High-Fidelity 3D Head Avatars Reconstruction through Spatially-Varying Expression Conditioned Neural Radiance Field

Minghan Qin*, **Yifan Liu***, **Yuelang Xu**, **Xiaochen Zhao**, **Yebin Liu** †, **Haoqian Wang** †
Tsinghua University
{qmh21,yf-liu21,xll20,zhaoxc19}@mails.tsinghua.edu.cn, liuyebin@mail.tsinghua.edu.cn, wanghaoqian@tsinghua.edu.cn

Abstract

One crucial aspect of 3D head avatar reconstruction lies in the details of facial expressions. Although recent NeRF-based photo-realistic 3D head avatar methods achieve high-quality avatar rendering, they still encounter challenges retaining intricate facial expression details because they overlook the potential of specific expression variations at different spatial positions when conditioning the radiance field. Motivated by this observation, we introduce a novel Spatially-Varying Expression (SVE) conditioning. The SVE can be obtained by a simple MLP-based generation network, encompassing both spatial positional features and global expression information. Benefiting from rich and diverse information of the SVE at different positions, the proposed SVE-conditioned NeRF can deal with intricate facial expressions and achieve realistic rendering and geometry details of high-fidelity 3D head avatars. Additionally, to further elevate the geometric and rendering quality, we introduce a new coarse-to-fine training strategy, including a geometry initialization strategy at the coarse stage and an adaptive importance sampling strategy at the fine stage. Extensive experiments indicate that our method outperforms other state-of-the-art (SOTA) methods in rendering and geometry quality on mobile phone-collected and public datasets. Code and data can be found at https://github.com/minghanqin/AvatarSVE.

Introduction

Reconstructing controllable and realistic 3D head avatars is beneficial in many applications, such as VR/AR, games, and teleconferencing. The facial expression details are crucial for realistic 3D head avatars. Current 3D head avatar reconstruction methods can generate controllable head avatars from monocular videos. However, achieving an accurate geometry of facial expressions and nuanced and individualized details remains a substantial challenge.

To reconstruct expressive 3D head avatars, some methods (Gafni et al. 2021; Gao et al. 2022; Athar et al. 2022; Zheng et al. 2022; Zielonka, Bolkart, and Thies 2023; Xu et al. 2023) based on neural radiance fields (NeRF) (Mildenhall et al. 2020) achieves the photo-realistic rendering. However, these implicit neural radiance field-based approaches exhibit the **insufficient ability to render detailed complex expressions and the corresponding geometry**. As shown in Fig. 1, these methods typically employ an optional deformation network $\mathcal{D}$ to represent face expression motions and a NeRF $\mathcal{F}$ to model head geometry and appearance. Both $\mathcal{D}$ and $\mathcal{F}$ are conditioned on the **global expression** from 3D Morphable Models (3DMMs). Although recent methods (Athar et al. 2022; Zielonka, Bolkart, and Thies 2023) focus on better exploiting the global expression in an elaborately-designed deformation $\mathcal{D}$, they still directly utilize the global expression as the conditioning for NeRF $\mathcal{F}$. This direct global expression conditioning struggles to provide fine-grained control over the geometry and rendering at different positions within the 3D space. As a result, these methods struggle to obtain detailed rendering and accurate geometry when dealing with complex expressions.

To solve the above limitation, we propose **Spatially-Varying Expression** (SVE) as the conditioning. As indicated in Fig. 1(a), the global expression ε from 3DMM stays constant across 3D space. The global expression ε solely encompasses 3DMM template expressions, limiting $\mathcal{D}$ and $\mathcal{F}$ in capturing nuances and spatial intricacies, e.g., eyes, teeth, wrinkles (as shown in Fig. 3). In contrast, SVE ε' in Fig. 1(b) varies across 3D space, encompassing both spatial positional information and expression information. Therefore, SVE helps NeRF capture the intricate movement of wrinkles, eyes, mouth, eyebrows, etc. Specifically, to generate SVE, we design a simple generation network $\mathcal{G}$ to integrate ε with the spatial positional features. To reduce errors in geometry reconstruction from inadequate constraints and enhance overall quality, we also introduce a coarse-to-fine training strategy to enhance the geometry at the coarse stage via the geometry initialization and improve the rendering quality at the fine stage by an adaptive importance sampling strategy. Extensive experiments show that our method achieves significantly superior results by employing the SVE as conditioning in terms of accurate geometry and detailed rendering, especially when dealing with intricate expressions. We present the contributions of our method as follows:

- We propose a 3D head avatar method based on the Spatially-Varying Expression conditioned neural radiance field. The proposed Spatially-Varying Expression (SVE) enables the radiance field to capture intricate expressions and detailed geometry faithfully.

*These authors contributed equally.
†These authors are co-corresponding authors.

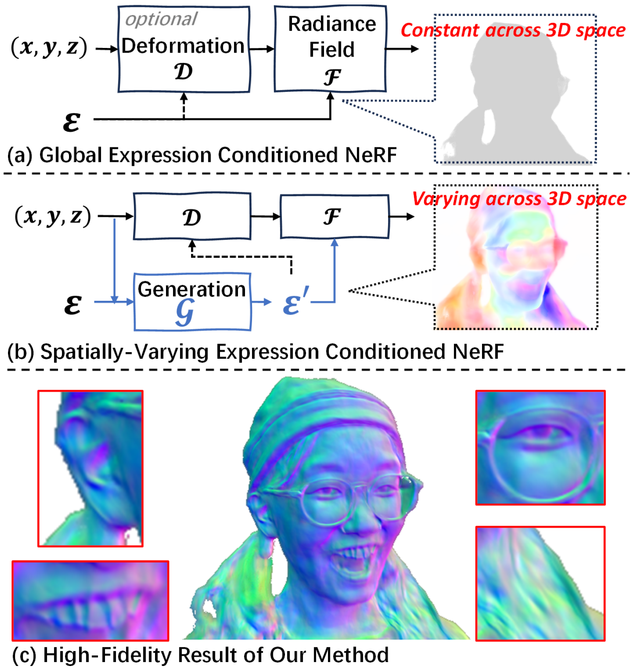

Figure 1: Previous 3D head avatar methods based on (a) 3DMM global expression ε conditioned NeRF *vs* (b)(c) Ours based on Spatially-Varying Expression ε' conditioned NeRF. We visualize ε and ε' via volume rendering by replacing the RGB with ε and ε'. The template expressions ε used in (a) limit the $\mathcal{F}$ to capture expression and geometry details. While ε' in (b) incorporates expression and spatial positional features, guiding the $\mathcal{F}$ for enhanced expression rendering and geometry. We present high-quality geometric results of our method in (c).

- We design a simple generation network to generate the Spatially-Varying Expression by integrating the spatial positional features into the global expression.
- We introduce a novel coarse-to-fine training strategy involving geometry initialization for credible reconstruction and adaptive importance sampling for enhanced rendering and geometry, thus refining avatar expression details.

Related Work

3D Head Avatar Reconstruction. Reconstructing 3D face models and head avatars has gained extensive research in recent years (Ichim, Bouaziz, and Pauly 2015; Cao et al. 2015, 2016; Hu et al. 2017; Nagano et al. 2018; Athar, Shu, and Samaras 2023; Zheng et al. 2022; Athar et al. 2022; Grassal et al. 2022; Chan et al. 2022; Zielonka, Bolkart, and Thies 2023; Zheng et al. 2023; Kirschstein et al. 2023; Sun et al. 2023). Traditional 3DMM (Blanz and Vetter 1999; Gerig et al. 2018) models appearance and geometry on linear space by PCA analysis. FLAME (Li et al. 2017) and other extension methods (Feng et al. 2021; Zielonka, Bolkart, and Thies 2022; Danecek, Black, and Bolkart 2022) additionally incorporate the eye and neck modeling, achieving the reconstruction of the whole head with vivid expressions and optimized texture.

As NeRF (Mildenhall et al. 2020) shows great potential in photo-realistic rendering, some methods (Gafni et al. 2021; Xu et al. 2023; Gao et al. 2022) explore NeRF to reconstruct controllable 3D head avatars by conditioning the NeRF with the global 3DMM tracked expression parameters. (Gafni et al. 2021) directly using the conditional neural radiance field without using any deformation networks. (Gao et al. 2022) utilize multiple multi-level hash tables to represent a specific expression. (Xu et al. 2023) leverages a lightweight deformation network with voxel features generated by the global expression.

Some methods (Athar et al. 2022; Zheng et al. 2022; Zielonka, Bolkart, and Thies 2023) have also explored the combination of traditional 3DMM and neural rendering by leveraging the 3DMM face template as a prior for deformation. (Athar et al. 2022) and (Zielonka, Bolkart, and Thies 2023) leverage the tracked face mesh templates to guide the deformation network. (Zheng et al. 2022) proposes implicit mophorable models to incorporate 3DMM into volume rendering framework. These methods directly exploit the head geometry estimated by 3DMM, thus avoiding incorrect head geometry, such as facial concavities. However, since these methods rely on the 3DMM geometry prior, incorrect and excessively smooth surface estimation prevents the model from learning detailed, intricate expressions, hair, accessories, and clothing, leading to coarse geometry and rendering without rich details.

Overall, the methods discussed above neglect leveraging the expression variations at different spatial positions for modeling detailed geometry and appearances. In contrast, our methods explore the potential of the Spatially-Varying Expression conditioning, leading to detailed geometry and rendering results even with intricate expressions.

Recently researchers also have explored training a general generative human head model from large-scale datasets (Chan et al. 2022; Sun et al. 2023, 2022; Wang et al. 2022; Zhuang et al. 2022; Hong et al. 2022), audio-driven talking head avatars (Guo et al. 2021; Liu et al. 2022), avatars from dense multi-view data (Ma et al. 2021; Lombardi et al. 2018, 2019; Chu et al. 2020; Lombardi et al. 2021; Cao et al. 2022), and one-shot head avatars (Drobyshev et al. 2022), which is beyond the research topic of this work.

Dynamic Neural Radiance Field. We aim to reconstruct controllable 3D head avatars from monocular RGB videos. To model monocular RGB videos, we employ NeRF-based dynamic scene modeling methods (Cao and Johnson 2023; Sara Fridovich-Keil and Giacomo Meanti et al. 2023; Pumarola et al. 2021; Park et al. 2021a,b; Fang et al. 2022) extend static NeRF (Mildenhall et al. 2020) to model dynamic scenes by adding additional temporal information and deform fields. D-NeRF (Pumarola et al. 2021) leverages the scene encoder to estimate the scene offsets between predefined canonical space and the current observation space from temporal embedding. Deformable NeRF (Park et al. 2021a) explores a dense SE(3) deform field conditioned on frame-wise learnable latent codes. HyperNeRF (Park et al. 2021b) extends Deformable NeRF in terms of topological changes problem since the continuity of dense deform field can not model discontinuous topological changes.

Recent methods (Cao and Johnson 2023; Sara Fridovich-Keil and Giacomo Meanti et al. 2023) extend efficient triplane representation (Chan et al. 2022) to dynamic scenes and model dynamic scenes without explicit deform fields for accelerating. Unlike existing dynamic NeRFs, we design the deformation as a tiny network, directly utilizing the Spatially-Varying Expression to predict 6D motions. Benefiting from the rich information of Spatially-Varying Expression, our method achieves competitive results compared to methods with well-signed deform networks. In addition to the former dynamic scene modeling methods, our method can control various expressions for self and cross-identity reenactment.

Preliminary

Our method is based on the neural radiance field (NeRF), combining the conditional NeRF and the deformable NeRF for better avatar motion control.

NeRF (Mildenhall et al. 2020) is an implicit function-based volumetric rendering technique which enables photo-realistic novel view synthesis. The radiance field $\mathcal{F}$ maps 3D spatial query points $p = (x, y, z)$ and the corresponding 2D view direction d to density σ and color c.

$$\sigma, c = \mathcal{F}_{\theta_F}(p, d) \tag{1}$$

where θ_F is learnable parameters of the radiance field $\mathcal{F}$. By computing the density σ and color c of each query point p along a ray from the camera origin o through a pixel $p_{2d} = (u, v)$, the RGB value of the pixel p can be obtained through integration of volume rendering.

Conditional NeRF (Gafni et al. 2021) with additional parameters ε enables NeRF's adaptability to varying condition information. In 3D head avatar reconstruction, most methods utilize pre-tracked global 3DMM expression parameters ε as the conditioning. The condition process is usually implemented by direct concatenation or addition of encoded p and ε.

$$\sigma, c = \mathcal{F}_{\theta_F}(p, d, \varepsilon) \tag{2}$$

Deformable NeRF (Park et al. 2021a) disentangles the shape and motion for dynamic scenes. Deformable NeRF introduces a dense deformation field $\mathcal{D}$ to deform query points $p_o = (x, y, z)$ from the observation space of the current frame to a canonical space $p_c = (x', y', z')$, then estimate the radiance field $\mathcal{F}$ in the canonical space.

$$\begin{aligned} p_c &= \mathcal{D}_{\theta_D}(p_o, t) \\ \sigma, c &= \mathcal{F}_{\theta_F}(p_c, d) \end{aligned} \tag{3}$$

where θ_D denotes learnable parameters of the dense deformation field $\mathcal{D}$. t represents a certain frame of the dynamic scene.

Method

Formulation

Previous methods have exploited the power of NeRF for photo-realistic 3D head avatar reconstruction. A typical NeRF-based head avatar model can be formulated as an expression-conditioned deformable NeRF as Eq. 4.

$$\begin{aligned} p_c &= \mathcal{D}_{\theta_D}(p_o|\varepsilon) \\ \sigma, c &= \mathcal{F}_{\theta_F}(p_c, d|\varepsilon) \end{aligned} \tag{4}$$

where ε denotes the global expression parameters obtained by tracking the input video using 3DMM face templates. Previous methods condition the radiance field $\mathcal{F}$ directly on the global expression parameters ε, neglecting the expression conditioning for specific spatial positions. Therefore, this per-frame global expression conditioning is insufficient to obtain detailed expression rendering and geometry.

Motivated by this observation, we present a novel approach utilizing the Spatially-Varying Expression (SVE) conditioned NeRF. Our method can be formulated as Eq. 5.

$$\begin{aligned} \varepsilon' &= \mathcal{G}_{\theta G}(\varepsilon|p_o) \\ p_c &= \mathcal{D}_{\theta_D}(p_o|\varepsilon') \\ \text{SDF}, c &= \mathcal{F}_{\theta_F}(p_c, d|\varepsilon') \end{aligned} \tag{5}$$

where $p_o = (x, y, z)$ is the query points in the observation space. $p_c = (x', y', z')$ is the query points deformed by network $\mathcal{D}$ in the canonical space. We select NeuS (Wang et al. 2021) as $\mathcal{F}$, which incorporates the Signed Distance Function (SDF) as the implicit representation. And c is the predicted color of each query point. $C = f(A|B)$ means the function f maps A to C conditioned on B. In contrast to global expression parameters ε, ε' denotes the generated compressed Spatially-Varying Expression parameters via a simple generation network $\mathcal{G}$. Through the network $\mathcal{G}$, we effectively integrate a certain frame's global expression with the spatial positional information, acquiring the Spatially-Varying Expression.

Spatially-Varying Expression Conditioned NeRF

Overview. As depicted in Fig. 2, given a certain training frame, we first extract the per-frame global expression parameters ε by 3DMM pre-processing. During the training, We first obtain rays according to camera poses and head poses. Then, after sampling a set of points p_o in the current frame observation space, we utilize the proposed generation network $\mathcal{G}$ to generate the Spatially-Varying Expression (SVE) parameters ε' with p_o as the additional condition to provide spatial positional features. Subsequently, we leverage a tiny deformation network $\mathcal{D}$ to specify the 6D motion R, T of each query point p_o according to the generated SVE parameters ε', and deform p_o to the query points in canonical space p_c. The deformation module $\mathcal{D}$ is designed as a remarkably simple structure without leading to performance degradation due to benefiting from the spatial information contained in ε'. Subsequently, the neural radiance field $\mathcal{F}$ based on NeuS takes the deformed query points p_c as inputs and the SVE parameters ε' as the conditioning, yielding the SDF and color values c. Finally, by employing the volumetric rendering technique, we integrate the color values c of each ray to obtain the rendered RGB of each pixel. In the meanwhile, we also integrate the gradient of each point's SDF value to get the rendered normal map.

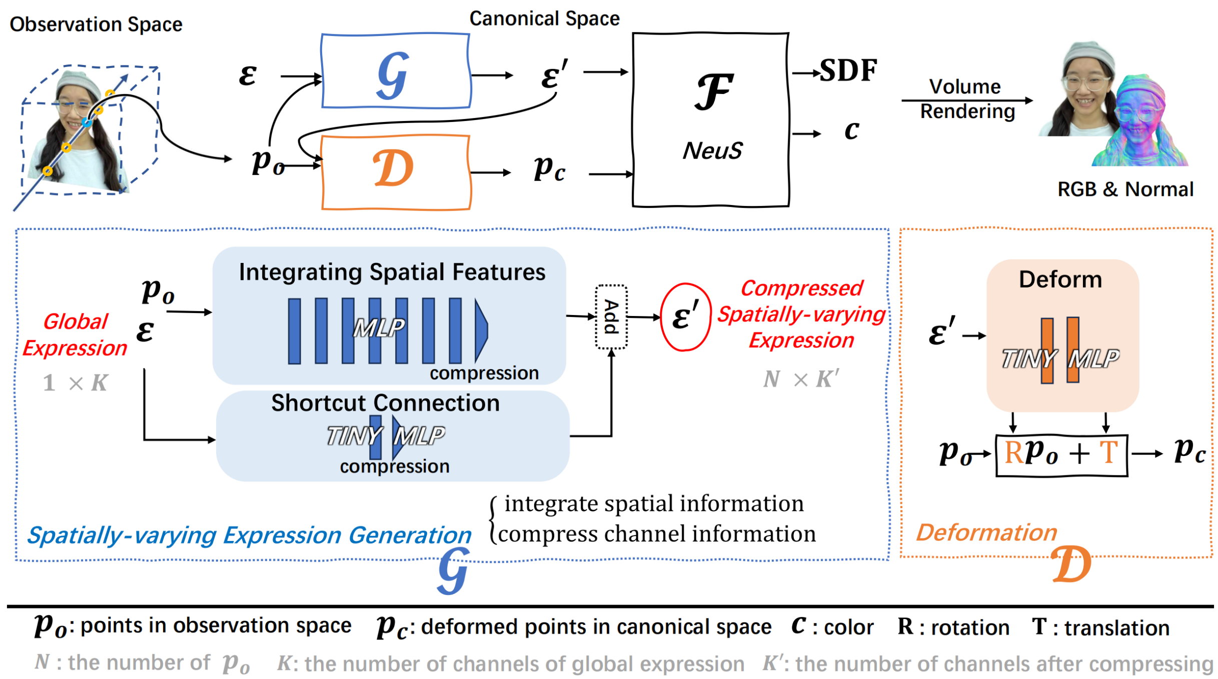

Figure 2: Method Overview. Given a portrait video, We first track the global expression parameters ε using 3DMM (Gerig et al. 2018). After the pre-processing, given the sampled 3D points p_o in observation space, we apply the generation network $\mathcal{G}$ to extend the global expression parameters ε with the spatial positional features of each position p_o in 3D space. Then, through a deformation network D, we transform p_o from the observation space to the p_c in the canonical space conditioned on ε'. Subsequently, we use ε' conditioned NeuS (Wang et al. 2021) to predict the SDF values and color c corresponding to p_c. Finally, we obtain the rendered RGB image and normal using volumetric rendering.

Spatially-Varying Expression generation. The proposed term **Spatially-Varying** refers to changing across different spatial positions. Previous methods (Athar et al. 2022; Zielonka, Bolkart, and Thies 2023) directly utilize the global expression parameters from 3DMM when modeling avatars' geometry and appearance. These methods focus on enhancing the utilization of the global expression parameters during the deformation $\mathcal{D}$, neglecting the potential to fully leverage the information contained in expressions for modeling both geometry and appearance. However, various facial expressions influence the intricacies of facial geometry and texture. For instance, a laughing expression does not only involve the mouth being open. The muscles' motion throughout the face, the raised angle of eyebrows and skin-creasing should exhibit varying levels of geometric and textural changes. Even with well-designed deformation networks, previous methods are challenging to capture such nuanced expression changes. Consequently, these prior methods based on the global expression inhibit the capacity of the neural radiance field to learn each 3D position's distinct features of the expression.

In contrast, our proposed expression parameters explore the potential of the radiance field to learn distinct features for different positions across the 3D space. To elaborate, the deformations (R and T) and characteristics (SDF and c) associated with each point $p = (x, y, z)$ in 3D space should rely not only on its spatial position (x, y, z) but also on the global expression parameters ε. Considering that this variable changes in conjunction with alterations in 3D positions, it is referred to as **Spatially-Varying Expression**.

Specifically, as shown in Fig. 2, the generation module $\mathcal{G}$ comprises two networks. The spatial feature integrating network amalgamates the spatial positional information contained in query points p_o with the expression parameters. This network learns the influence of the global expression parameters on these specific positions $p_o = (x, y, z)$. Simultaneously, the shortcut connection compresses the global expression parameters ε for residual addition. While the integrating network is parameterized as a Multilayer Perceptron (MLP) consisting of 8 fully-connected layers, the shortcut connection only consists of 2 fully-connected layers, serving as dimensional mapping for residual addition. Through the generation module $\mathcal{G}$, we obtain the Spatially-Varying Expression parameters, which encapsulate both the global expression information and the spatial positional features.

Note that in both the integrating network and the shortcut connection, we reduce the dimension of the outputs from K to K' as shown in Fig. 2 to avoid over-fitting. Intuitively, one high-dimensional global expression parameter can sufficiently capture the coarse changes of facial expressions validated by previous methods discussed above. However, when employing Spatially-Varying Expression as conditioning for D and F, it degrades performance when generalizing to new expressions due to the excessive information for each query point p_o. Therefore, by reducing the dimension of the Spatially-Varying Expression codes, we effectively avoid over-fitting without performance degradation.

Expression conditioned deformation. As we discussed above, the 6D motion deformation R, T of each point $p_o = (x, y, z)$ in the 3D observation space should rely on the generated Spatially-Varying Expression ε'. Therefore, we design the deformation as a lightweight tiny MLP consisting of two fully-connected layers. Compared to the well-designed deformation modules proposed by prior methods, our deformation network $\mathcal{D}$ benefits from the rich position-dependent expression information from the Spatially-Varying Expression parameters, thus effectively predicting accurate 6D motions with only a lightweight network.

Coarse-to-Fine Training Strategy

Geometry initialization of coarse stage. In monocular 3D head avatar reconstruction, the geometry often encounters geometric collapse: the facial structure deviations from the intended geometry, leading to varying degrees of concavity. Previous methods (Zheng et al. 2022, 2023) try to address this issue by incorporating 3DMM templates as a prior into the deformation network $\mathcal{D}$ to constrain the geometry. Nonetheless, the inherent smoothness and limitation of the 3DMM itself lead to lacking intricate geometry details of facial features, hair, clothing, etc.

To tackle this issue, we present an innovative geometry initialization strategy at the coarse training stage to achieve a harmonious equilibrium between intricacies and smoothness geometries. We harness the pseudo-depth of the tracked 3DMM model to safeguard the geometry against geometric collapse. Instead of using strong 3DMM constraints throughout the training, the initialization strategy avoids excessively smooth geometry.

Specifically, in the geometry initialization, we employ the rendered pseudo-depth D of the tracked 3DMM template, as shown in the supplementary appendix. We map the sampled pixels of rendered pseudo-depth into the 3D observation space points p_o^d. Then, we predict the SDF values SDF^d corresponding to p_o^d according to Eq. 5. Because p_o^d lie on the approximate surface of the face, their associated SDF values SDF^d should naturally tend towards zero. To enforce this, we employ a geometry loss to constrain the predicted SDF^d. Furthermore, by utilizing an L1 loss, we also guide the alignment of the rendered depth $\hat{D}$ with its corresponding pseudo-depth D.

Adaptive importance sampling strategy of the fine stage. The adaptive importance sampling strategy is proposed to achieve a sensitive perception of infrequent areas and small areas, e.g., teeth and rim glasses. In contrast to commonly used random pixel sampling (Mildenhall et al. 2020), importance sampling with fixed-weight (Gafni et al. 2021) or with the pre-computed weight (Li et al. 2022), our strategy automatically adapts to different training data, and dynamically adjust the weight during the training.

Specifically, for each training frame, we first segment the frames into $N = 19$ semantic regions, encompassing various components such as eyes, face, hair, lips, etc. Specific classification criteria are detailed in the supplementary materials. We assign weights $w_i^s, i = 1, 2, \ldots, N$ to each region at the s-th training step. During the s-th training step, we calculate the guidance loss L_i^s for sampled points within each region based on Eq. 6 and the area of each region A_i^s of the current frame.

$$\begin{aligned} L_i^s &= \lambda_1 L^s_{i,\text{render}} + \lambda_2 L^s_{i,\text{depth}} \\ L^s_{i,\text{render}} &= M_i \|\hat{C}_i - C_i\|_1 + BCE(\hat{M}_i, M_i) \\ L^s_{i,\text{depth}} &= M_i \|\hat{D}_i - D_i\|_1 \end{aligned} \tag{6}$$

where M_i, C_i, and D_i stand for the ground-truth mask, color, and pseudo-depth of the region i. $\hat{M}$, $\hat{C}$, and $\hat{D}$ represent the corresponding predictions. The values λ_1 and λ_2 help decide whether L_i^s should concentrate more on parts with unsatisfactory rendering or areas with subpar geometry.

Next, we utilize the loss L_i^s for guidance to update w_i^s using exponential moving average (EMA) according to Eq. 7. The EMA updating stabilizes the updating and also addresses the issue of not being able to resample region i in a training step when its area A_i^s becomes 0. Because when A_i^s is 0, the updated weight w_i^{s+1} also becomes 0 without EMA. Consequently, in subsequent training steps, both L_i^{s+m} and w_i^{s+m} with $m \geq 1$ remain 0. In such cases, the importance sampling will not sample points within region i.

$$w_i^{s+1} = \left(\frac{L_i^s}{w_i^s A_i^s \sum_i L_i^s}\right) \cdot \alpha + w_i^s \cdot (1 - \alpha) \tag{7}$$

where α is the updating ratio. This loss-guided sampling strategy adaptively encourages the model to prioritize previously inadequately learned regions, leading to improved expression realism and rendering quality.

Experiments

Datasets and Preprocessing

Datasets. We collect seven monocular RGB sequences of different subjects. All videos are collected using an iPhone 12 front camera with a fixed camera pose and a length of 2000-4000 frames. The image resolution of each video is 480×480. The content of each collected video includes facial expression changes, head pose changes, and talking. Additionally, to evaluate the effectiveness of our method on public datasets, we also conduct experiments on two open-source datasets from IMAvatar (Zheng et al. 2022) and NeRFace (Gafni et al. 2021).

Pre-processing. During the pre-processing, we exploit PP-Matting (Chen et al. 2022) to generate foreground masks and the face parsing method (zllrunning 2019) to obtain coarse semantic segmentation maps for subsequent adaptive importance sampling. The BFM Model (Gerig et al. 2018) is used to track the collected video's head poses and expression parameters. We render the predicted pseudo-depth from BFM-tracked face meshes.

Comparison on Avatar Reconstruction Quality

We conducted qualitative and quantitative comparisons with three SOTA NeRF-based 3D head avatar reconstruction

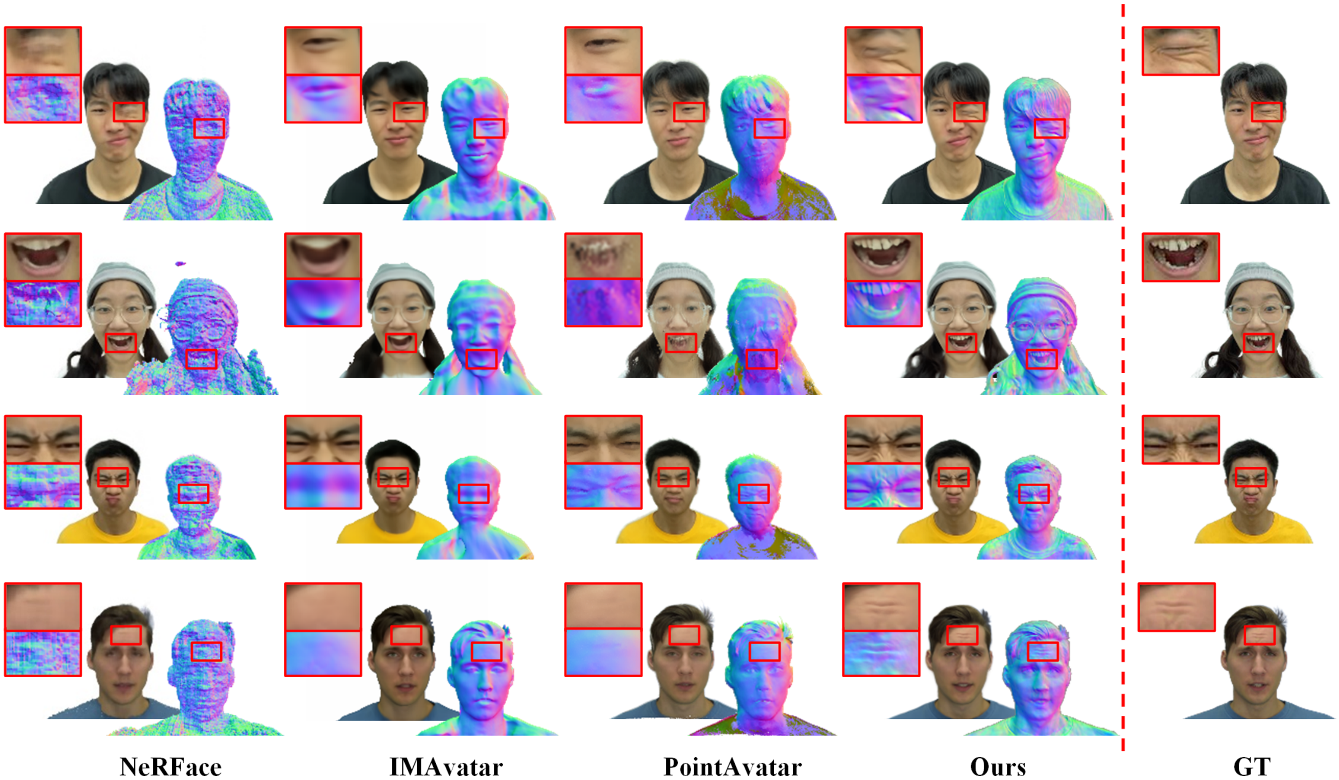

Figure 3: Qualitative comparisons on self reenactment task. From left to right: NeRFace (Gafni et al. 2021), IMAvatar (Zheng et al. 2022), PointAvatar (Zheng et al. 2023), and Ours. Our method reconstructs high-quality rendering and geometric details of wrinkles, teeth, hairs, and accessories. We recommend zooming in to see more details.

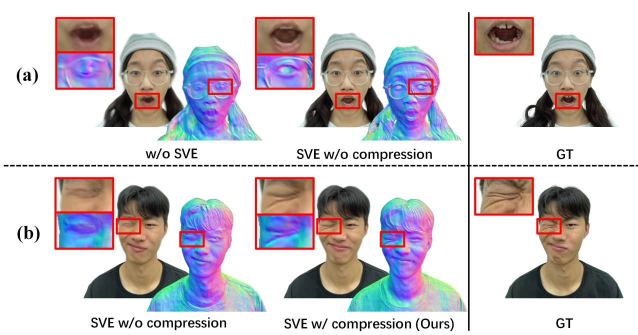

Figure 4: Qualitative ablation results of Spatially-Varying Expression (SVE).

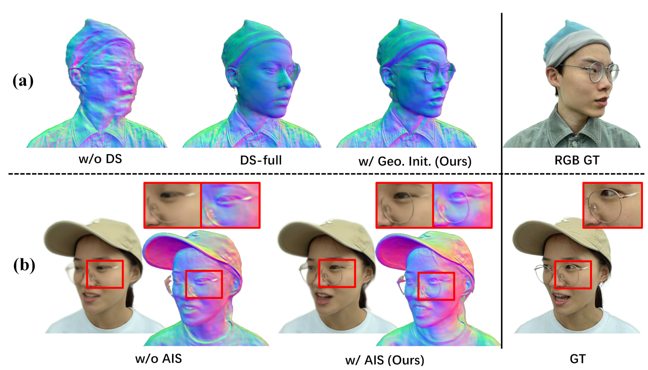

Figure 5: Qualitative ablation results of the geometry initialization (Geo. Init.) strategy and the adaptive importance sampling (AIS). In the case w/o DS, the concavity and convexity of the face are incorrect.

methods: NeRFace (Gafni et al. 2021), IMAvatar (Zheng et al. 2022), and PointAvatar (Zheng et al. 2023). NeRFace directly uses global expression conditioned NeRF without explicit deformation. IMAvatar employs implicit morphing based on the FLAME head templates (Athar, Shu, and Samaras 2023) in the deformation network to utilize the expression parameters. PointAvatar extends IMAvatar by utilizing a point-based neural representation approach for efficient training. We recommend reading the supplementary appendix and video for more experimental results, including novel view synthesis, cross-identity reenactment, and additional comparison results.

The qualitative results of the self-reenactment task are shown in Fig. 3. The results suggest that IMAvatar and PointAvatar reconstruct coarse head geometry. Constrained by the smooth 3DMM template, these two approaches struggle to capture intricate expression details, especially on our

	L1↓	PSNR↑	SSIM↑	LPIPS↓
NeRFace	0.0195	24.976	0.933	0.122
IMavatar	0.0187	25.360	0.927	0.145
PointAvatar	0.0207	24.799	0.918	0.130
Ours	**0.0148**	**27.748**	**0.944**	**0.0925**

Table 1: Quantitative evaluations on self-reenactment task. We show the Average performance on all nine subjects. Our method notably outperforms all SOTA approaches. For separate results of individual subjects, please refer to the supplementary appendix.

Exp	SVE	SVE-C	DS	GI	AIS	L1↓	PSNR↑	SSIM↑	LPIPS↓
(1)			✓	✓	✓	0.0198	24.098	0.922	0.0963
(2)	✓		✓	✓	✓	0.0188	24.090	0.925	0.0923
(3)	✓	✓			✓	0.0199	23.707	0.920	0.0933
(4)	✓	✓	✓		✓	0.0190	24.069	0.924	0.0925
(5)	✓	✓	✓	✓		0.0191	24.011	0.921	0.0931
Ours	✓	✓	✓	✓	✓	**0.0187**	**24.335**	**0.927**	**0.0911**

Table 2: Quantitive results of ablation studies. SVE-C denotes SVE with compression. DS denotes depth supervision. GI denotes geometry initialization. AIS denotes the adaptive importance sampling strategy.

datasets with complex expression variations. NeRFace relies on global expression parameters, limiting its incorporation of distinct expression information across spatial positions. Consequently, it struggles to render accurate, fine-grained expression details. Moreover, without reliance on prior 3DMM template knowledge and geometric optimization, NeRFace fails to accurately reconstruct facial geometry. In contrast, our method leverages the Spatially-Varying Expression as the conditioning of the radiance field, effectively harnessing the information embedded within 3DMM expression parameters and spatial positional features. As a result, our method markedly outperforms the previous SOTA methods in terms of expression detail refinement and geometric reconstruction quality.

As for quantitative comparisons, We evaluate the rendering quality and expression similarity of all the methods discussed above. Tab. 1 reports several commonly used metrics for rendering quality evaluation, including Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). We computed the average performance across nine subjects. Our method notably outperforms all SOTA approaches. Furthermore, in the supplementary appendix, we provide individual quantitative comparison results for all subjects to offer more compelling evidence for the comparisons.

Ablation Study

Effectiveness of the Spatially-Varying Expression (SVE). To validate the effectiveness of SVE, we design these two baselines: (1) **w/o SVE.** We use the tracked 64-dimensional 3DMM expression parameters as the NeRF's condition. (2) **SVE w/o compression.** We replace the shortcut compression in the generation network $\mathcal{G}$ to an identity mapping and modify the integrating network of $\mathcal{G}$ to retain 64-dimensional features. Then we consider the addition of the output of the two branches as the condition. Please refer to the supplementary appendix for detailed illustrations of these two baselines. Fig. 4(a) demonstrates that employing SVE without compression can exploit spatial positional information, resulting in clearer reconstructed expression details than using global expression only. As depicted in Fig. 4(b), compression of SVE mitigates the risk of over-fitting and improves expression detail reconstruction quality.

Effectiveness of the coarse-to-fine training strategy. To validate the effectiveness of the geometry initialization (GI) and the adaptive importance sampling (AIS) in coarse-to-fine training strategy, we design these three baselines: (3) **w/o Depth Supervision (w/o DS).** We refrain from employing the predicted depth as supervision. (4) **Depth Supervision during full training stage (DS-full).** We incorporate the predicted depth as supervision throughout the training process. (5) **w/o AIS.** We employ a random sampling approach to sample pixels in training. As depicted in Fig. 5(a), the approach employing geometry initialization accomplishes refined geometric reconstruction. Also, Fig. 5(b) shows that the adaptive importance sampling directs the network's focus towards overlooked intricate regions.

Conclusions

In this paper, we have proposed a 3D head avatar reconstruction method through Spatially-Varying Expression (SVE) conditioned NeRF. The SVE integrates global expression with localized spatial positional features, enabling the radiance field to capture intricate expressions and geometric details accurately. We employ a concise yet effective MLP-based generation network to produce the compressed SVE by integrating spatial positional features with the global expression from 3DMM. Furthermore, we introduce an innovative coarse-to-fine training strategy, including a geometry initialization technique and adaptive importance sampling strategy, thus further refining the expression details of avatars. Prior to this work, there has been a lack of focus on efficiently leveraging the global expression to achieve improved conditioned NeRF for reconstruction quality. We aspire for this study to garner attention from researchers and instigate ongoing explorations in this direction.

Discussion

Limitation. Despite achieving high-quality reconstruction of 3D head avatars, the generalization capacity of our method remains constrained by the distribution of data. Our method encounters challenges in generating distinct teeth when the dataset predominantly comprises instances of mouth closing. The failure cases are shown in the supplementary appendix.

Future Work. To address the above shortcomings, we will try to train a 3D head avatar reconstruction model with enhanced expression generalization capabilities using large-scale facial video datasets.

Acknowledgements

This research was funded through National Key R&D Program of China (Project No.2022YFB36066 and 2021ZD0113502), in part by the Shenzhen Science and Technology Project under Grant (JCYJ20220818101001004 and JSGG20210802153150005).

References

Athar, S.; Shu, Z.; and Samaras, D. 2023. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–8.

Athar, S.; Xu, Z.; Sunkavalli, K.; Shechtman, E.; and Shu, Z. 2022. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20364–20373.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Cao, A.; and Johnson, J. 2023. HexPlane: A Fast Representation for Dynamic Scenes. *CVPR*.

Cao, C.; Bradley, D.; Zhou, K.; and Beeler, T. 2015. Real-Time High-Fidelity Facial Performance Capture. *ACM Trans. Graph.*, 34(4).

Cao, C.; Simon, T.; Kim, J. K.; Schwartz, G.; Zollhoefer, M.; Saito, S.-S.; Lombardi, S.; Wei, S.-E.; Belko, D.; Yu, S.-I.; Sheikh, Y.; and Saragih, J. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Trans. Graph.*, 41(4).

Cao, C.; Wu, H.; Weng, Y.; Shao, T.; and Zhou, K. 2016. Real-Time Facial Animation with Image-Based Dynamic Avatars. *ACM Trans. Graph.*, 35(4).

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; Mello, S. D.; Gallo, O.; Guibas, L.; Tremblay, J.; Khamis, S.; Karras, T.; and Wetzstein, G. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16102–16112.

Chen, G.; Liu, Y.; Wang, J.; Peng, J.; Hao, Y.; Chu, L.; Tang, S.; Wu, Z.; Chen, Z.; Yu, Z.; et al. 2022. PP-Matting: High-Accuracy Natural Image Matting. *arXiv preprint arXiv:2204.09433*.

Chu, H.; Ma, S.; Torre, F.; Fidler, S.; and Sheikh, Y. 2020. Expressive Telepresence via Modular Codec Avatars. In *Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV)*, 330–345.

Danecek, R.; Black, M. J.; and Bolkart, T. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 20311–20322.

Drobyshev, N.; Chelishev, J.; Khakhulin, T.; Ivakhnenko, A.; Lempitsky, V.; and Zakharov, E. 2022. MegaPortraits: One-shot Megapixel Neural Head Avatars.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers*.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4): 88:1–88:13.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 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.

Gao, X.; Zhong, C.; Xiang, J.; Hong, Y.; Guo, Y.; and Zhang, J. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 41(6).

Gerig, T.; Morel-Forster, A.; Blumer, C.; Egger, B.; Luthi, M.; Schönborn, S.; and Vetter, T. 2018. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 75–82. IEEE.

Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18653–18664.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5764–5774.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. HeadNeRF: A Real-Time NeRF-Based Parametric Head Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20374–20384.

Hu, L.; Saito, S.; Wei, L.; Nagano, K.; Seo, J.; Fursund, J.; Sadeghi, I.; Sun, C.; Chen, Y.-C.; and Li, H. 2017. Avatar Digitization from a Single Image for Real-Time Rendering. *ACM Trans. Graph.*, 36(6).

Ichim, A. E.; Bouaziz, S.; and Pauly, M. 2015. Dynamic 3D Avatar Creation from Hand-Held Video Input. *ACM Trans. Graph.*, 34(4).

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023. NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. arXiv:2305.03027.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6): 194:1–194:17.

Li, T.; Slavcheva, M.; Zollhoefer, M.; Green, S.; Lassner, C.; Kim, C.; Schmidt, T.; Lovegrove, S.; Goesele, M.; Newcombe, R.; et al. 2022. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5521–5531.

Liu, X.; Xu, Y.; Wu, Q.; Zhou, H.; Wu, W.; and Zhou, B. 2022. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lombardi, S.; Saragih, J.; Simon, T.; and Sheikh, Y. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.*, 37(4): 68:1–68:13.

Lombardi, S.; Simon, T.; Saragih, J.; Schwartz, G.; Lehrmann, A.; and Sheikh, Y. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.*, 38(4): 65:1–65:14.

Lombardi, S.; Simon, T.; Schwartz, G.; Zollhoefer, M.; Sheikh, Y.; and Saragih, J. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.*, 40(4).

Ma, S.; Simon, T.; Saragih, J.; Wang, D.; Li, Y.; La Torre, F. D.; and Sheikh, Y. 2021. Pixel Codec Avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 64–73.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Nagano, K.; Seo, J.; Xing, J.; Wei, L.; Li, Z.; Saito, S.; Agarwal, A.; Fursund, J.; and Li, H. 2018. PaGAN: Real-Time Avatars Using Dynamic Textures. *ACM Trans. Graph.*, 37(6).

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021a. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5865–5874.

Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021b. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Sara Fridovich-Keil and Giacomo Meanti; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *CVPR*.

Sun, J.; Wang, X.; Shi, Y.; Wang, L.; Wang, J.; and Liu, Y. 2022. IDE-3D: Interactive Disentangled Editing for High-Resolution 3D-aware Portrait Synthesis. *ACM Transactions on Graphics (TOG)*, 41(6): 1–10.

Sun, J.; Wang, X.; Wang, L.; Li, X.; Zhang, Y.; Zhang, H.; and Liu, Y. 2023. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, D.; Chandran, P.; Zoss, G.; Bradley, D.; and Gotardo, P. 2022. MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393379.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS*.

Xu, Y.; Wang, L.; Zhao, X.; Zhang, H.; and Liu, Y. 2023. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zheng, Y.; Yifan, W.; Wetzstein, G.; Black, M. J.; and Hilliges, O. 2023. PointAvatar: Deformable Point-based Head Avatars from Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. MoFaNeRF: Morphable Facial Neural Radiance Field. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zielonka, W.; Bolkart, T.; and Thies, J. 2022. Towards Metrical Reconstruction of Human Faces.

Zielonka, W.; Bolkart, T.; and Thies, J. 2023. Instant Volumetric Head Avatars.

zllrunning. 2019. face-parsing.PyTorch. https://github.com/zllrunning/face-parsing.PyTorch. Accessed: 2023-04-01.