation which comes from estimated head pose $\mathbf{p_t}$, and R_{torso} and t_{torso} means the torso movement which is static by default. In order to avoid overfitting caused by learning backward skinning [Chen et al. 2021; Zheng et al. 2022], following HumanNerf [Weng et al. 2022], we solve for the weight volume in canonical space to derive the w_p as:

$$w_p(\mathbf{x}) = \frac{w_c(R_{head}\mathbf{x} + t_{head})}{w_c(R_{head}\mathbf{x} + t_{head}) + (1 - w_c(\mathbf{x}))} \tag{5}$$

Concretely, we set a 3D convolutional network W_c which inputs a constant random vector and generates the canonical weight volume $w_c(x)$ with limited resolution that can be resampled via trilinear interpolation. With the optimized motion decoupling module, our method can separate out the head movement and stabilize the torso motion.

4.3 Photo-Realistic 3D-Aware Portrait Synthesis

Although the aforementioned hybrid NeRF-based representation is more expressive than available methods, only relying on pixel supervision (MSE/l_1 RGB loss) can hardly yield high-frequency details in the rendered images. Hence, we incorporate the 3D representation into an image2image translation architecture and train the overall network jointly in an adversarial manner to enhance facial details and recover realistic portrait images.

Based on the established appearance volume, volume rendering is implemented using two-pass importance sampling as in [Mildenhall et al. 2020]. In order to remain more 3D-aware information for the subsequent module to generate view-consistent images, similar to previous works [Chan et al. 2022; Gu et al. 2021; Hong et al. 2022;

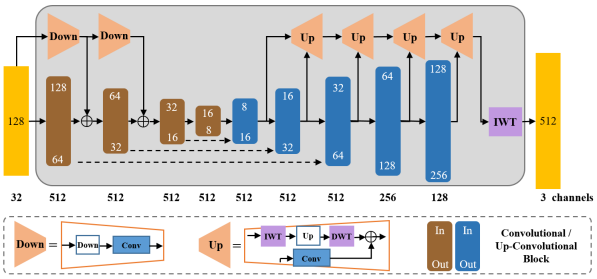

Fig. 8. The architecture of image2image translation network.

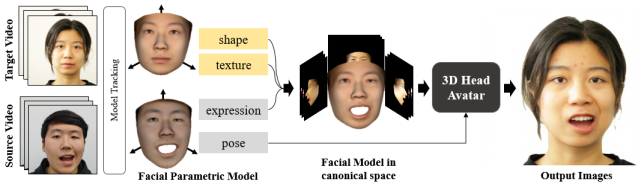

Fig. 9. Re-animation. For an established head avatar, we implement re-animation by transferring the pose and expression parameters from a facial model estimated from a source video to the avatar facial model.

Niemeyer and Geiger 2021], we predict a low-resolution feature map $128 \times 128 \times 64$ from a given camera pose via volumetric rendering, instead of directly rendering an RGB image. However, different from these methods leveraging up-sample super-resolution (SR) module, our approach chooses a UNet-style image2image translation network to transfer the raw feature maps to the final RGB images. The down-sampling encoding process in UNet helps the 2D network learn the global portrait features, which conduces to the view-consistent images generation.

Our architecture is presented in Fig. 8, which includes two main modifications. First, we incorporate skip connections in the decoder, which map each intermediate feature image to an RGB image and integrate the previous output with the next output through addition. Second, we represent the output image as a wavelet (WT) following [Gal et al. 2021], and the RGB image is generated through an inverse wavelet transform (IWT). This design choice helps reduce the number of parameters and speed up network computations.

The joint training of the overall network can guide NeRF module to provide sufficient and appropriate information for the image2image translation module to raise 3D awareness, for regularizing time- and view- inconsistent tendencies. In the next section, we will explain the training procedure and the used loss functions in detail.

4.4 Network Training and Avatar Re-Animation

4.4.1 Training Strategy. Given the tracked facial models of the training sequence and segmented mask images, we employ a two-stage training procedure to optimize the neural head avatar, including the pretraining of the NeRF-based appearance volume and the overall joint training. Firstly, we train only the volume renderer part, the parametric model conditioned NeRF along with the motion decoupling module, to preliminarily establish 3D representation. The objective of the training at the first stage is composed of two components, including an RGB reconstruction loss and a mask loss:

$$\mathcal{L}_{nerf} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} \tag{6}$$

For ease of notation, we drop the subscript (t) of all variables in this subsection.

RGB Reconstruction loss: We additionally set a single linear layer for converting the 64-channel color feature output by the MLP decoder to a 3-channel RGB, and calculate the pixel color via volume rendering [Mildenhall et al. 2020]. The main supervision is $\mathcal{L}_{rgb}$ that measures the mean squared error between the rendered and true pixel colors:

$$\mathcal{L}_{rgb} = \sum_{r\in R} \left\| \hat{C}_r - C(r|M, \mathbf{p}, \gamma) \right\|_2^2 \tag{7}$$

where R is the set of rays in each batch, $\hat{C}_r$ the ground truth pixel color,$C(r|M, \mathbf{p}, \gamma)$ the corresponding reconstructed color determined by parametric model $(M, \mathbf{p})$ and conditional variables (γ) and the network (H) via volume rendering function.

Silhouette Mask loss: Additionally, we utilize the foreground mask that can be easily obtained with BgMatting [Lin et al. 2020] algorithm to provide supervision:

$$\mathcal{L}_{mask} = \sum_{r\in R} BCE(\hat{S}_r, S(r\|\delta, \mathbf{p}, \gamma)) \tag{8}$$

where $BCE(\cdot)$ is the binary cross entropy loss calculated between the rendered silhouette mask value $S(r\|\delta, \mathbf{p}, \gamma)$ and the ground truth mask $\hat{S}_r$.

Next, we train the whole network end-to-end in an adversarial manner with a discriminator [Gal et al. 2021], using the non-saturating GAN loss [Goodfellow et al. 2014] with R1 regularization [Mescheder et al. 2018], denoted $\mathcal{L}_{adv}$. On top of that, the additional loss terms, an l_1-norm reproduction loss $\mathcal{L}_{recon}$ and a perceptual loss $\mathcal{L}_{percep}$, are utilized to penalize the distance between the synthesized image and the ground-truth image.

$$\mathcal{L}_{total} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{percep}\mathcal{L}_{percep} + \lambda_{adv}\mathcal{L}_{adv} \tag{9}$$

4.4.2 Full Head Re-Animation. After network training, the neural head avatar is obtained and can be used to faithfully reconstruct the 4D training sequence and be observed under novel viewpoints. As shown in Fig. 9, facial reenactment can be achieved by transferring expression and pose information from the actor to the avatar. Specifically, given a monocular source video, we only need to extract pose and expression parameters from the estimated parametric model for each frame and combine these parameters with our pre-established avatar-specific facial model to generate the sequence of deformed mesh models serving as the network input. As for the conditional embedding vectors, we use the average of all learned embeddings and fix it during the test phase. Finally, the photo-realistic head appearance, which shares the same identity with the modeled avatar but has the novel poses and expressions from the actor in the source video, is generated.

4.4.3 Implementation Details. We use Adam optimizer to train our networks with the learning rate to 1×10^{-3} for the image-to-image translation module and 5×10^{-4} for all the others. We use 80 samples

(64 from coarse sampling and 16 from fine sampling) per ray. The first stage of training takes about 12 hours and the joint training takes about 36 hours using two NVIDIA 3090 GPUs, while rendering an color image with resolution of 512×512 typically takes 0.15 seconds on one NVIDIA 3090 GPU.

5 EXPERIMENTS

Dataset and Metrics. We separate the evaluation and comparison into two parts: monocular-video-based and multi-view-videos-based experiments. Our monocular dataset contains the public sequence from I M Avatar [Zheng et al. 2022] and a self-made sequence captured with a phone. We collect multi-view sequences with six cameras focusing on the frontal face. All images are cropped and scaled to 512x512. We calculate the foreground masks with BgMatting [Lin et al. 2020] and estimate th per-frame parametric facial model FaceVerse [Wang et al. 2022b] using thir released code. Not that we also track eye gaze and additionally draw the position of pupils on top of the RGB renderings. With each sequence split into training frames and testing ones, we train the networks using the training frames from all viewpoints, and test the animation quality using the testing frames. For quantitative evaluation, we use two standard metrics: peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS).

5.1 Comparisons

We mainly compare our method with the state-of-the-art 3D head avatar modeling methods: Nerface [Gafni et al. 2021], IM Avatar (IMA) [Zheng et al. 2022], RigNeRF [Athar et al. 2022] and Neural Head Avatar (NHA) [Grassal et al. 2022]. For the monocular settings, we also compare with 2D facial reenactment method Head2Head++ (H2H++) [Doukas et al. 2021]. We conduct the comparison on the dataset of [Zheng et al. 2022] and our own data. For IMA, NHA and H2H++, the released data preprocessing codes are utilized to process the monocular videos, and we use our tracked data for Nerface. [2] As the code of RigNeRF is not open-source, we re-implement it and leverage the tracked data of IMA's preprocessing codes to train. To validate the expressiveness of our synthetic-rendering-based NeRF, we also provide a NeRF-baseline(SynR-NeRF) without the image2image translation module.

	case 1		case 2	
Method	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Nerface	26.47	0.221	24.45	0.164
IMA	25.59	0.208	23.90	0.166
NHA	19.54	0.154	18.21	0.158
RigNerf	27.12	0.202	27.92	0.118
H2H++	24.39	0.258	27.12	0.154
Ours(SynR-NeRF)	27.24	0.125	**28.78**	0.109
Ours	**27.58**	**0.070**	28.476	**0.058**

Table 1. Quantitative Evaluation for monocular-videos datasets. Case 1 refers to the top two rows of the Fig. 10 and case 2 refers to the bottom two rows.

	case 1		case 2	
Method	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Nerface-MV	21.77	0.239	19.90	0.247
NHA-MV	16.39	0.238	14.96	0.216
Ours(SynR-NeRF)	22.66	0.122	20.06	0.15
Ours	**23.83**	**0.078**	**21.65**	**0.095**

Table 2. Quantitative Evaluation for sparse-views-videos datasets. Case 1 refers to the top two rows of the Fig. 11 and case 2 refers to the bottom two rows.

Qualitative results are presented in Fig. 10. For IMA and NHA, their texture heavily relies on shape reconstruction and the performance of appearance recovery is inferior. Nerface cannot split the head motion and is more prone to generate unstable results for unseen expressions as it lacks structure prior from the parametric model. H2H++ suffers from unrealistic image artifacts especially when dealing with challenging head poses. As RigNeRF is built on a backward deformation field guided by a coarse 3DMM mesh, for the unconstrained area such as the mouth interior, RigNeRF tends to establish ambiguous correspondences and generates blurry appearance. Compared with the above approaches, our SynR-NeRF baseline is capable of full-head control and accurate reconstruction of the expressions and head poses. Our full pipeline can moreover recover high-frequency details. The quantitative results presented in Tab. 1 further demonstrates the superiority of our method. Note that instead of focusing on the pixel-wise similarity, our full pipeline further improves the strength in detail generation and increase the perceptual similarity, which is proven by the gap of LPIPS scores between our method and the other methods. We also illustrate the comparison on monocular-based animation in Fig. 12. In this experiment, we utilize part of the video from IMA dataset to drive an established head avatar. While dealing with novel expressions and poses obviously different from the training dataset, our approach shows significantly superior performance and robustness.

For the multi-view settings, as far as we know, there is no method focusing on sparse-views-based head avatar modeling available. To this end, similar to our extension to multiview scenario, we extend Nerface-MV and NHA-MV, by adopting multiview parametric face model tracking and optimizing the avatar according to multi-view image evidence. [3] Compared with monocular data, multi-view observations can help model a more complete 3D head avatar, but also cause more obvious misalignment between the estimated mesh models and images due to the limited expressiveness of the parametric model, which raises more challenges for high-quality appearance recovery. Fig. 11 illustrates the qualitative results of two different views, which demonstrates that our method can achieve fine-grained expression control and generate a view-consistent appearance. Nerface tends to produce view-inconsistent artifacts and NHA struggles to describe the topology-varying parts like glasses. The numeric results in Tab 2 show that our method achieves higher

[2] For fair comparisions, we additionally take the position of pupils besides expression parameters as input.

[3] As IMA relies on DECA [Feng et al. 2021b] for tracking, which cannot straightforwardly accommodate to multi-view setting, we do not include it in multi-view experiments.

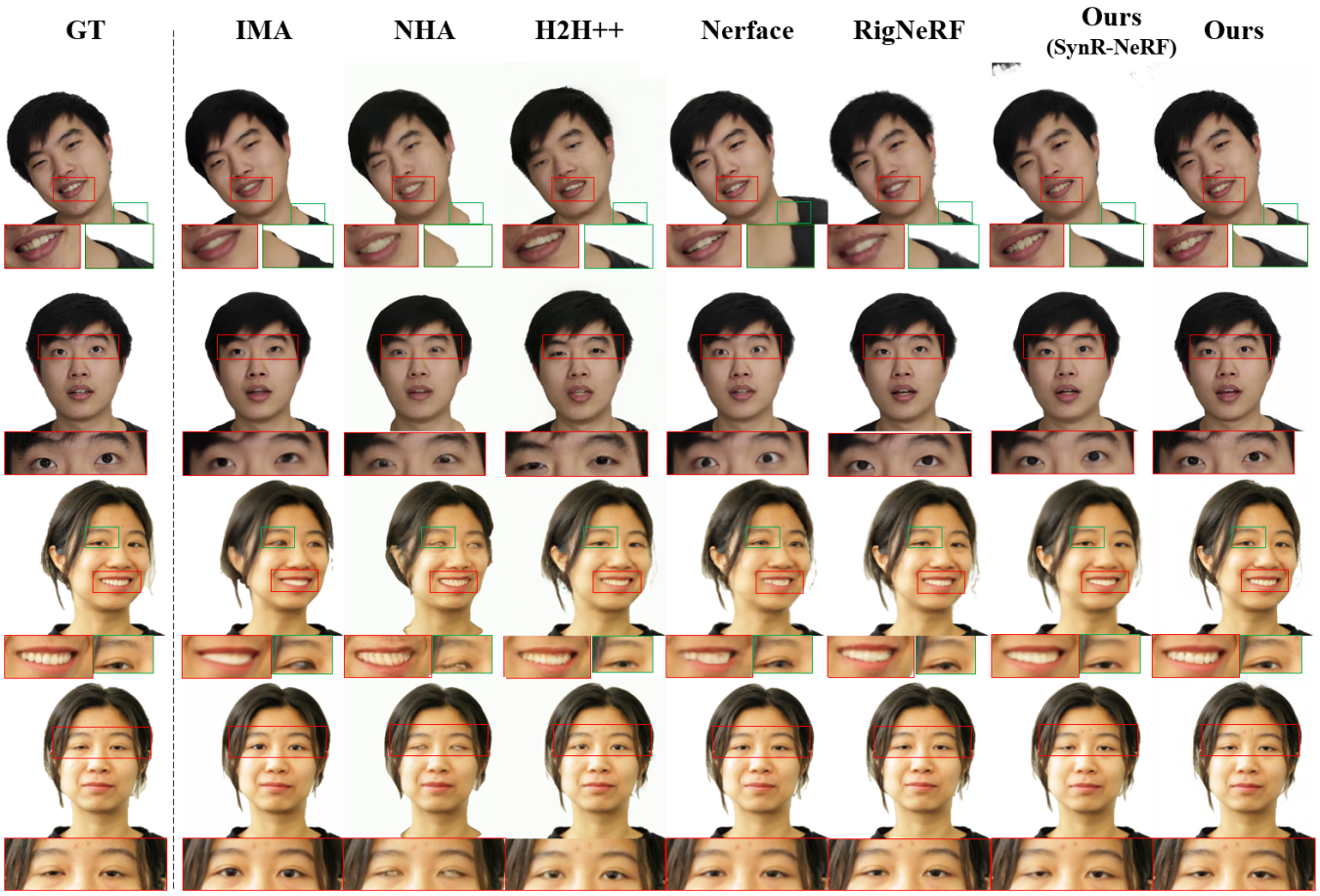

Fig. 10. Comparison with the state-of-the-art methods on monocular video datasets. From left to right: ground truth images, I M Avatar [Zheng et al. 2022], Neural Head Avatar [Grassal et al. 2022], Head2Head++ [Doukas et al. 2021], Nerface: [Gafni et al. 2021], RigNeRF [Athar et al. 2022], our NeRF-baseline and ours. The results demonstrate the superior performance of our method in terms of realistic appearance recovery and fine-grained expression control.

accuracy in both metrics. We present the monocular-based animation results in Fig. 12, which demonstrates our better performance on 3D reenactment.

5.2 Ablation Study

Synthetic-rendering based condition In this part, two modified baselines were implemented for this ablation study. The first one, named 'ExprPlanes-NeRF', replaces our synthetic-rendering-based condition with the implicit vector-based condition used in Nerface, with all other things being equal. The second baseline, named 'ExprMLP-NeRF', further replaces the orthogonal-planes-based neural representation with a deep MLP backbone used in Nerface. To evaluate the effectiveness of our synthetic-rendering-based volumetric representation, we optimized a head avatar on a monocular video dataset using these two variants, and the results are presented in Fig. 13 and Tab. 3. Comparing 'ExprPlanes-NeRF' and 'ExprMLP-NeRF', we found that only applying the orthogonal-planes representation does not significantly improve performance.

	PSNR	SynR-NeRF (Ours)	ExprPlanes NeRF	ExprMLP NeRF
Backbone	MLP			✓
	Orthogonal planes	✓	✓	
Condition	Expression vector		✓	✓
	Synthetic rendering	✓		
	PSNR	26.05	22.93	22.10
	LPIPS	0.1516	0.1683	0.1780

Table 3. Ablation study on our orthogonal synthetic-rendering based volumetric representation.

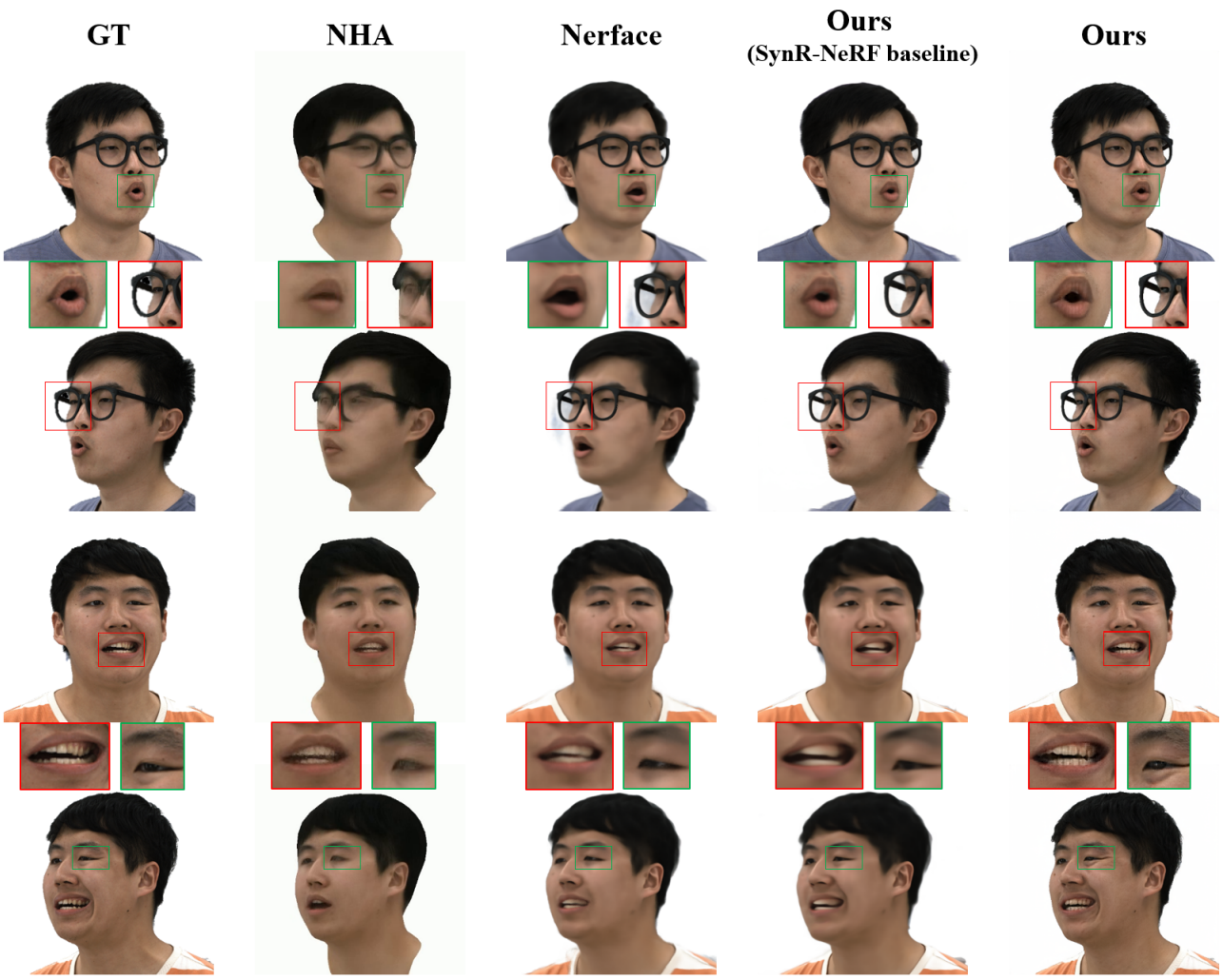

Fig. 11. Comparison with other methods on multi-view video datasets. From left to righ: ground truth images, Neural Head Avatar [Grassal et al. 2022], Nerface: [Gafni et al. 2021], our NeRF-baseline and ours. The results prove our ability in representing topology-varying objects (glasses) and recovering view-consistent high-fidelity appearance.

However, by using synthetic renderings for explicit condition, our method contributes to more accurate expression control.

Image-to-Image Translation Module Results in sec. 5.1 have proven that the image translation module effectively enhances the fine-level details. We implement other two baselines for the ablation study to separately validate the choice of the 2D neural rendering network and the strategy of joint training: 1) We replace the image translation network with the up-sample SR module used by [Chan et al. 2022; Niemeyer and Geiger 2021] and train the whole pipeline end-to-end. However, when attempting to train the network with adversarial loss functions, we empirically find it hard to maintain stable training. We argue that, without the encoder part, the up-sample SR module alone is not suitable for the person-specific dataset with insufficient diversity. Instead, l_1 loss and perceptual loss are adopted in this experiment. 2) We independently train the image translation module with GAN loss to super-resolution the rendered images from a frozen pretrained SynR-NeRF. Experiments are conducted on a multi-view sequence, using 5 views for training and leaving one view for evaluation. As shown in Fig. 14 and Tab. 4, up-sample SR based baseline fails to generate fine details. For separate training baseline, it operates primarily in image-space and introduces undesirable inconsistent artifacts, when dealing with the complex distribution of multi-view images. Through end-to-end training the whole framework, our pipeline contributes to guaranteeing realistic detail generation performance.

Zero-posed Orthogonal Mesh Rendering In our method, we render the 3D facial model orthogonally to create a canonical feature volume in zero pose for feature conditioning. In this part, we introduce a baseline method called 'Posed Rendering,' which involves rendering the pose-dependent meshes to condition the orthogonal

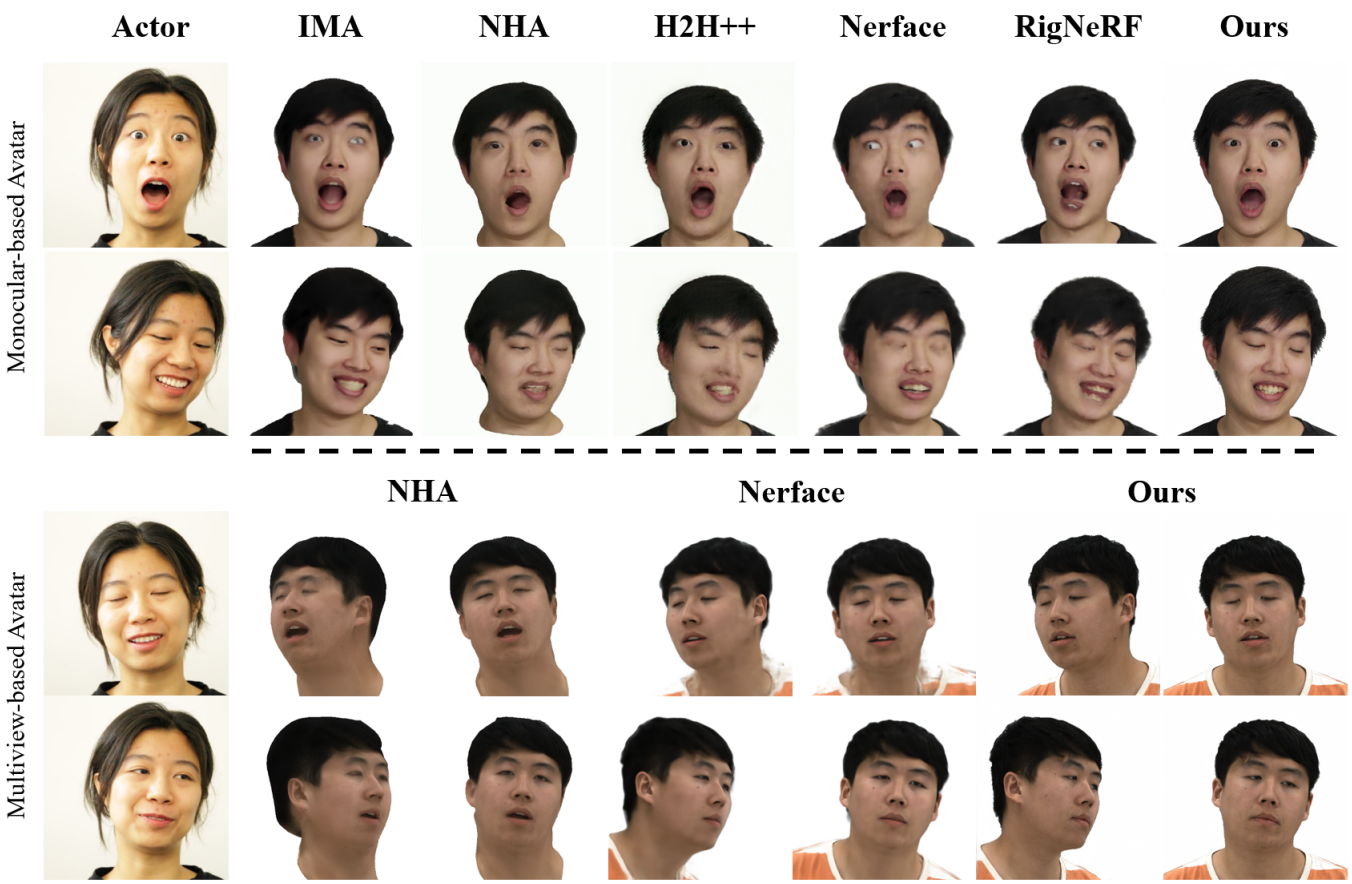

Fig. 12. Comparison with other methods for the task of head animation. From left to right: actor images, I M Avatar [Zheng et al. 2022], Neural Head Avatar [Grassal et al. 2022], Head2Head++ [Doukas et al. 2021], Nerface: [Gafni et al. 2021], RigNeRF [Athar et al. 2022] and ours. Results demonstrate the generalization of our method to novel expressions and poses.

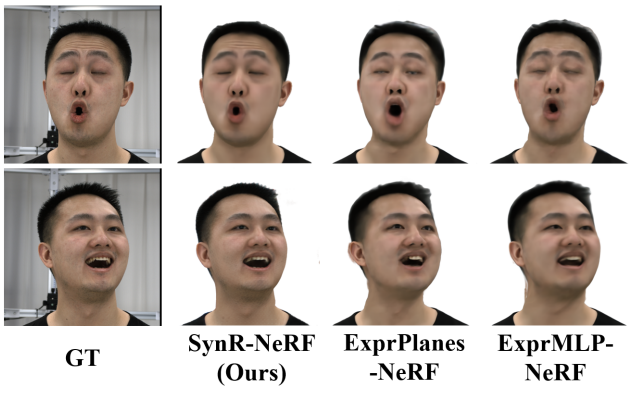

Fig. 13. Ablation study on our orthogonal synthetic-rendering based volumetric representation.

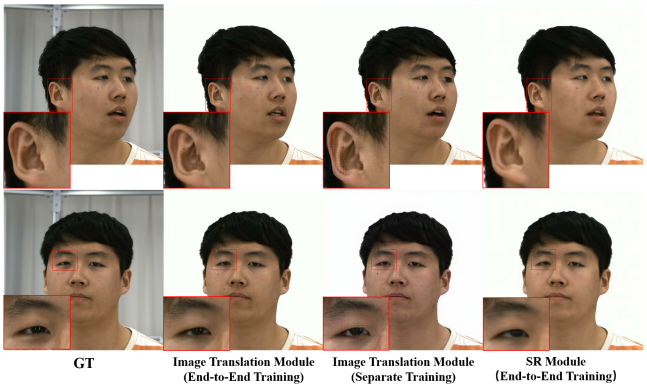

Fig. 14. Ablation study on the Image Translation Module. Using the image translation module instead of up-sample SR module contributes to recovering fine-scale details and the joint training strategy further helps eliminate image-space artifacts. Please zoom in and also refer to our video for more clear comparisons.

feature planes. The results, as shown in Fig. 15 and Tab. 5, indicate

	Ours Image Translation End-to-End	baseline 1 Up-Sample SR End-to-End	baseline 2 Image Translation Separate
FID	7.67	8.62	17.34

Table 4. Ablation study on Image Translation Module.

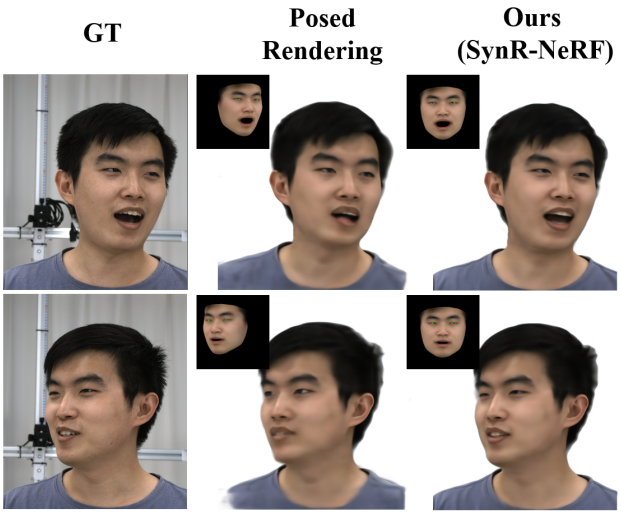

Fig. 15. Ablation study on orthogonal mesh rendering. Besides the generated portrait images, we also show the front-view orthogonal renderings in the top left corner of the picture.

	Posed Rendering	Ours (SynR-NeRF)
PSNR	27.67	29.79
LPIPS	0.1324	0.1219

Table 5. Ablation study on orthogonal mesh rendering.

that 'Posed Rendering' performs worse on the testing set. We attribute this to the coupling of expressions and poses, which creates false correlations between the facial appearance and the face location in the renderings. Plane feature generators have to remember the potential diverse locations in the input image space, leading to reduced performance. In contrast, our method orthogonally renders zero-posed meshes, enabling the generators to concentrate on extracting expression-related information from the renderings and achieve fine-grained control.

Conditional Learnable Embeddings One strength of our method is that we solve the expression-shape coupling issue presented in previous NeRF-based head avatar methods. We owe it to our strategy of modulating plane feature generators, as metioned in Sec. 4.1.2. For evaluation, similar to Fig. 6, we implement a baseline that the NeRF is conditioned in an auto-decoding fashion by inputting the learnable embeddings into the MLP decoder. To fully explore the expression-pose coupling issues, we fix the head pose and only transfer the expression to animate the avatar. The results are presented in the video. Nerface and our modified baseline both show

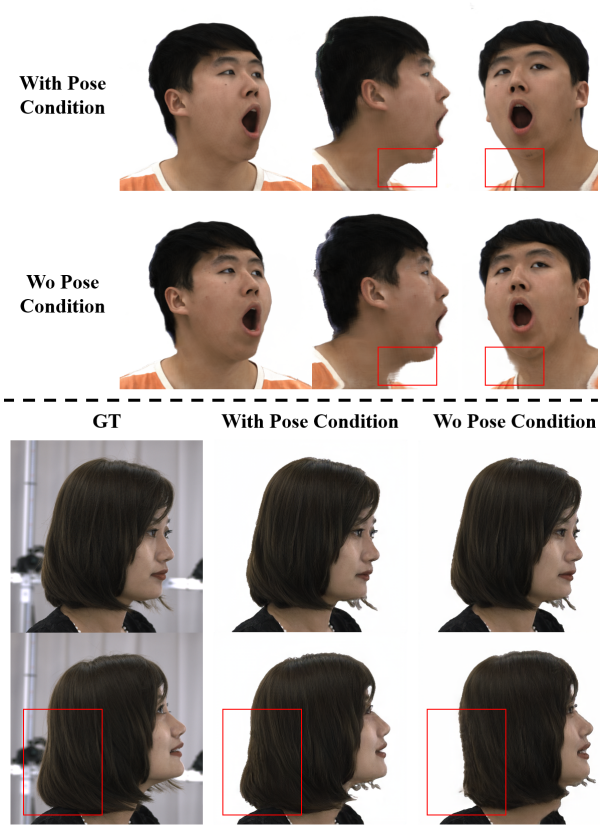

Fig. 16. Ablation study on pose condition.

obvious jitter, while our method illustrates stable animation results and better appearance quality.

Pose Condition As mentioned in Sec. 4.1.3, by introducing pose vectors into condition, our method is able to describe pose-related non-rigid deformation in the canonical space. Fig. 16 illustrates two cases. In the first case, without considering pose, artifacts are visible in the side-view observation when the head turns around. Our method, which includes pose condition, eliminates these artifacts and enhances view-consistency. In the second case, we demonstrate the ability of our method to describe simple pose-related movements of hair.

6 DISCUSSION AND CONCLUSION

Limitation. Although our approach is able to synthesize high-quality 3D-aware portrait images, the proxy shapes produced by our method cannot be competitive with the state-of-the-art alternative approaches [Grassal et al. 2022; Zheng et al. 2022], as shown in Fig. 19 and Fig. 17. While this is not important for the photo-realistic stable view-consistent head image synthesis application we consider in this paper, other applications may benefit from reconstructing more accurate morphable geometry.

Compared to surface-based avatar modeling methods [Zheng et al. 2022], our method struggles with out-of-distribution head poses.

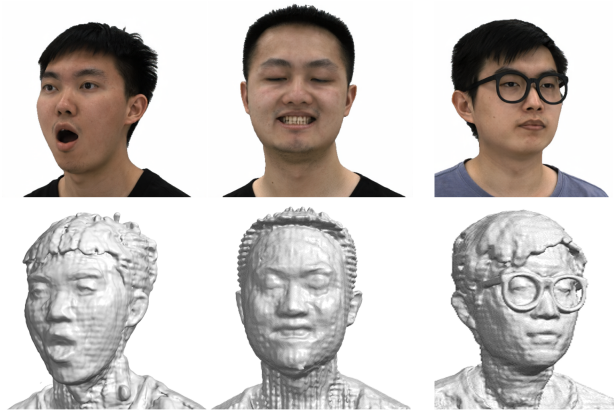

Fig. 17. The proxy shapes of multiview-based avatars produced by our method. We visualize the 3D shape by using the marching cubes algorithm [Lorensen and Cline 1987] on the density output of our implicit radiance filed to produce a surface mesh.

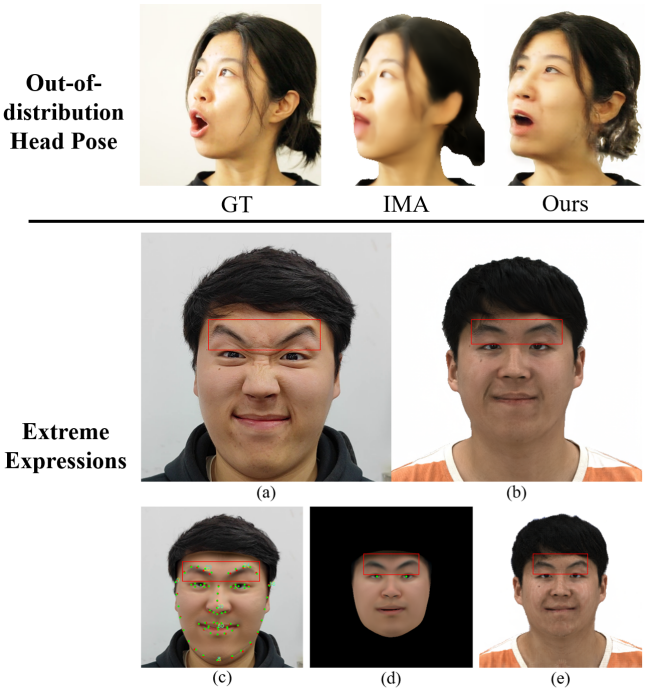

Fig. 18. Failure cases caused by out-of-distribution head poses and extreme expressions that cannot be expressed by the facial model. (a) Actor image. (b) Animation result. (c) Estimated facial model and keypoints (red: detected landmarks; green: projected 3D keypoints of the model). (d) Orthogonal front-view synthetic rendering. (e)Synthesized orthogonal front-view result. Note that the eyebrows of our result in (e) align well with the rendering in (d), but the parametric model cannot fully express the narrow eyebrow of the actor image.

Additionally, because our method relies on a parametric model to control facial expressions, it is challenging to handle extreme expressions that cannot be expressed by the facial model, as depicted in Fig. 18. Furthermore, while our method can capture simple pose-related deformation of long hair, it faces difficulties in dealing with challenging topology-varying cases caused by large hair movements. Special treatment of the hair region is an important problem in the future.

Conclusion. We introduce a novel modeling method that firstly achieves high-resolution, photo-realistic and view-consistent portrait synthesis for controllable human head avatars. By integrating the parametric face model with the neural radiance field, it has expressive representation power for both topology and appearance, as well as the fine-grained control over head poses and facial expressions. Utilizing learnable embeddings to modulate feature generators, our method further stabilizes animation results. Besides monocular-video-based avatar modeling, we also present high-fidelity head avatar based on a sparse-view capture system. Compared to existing methods, the appearance quality and animation stability of our head avatar is significantly improved.

ACKNOWLEDGMENTS

This paper is supported by National Key R&D Program of China (2022YFF0902200), the NSFC project No.62125107 and No.61827805.

REFERENCES

2016. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics* 35, 4 (2016).

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully Controllable Neural 3D Portraits. In *Computer Vision and Pattern Recognition (CVPR)*.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*.

Alan Brunton, Timo Bolkart, and Stefanie Wuhrer. 2014. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*.

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

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *TVCG* (2014).

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.

Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. 2021. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11594–11604.

Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. 2020. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13518–13527.

Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2019. Meshgan: Non-linear 3d morphable models of faces. *arXiv preprint arXiv:1903.10384* (2019).

Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *CVPR*.

Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. 2021. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (2021), 31–43.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021a. Learning an animatable detailed 3D face model from in-the-wild images. *TOG* (2021).

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021b. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 40, 8. https://doi.org/10.1145/3450626.3459936

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

Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. 2021. SWAGAN: A style-based wavelet-driven generative model. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2021. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 4879–4893.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

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

Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. *CVPR*.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 1–14.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

Tero Karras, Samuli Laine, and Timo Aila. 2020. A Style-Based Generator Architecture for Generative Adversarial Networks. *TPAMI* (2020).

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020a. Analyzing and improving the image quality of stylegan. In *CVPR*.

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

Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. 2021. Neural Lumigraph Rendering. In *CVPR*.

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 Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 16–23.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. 2021. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2021), 9269–9284.

Hao Li, Thibaut Weise, and Mark Pauly. 2010. Example-based facial rigging. *TOG* (2010).

R. Li, K. Bladin, Y. Zhao, C. Chinara, and H. Li. 2020. Learning Formation of Physically-Based Face Attributes. (2020).

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *TOG* (2017).

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Real-Time High-Resolution Background Matting. *arXiv* (2020), arXiv–2012.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)* 37, 4 (2018), 1–13.

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

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

William E Lorensen and Harvey E Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169.

Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. 2021. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11662–11672.

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 64–73.

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

Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. 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. In *European conference on computer vision*. Springer, 405–421.

Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. 2019. Deep face normalization. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.

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. *TOG* 37, 6 (2018), 1–12.

Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. 2013. Sparse localized deformation components. *TOG* (2013).

Michael Niemeyer and Andreas Geiger. 2021. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.

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. *arXiv preprint arXiv:2106.13228* (2021).

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv* (2020).

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Trans. Graph.* (2021).

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 *CVPR*.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*. Springer, 716–731.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*.

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3D Face Morphable Model. In *CVPR*.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In *SIGGRAPH*.

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

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022b. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20333–20342.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5704–5713.

Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16210–16220.

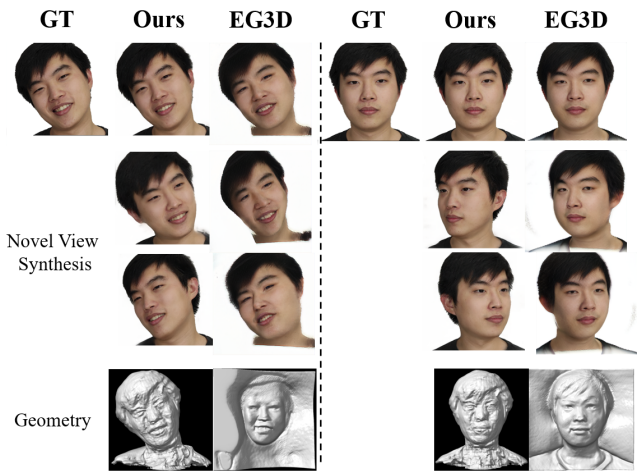

Fig. 19. Comparison with EG3D on novel view synthesis.

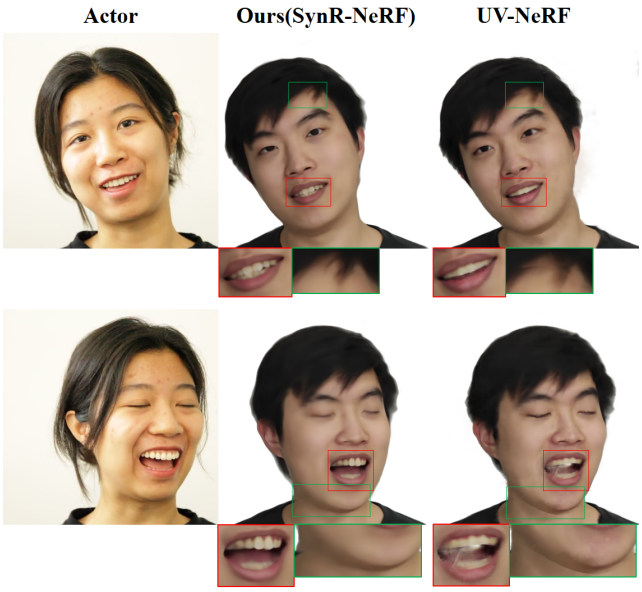

Fig. 20. Ablation study on model condition manner. Representing the dynamic feature on the mesh surface causes unrealistic artifacts at the edge region of the model mesh. Notice the ambiguous appearance in the mouth and the sharp seam around the neck.

Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. 2020. Deep 3d portrait from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7710–7720.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 601–610.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems* 33 (2020), 2492–2502.

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*. 12803–12813.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9459–9468.

Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C Bühler, Michael J Black, and Otmar Hilliges. 2022. IM Avatar: Implicit Morphable Head Avatars from Videos. *CVPR*.

A COMPARISON WITH EG3D

EG3D [Chan et al. 2022] is a state-of-the-art powerful generative model for HD 3D portrait, hence we compare it with our method to demonstrate the ability on novel view synthesis. After fitting the pretrained EG3D model with a single reference frame [Roich et al. 2021], we render the reconstructed 3D head in different views. As Fig. 19 shows, our method models a more vivid head avatar and presents more convincing novel view synthesis, which benefits from the joint learning of the temporal observation of the person-specific video data. Besides, EG3D performs worse on poses that are rare in FFHQ dataset. As for geometry, as we only utilize monocular view observation and do not apply any depth or sigma regularization, the geometry of our head avatar is noisier than EG3D's result.

B COMPARISON WITH UV-BASED NERF BASELINE

To validate the effectiveness of our adopted synthetic-rendering-based condition, we implement a mesh conditioned baseline (UV-NeRF) that encodes the feature map defined based on the UV parameterization [4] instead of our orthogonal plane features. For each sample point, the local feature, obtained from the nearest surface vertices, serves as the input to the MLP decoder. This baseline network is trained under the same setting as our network. The experiment is conducted based on a monocular dataset and the results are presented in Fig. 20. Not surprisingly, though UV-NeRF can accurately reconstruct the expression and reproduce reliable facial appearance, it generates unrealistic artifacts at the edge region of the model mesh. Notice the ambiguous appearance in the mouth and the sharp seam around the neck. Our synthetic-rendering-based condition fully utilizes the powerful convolutional network to learn the reasonable correspondence between the facial model and the entire head appearance, synthesizing consistent and stable images.

C ABLATION STUDIES ON THE TEXTURED MESH RENDERING CONDITION

To further validate our synthetic-rendering-based condition in detail, we implement two baselines that generate the feature volume directly from latent codes. The first one, called 'VectorPlane', uses the

[4]We set a learnable feature map in UV space which is shared for all frames, and utilize a U-Net with 7 layers and the number of channels used are 64, 128, 256, 512, *and*512. For each frame, the U-Net is input with the shared learnable feature map and per-frame UV normal map to generate an expression-related feature map on the UV parameterization. Specifically, the size for our shared UV feature map and generated expression-related UV feature map is $256x256x64$.

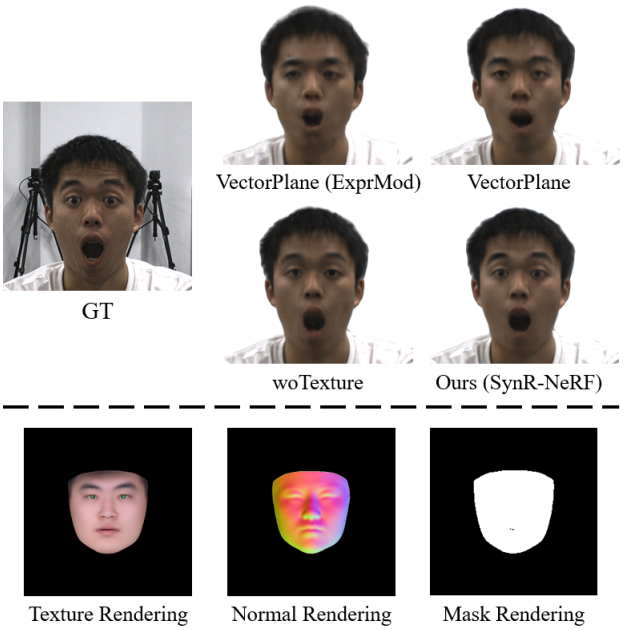

Fig. 21. Ablation studies on the textured mesh rendering condition. The last row visualizes the synthetic renderings of the front view.

expression parameters as input for the feature generation backbone, while the second one, 'VectorPlane(ExprMod)', feeds the expression parameters to a mapping network to modulate the convolutional kernels of the networks. Both baselines modulate the feature generation backbone with pose vectors and per-frame latent codes. As shown in Fig. 21 and Tab. 6, learning the mapping from the global vectors to appearances tends to overfit training sequences, and the ability of expression control degraded for out-of-distribution expressions.

We expound our analysis on the differences between the two conditional methods. Two similar expressions may be too close in the parameter space for a global-vector-conditioned feature generator to distinguish between them. Our synthetic-rendering-based conditional method preserves the spatial alignment between the mesh renderings and the feature planes, hence the local spatial changes in the rendering image space can be reflected in the feature volume. Additionally, as we align individual local features at the pixel level of renderings to the global context of the entire appearance, it is more likely to infer plausible results for unseen expressions.

Besides, we also implement a baseline, named 'woTexture', that only utilizes normal and mask renderings to condition feature volume generation. Removing texture rendering is feasible for person-specific avatar modeling and does not significantly affect the robustness of expression control. However, despite the similar numerical results, the visualization results of our method exhibit more detailed appearances around the eyes. We hypothesize that the texture rendering contains more high-frequency information around the eye region, as shown in the last row of Fig. 21, which may facilitate the network in effectively learning dynamic appearance around eyes.

	VectorPlane (ExprMod)	VectorPlane	woTexture	SynR-NeRF
PSNR	26.99	27.53	28.40	30.20
LPIPS	0.1446	0.1358	0.1268	0.1243

Table 6. Ablation studies on the textured mesh rendering condition.

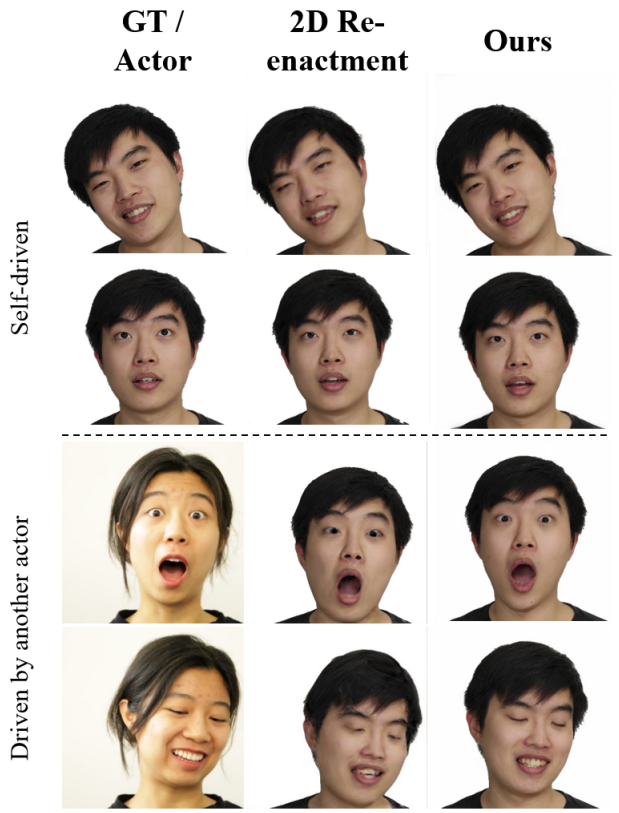

Fig. 22. Comparison with a 2D re-enactment baseline on monocular video datasets. The results demonstrate the superior performance of our method in terms of realistic appearance recovery and robust expression/pose control.

D COMPARISON WITH 2D RE-ENACTMENT BASED BASELINE

By omitting the Nerf module, we implement a 2D re-enactment method that only utilizes our image2image translation module. In our practice, the image2image translation network takes the renderings of the fitted 3DMM in the observed view as input and generates the corresponding 2D images. This 2D re-enactment method has limitations. Firstly, it cannot establish an avatar based on a multi-view dataset because it cannot differentiate between camera poses and head poses. Secondly, when applied to monocular videos, the 2D re-enactment method is sensitive to the location of the facial model in the image space, as shown in Fig. 22. The 2D re-enactment method is prone to generating artifacts or distorted faces, particularly at the edges of the image.

Data IO	Feature Plane Generation	NeRF Module	2D Image Module	**Total**
47.4ms	22.3ms	58ms	24.5ms	152.2ms

Table 7. Detailed time consuming during inference.

E MULTI-VIEW SETTING

While fitting per-frame 3DMM model, we use the detected landmarks in multi-view images at the same instant as supervision, and additionally estimate the scale parameter of 3DMM. As for NeRF optimization, we simply leverage multi-view images of the same frame to supervise the appearance, with all loss terms and the training strategy as same as monocular-based setting.

F DATA PREPROCESSING

We optimize the shape and texture parameters using the first few frames of the video and these parameters remain fixed for the remaining frames of the video. For each frame's 3DMM fitting, we optimize the pose, expression, and illumination parameters, which are initialized as the last frame's fitting results. Once the 3DMM model is tracked, we utilize PyTorch3D [Ravi et al. 2020] to render per-frame synthetic renderings onto orthogonal planes. The tracking code mainly comes from the open-source project Faceverse [5]. Besides, to perform eye gaze tracking, we segment the dark area within the eye region in the given frame. The centroid of the dark area is considered the pupil, and we calculate the pupil's relative position inside the eye based on the detected landmarks surrounding the eye. Finally, we mark the pupil as a small dot in the front-view orthogonal renderings.

G INFERENCE TIME

Tab. 7 shows the detailed time consuming during inference. Rendering a color image with a resolution of 512x512 takes 0.15 seconds on one NVIDIA 3090 GPU, and the most time-intensive part is the volume rendering of our NeRF module.

[5](https://github.com/LizhenWangT/FaceVerse)