Implicit Neural Head Synthesis via Controllable Local Deformation Fields

Chuhan Chen[1*]   Matthew O'Toole[1]   Gaurav Bharaj[2]   Pablo Garrido[2]

[1]Carnegie Mellon University   [2]Flawless AI

Abstract

High-quality reconstruction of controllable 3D head avatars from 2D videos is highly desirable for virtual human applications in movies, games, and telepresence. Neural implicit fields provide a powerful representation to model 3D head avatars with personalized shape, expressions, and facial parts, e.g.*, hair and mouth interior, that go beyond the linear 3D morphable model (3DMM). However, existing methods do not model faces with fine-scale facial features, or local control of facial parts that extrapolate asymmetric expressions from monocular videos. Further, most condition only on 3DMM parameters with poor(er) locality, and resolve local features with a global neural field. We build on part-based implicit shape models that decompose a global deformation field into local ones. Our novel formulation models multiple implicit deformation fields with local semantic rig-like control via 3DMM-based parameters, and representative facial landmarks. Further, we propose a local control loss and attention mask mechanism that promote sparsity of each learned deformation field. Our formulation renders sharper locally controllable nonlinear deformations than previous implicit monocular approaches, especially mouth interior, asymmetric expressions, and facial details. Project page:* https://imaging.cs.cmu.edu/local_deformation_fields/

1. Introduction

Monocular human head avatar reconstruction is a long standing challenge that has drawn a lot of attention in the last few decades due to its wide application in movie making [11, 17, 43], and virtual reality [2, 3, 29], among others. Traditional reconstruction methods in production pipelines create animatable and detailed avatars, often represented as 3D rigs, from high-quality face scans with predefined expressions and poses [56, 58]. However, such data is often expensive to acquire and process, and over the years has created the need for an easier capture pipeline, *e.g.*, based on high-definition images, or videos of human subjects. With the advancements in deep learning, much effort has gone into learning neural 3D face representations from 2D images and the research community has achieved impressive results [22, 27, 39, 45]. However, modeling 3D structures from 2D information alone is an ill-posed problem, which results in models that lack view consistency and details.

Both traditional and neural reconstruction pipelines based on the parametric mesh representation, 3DMM [9], are efficient, controllable, and well integrated into the graphics pipeline, though at the expense of lacking important facial features such as hair, eyes, and mouth interior. In the last couple of years, there has been a surge of research on generalized implicit face representations, *e.g.*, sign distance functions (SDFs) [34], neural radiance fields (NeRFs) [16] or hybrid volumetric representations [4], that allow accurate modeling of fine-grained facial, and non-facial features not possible with mesh-based representations alone, while preserving view-consistency.

Recently, several implicit models for human head avatars from monocular videos have demonstrated great progress [1, 3, 12, 15, 18, 35, 36, 41, 47, 57, 67, 68]. Several employ facial parameters from an existing face tracker to condition a multi-layer perceptron (MLP) to model expression changes [3, 12, 18, 47], use deformation fields to learn a mapping from an observed point deformed by expression to a point on the template face [1, 35, 67], or learn forward mapping functions that estimate implicit 3DMM functions, represented as facial deformation bases [15, 42, 68]. These approaches have done a good job in allowing control of expressions and poses, even for out-of-distribution training parameters. However, none of these approaches reconstruct animatable heads with high-fidelity details such as deep creases. Besides, since they heavily rely on creating implicit fields derived from linear 3DMMs, which are de-facto limited by global or large-scale expression decompositions, it is relatively difficult to control localized deformations at a finer level, *e.g.*, wrinkles that form around eyes when winking.

In this paper, we propose a novel approach that allows implicit 3D rig modeling, and local control of detailed facial deformations. We use expression parameters obtained from a 3DMM face tracker, *e.g.*, DECA [10], but learn to model more *local* deformations beyond linear 3DMM. To this end, instead of learning a single global deformation field, we

---

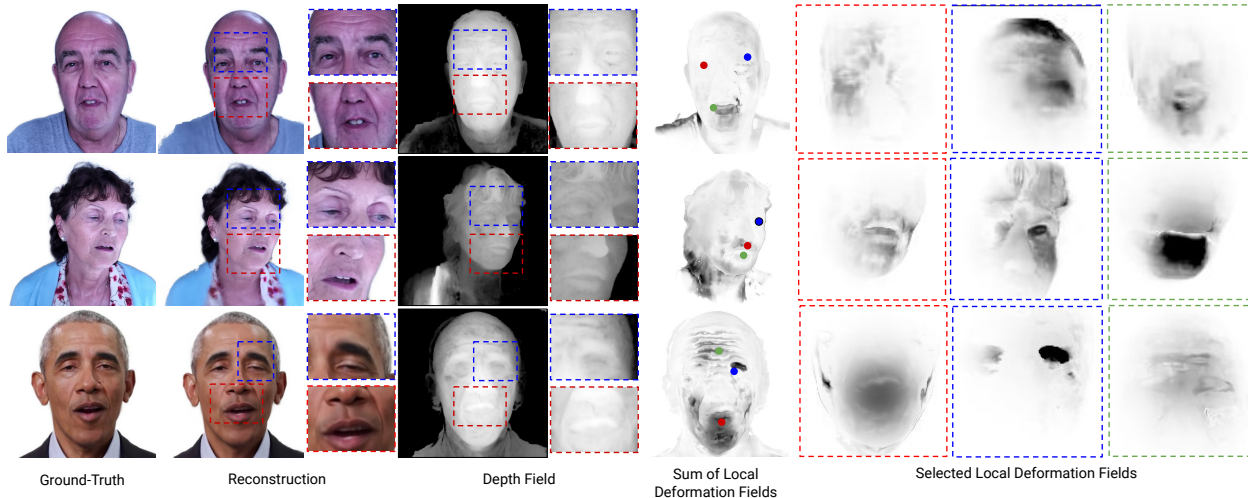*Work was done while interning at Flawless AI.

Figure 1. **Main results on test samples.** Our method models dynamic 3D deformations as an ensemble of local deformation fields, centered around 3D facial landmarks, shown as red, blue, and green dots in this example (details in Sec. 3). Our formulation synthesizes 3D neural heads from 2D videos with fine-grained geometric details, as shown in column 3 (depth field).

break it into multiple local fields with varying spatial support, triggered by sparse facial landmarks, with weak supervision on the 3DMM expressions. The local deformation fields are represented by nonlinear function within a certain radius, and conditioned by tracked expression parameters weighted by an attention mask that filters redundant parameters that do not influence the landmarks. Finally, the local deformations are summed with distance-based weights, which are used to deform the global point to the canonical space, and retrieve radiance and density for volumetric rendering.

While part-based field decomposition [46,63] approaches have been proposed, we demonstrate that decomposing implicit deformation fields into local fields improves representation capacity to model facial details. By filtering redundant input expression parameters to each local field and providing weak supervision via 3DMM, we achieve better, detailed local control, and modelling of asymmetric expressions (see Fig. 1). We provide qualitative and quantitative comparisons with state-of-the-art monocular head avatar synthesis methods and show that our approach reconstructs facial details more accurately, while improving local control of the face. In summary, our contributions are as follows:

1. A novel formulation that models local field deformation for implicit NeRF face rigs that provides fine-scale control with landmarks.

2. Local deformation fields surpass linear 3DMM's representation capacity via a novel local field control loss, and attention masks filtering.

3. We demonstrate the advantage of our approach in different applications, including detailed asymmetric expression control, and high-quality head synthesis with sharp mouth interior.

2. Related Work

The literature on detailed dynamic human avatars is vast. Thus, we mainly review methods for reconstructing controllable head models from unstructured 2D data using 3D aware neural rendering approaches. Please refer to Egger *et al*. [9] for a comprehensive overview on 3DMM and applications and to Tewari *et al*. [51, 52] for advances in neural rendering.

Neural Avatar Reconstruction and Portrait Synthesis
Advances in neural rendering have pushed the boundaries of human avatar digitization from unconstraint 2D images or video. An early trend are model-based neural reconstruction approaches that learn to regress detailed 3DMMs [39, 44, 49, 54] or detailed rendering layers of a 3D model [14,23,32,53,64] using autoencoders or GANs to reconstruct high-quality 3D faces with photo-realistic appearance. However, model-based approaches often generate avatars with coarse local control derived from 3DMMs and restricted to the inner face region, *i.e.*, they lack eyes, mouth interior, ears, and hair. Another line of research attempts to synthesize face portraits using controllable GAN-based generative approaches [20,30,59,65,71], driven either via sparse keypoints [30,59,65], dense mesh priors [6,20,50], or multi-modal input [62,71]. These methods can synthesize photo-realistic 2D face portraits, but struggle with large poses and cannot generate detailed localized deformations, *e.g.*, around mouth and eyes. The work by Lombardi *et al*. [28] and Ma *et al*. [29] can generate 3D avatars with very detailed facial features, but local facial control is non-intuitive and reconstructions require expensive multiview camera systems. Our approach reconstructs 3D avatars with detailed local control from unconstrained 2D videos.

Deformable Neural 3D Consistent Representations
Modeling 3D-aware objects especially heads has been an

active research in the last few years. Some approaches learn detailed 3D consistent implicit head models from a large corpus of unconstrained 2D images via 3D neural implicit surfaces [34, 40], 3D-aware generative NeRFs [5, 8, 16] or hybrid volumetric representations [4, 60]. NeRFs have particularly earned popularity since they can reconstruct complex scene structures while preserving fine scale details. Some recent NeRF-based approaches can model general dynamic objects, such as bodies and heads in general video scenes using dynamic embeddings and often by warping observed points into a canonical frame configuration [25, 35, 36, 38, 55]. These approaches can reconstruct fine grained avatars from novel viewpoints but offer no semantic control over the generated models. Our approach on the contrary can learn detailed 3D consistent representations with local facial control.

Controllable Neural Implicit Head Generation Although implicit deformable head models have showed great promise, most methods lack semantic control over the learned implicit fields. A recent line of work create animatable implicit blendshape rigs with semantic latent representations such as landmarks and expression parameters to enable downstream animation and editing tasks [1, 3, 12, 13, 15, 18, 47, 67, 68]. Some approaches learn deformation fields using dense mesh priors [1, 13, 68], sparse landmark priors [67] or hard mesh constraints [15, 42, 47]. Among them, [1, 68] model pose aware deformation fields to enable full head control with better generalization capabilities at the expense of smoother reconstructions and coarse local control. Finer local control is achieved using multiple local deformation fields, either rigged via sparse landmarks [67] or blendshape weights [13]. However, these methods can only reconstruct mid-scale dynamic details, *i.e.*, no wrinkles or sharp mouth interior. Hong *et al.* [18] boost rendering quality by integrating a coarse-to-fine 2D neural rendering module with NeRF-based rendering. However, none these approaches can generate fine-grained facial deformations with sparse local control. As far as we are aware, only implicit models derived from multiview imagery [3, 57] can produce crisp results. Cao *et al.* [3] partially achieve personalized fine-grained control by fine tuning a universal implicit morphable model on user-specific custom expressions. As in Gao *et al.* [13], local control granularity is limited to in-distribution expressions. On the contrary, we propose an implicit rig representation that learns to disentangle detailed deformation fields from arbitrary expressions and head poses while achieving good generalization power.

3. Our Method

In this section, we describe our method that enables detailed reconstruction of facial expressions with the ability to perform local control, as shown in Fig. 2. We propose a deformable NeRF approach conditioned on expression pa-

rameters from a 3D face tracker, which learns translation from each point in the observed space to the canonical space, as defined by a 3DMM, *e.g.*, FLAME [24]. To improve the ability to represent high-frequency change in the deformation field and allow local control, we break down the deformation field into multiple local fields conditioned on expression filtered by an attention mask. The locality is further enforced through a local control loss for each local field and the sum of the deformations from all local fields are weakly supervised by a mesh prior.

3.1. Deformable Neural Radiance Field for Faces

A neural radiance field is defined as a continuous function

$$F : (x, d) \rightarrow (\sigma(x), c(x, d)) \tag{1}$$

that maps view direction d and per-point location in the observed space x into per-point density σ and radiance c. Every pixel of the rendered scene is the accumulation of per-point density and radiance on each ray, cast through the pixel using the volumetric rendering equation [31].

To model dynamic scenes, we represent a continuous transformation from a point x_{obs} in the observed space to the corresponding point x_{can} in the canonical space [35] as

$$T = D(\gamma(x_{obs}), \omega_i), \tag{2}$$

where ω_i is a per-frame deformation code that learns the scene at frame i, γ are sinusoidal positional encodings of different frequencies [48], and D represents a neural deformation field capable of modeling high-frequency variations. A canonical point x_{can} can then be obtained as follows:

$$x_{can} = T(x_{obs}) . \tag{3}$$

Inspired by [1, 68], we model human face deformation in a monocular video setup instead of a per-frame deformation code. We condition the deformation field using a per-frame expression e_i and pose p_i code generated from an existing tracker as well as additional latent codes to account for pose inaccuracies ω_i^p and appearance inconsistencies ω_i^a, so that

$$T = D(\gamma(x_{obs}), e_i, p_i, \omega_i^p) \tag{4}$$
$$\sigma, c = F(x_{can}, d, \omega_i^a) \tag{5}$$

Here, T can be in $SE(3)$ to represent a 3D rigid deformation [35, 36] or in $\mathbb{R}^3$ to model a non-rigid deformation [26, 38]. Since facial expression deformations are commonly non-rigid, we predict T as translation $t \in \mathbb{R}^3$, and thus map x_{obs} into x_{can} as follows:

$$x_{can} = t + x_{obs} . \tag{6}$$

While this disentanglement of deformation and canonical space provides better generalization to in- and out-of-distribution deformations, the additive model still struggles
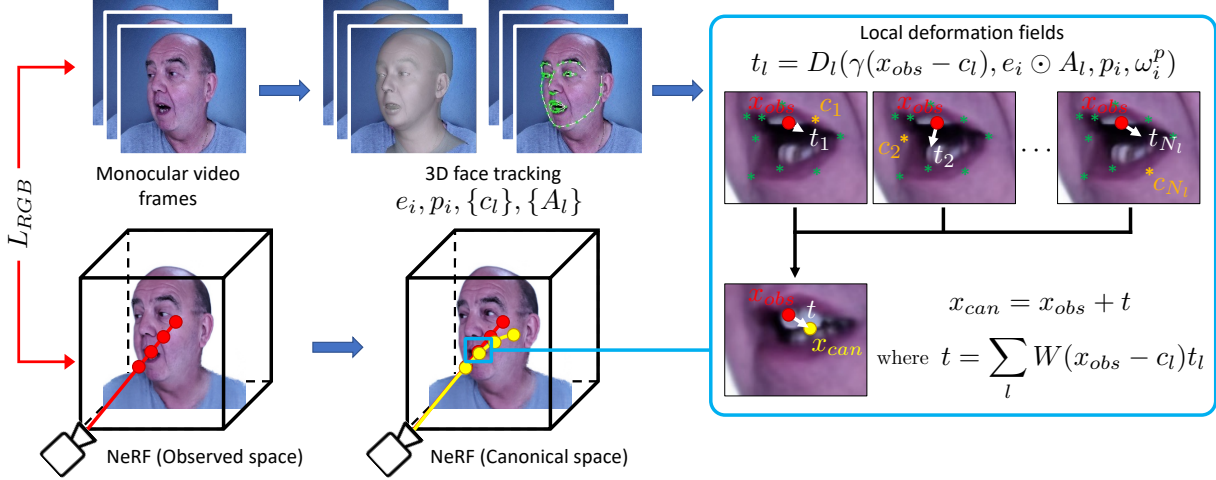
Figure 2. **Method overview.** Given an input video sequence, we run a face tracker [10] to get at each frame i the following parameters of a linear 3DMM [24]: expression e_i, pose p_i, and sparse 3D facial landmarks c_l. An attention mask with local spatial support is also pre-computed from the 3DMM. We model dynamic deformations as a translation of each observed point x_{obs} to the canonical space, represented as x_{can} (Sec. 3.1). We decompose the global deformation field t into multiple local fields $\{t_l\}$, each centered around representative landmarks c_l (Sec. 3.2). We enforce sparsity of each field t_l via an attention mask A_l that modulates e_i (Sec. 3.3). Our implicit representation is learned using RGB information, geometric regularization and priors, and a novel local control loss (Sec. 3.4).

to represent details such as deep skin creases, crowfeet around eyes, and facial hair, especially when no ground truth mesh is available. The main reasons are that (i) a global model often lack expressibility [67] and (ii) the conditioning with 3DMM expression parameters represent linear and non-sparse local deformations (Sec. 5).

3.2. Local Deformation Fields

Inspired by advances in part-based implicit rigging [33, 37, 67, 70], we overcome limitations of previous work by decomposing the global deformation field into multiple local fields, each centered around a pre-defined facial landmark location, to model non-linear local expression deformations with higher level of details, as shown in Fig. 2.

Mathematically, we first define a set of N_l landmarks on the face mesh and a set of local deformation fields D_l, $\forall l \in [1, ..., N_l]$. We denote c_l to be the 3D position of the lth landmark on the deformed mesh. To map a point x_{obs} from the observed space to its corresponding point in the canonical space, we first compute the position of this point relative to a landmark location (i.e., $x_l = x_{obs} - c_l$) and pass this to the local deformation field $D_l(\gamma(x_l))$. We repeat this process for all landmarks, and use a weighted average to compute the deformation:

$$t = \sum_l W(x_l)t_l , \qquad (7)$$

$$\text{where} \quad t_l = D_l(\gamma(x_l), e_i, p_i, \omega_i^p) . \qquad (8)$$

Similar to [70], we define the weight $W(x)$ as a scaled

Gaussian function with zero mean and R standard deviation based on the point's distance to each local center, i.e.,

$$W(x) = \max \left(\exp \left(\frac{-\|x\|_2^2}{2R^2} \right) - \tau, 0 \right) s , \qquad (9)$$

where R is a user-defined parameter representing the spatial support of t_l, $\forall l$. The threshold τ prevents points farther away from each landmark to query the local neural network, improving computing efficiency. The scale $s < 1$ adjusts the magnitude of the local fields' outputs to the range of valid values for facial deformation.

3.3. Attention Mask Based on 3DMM Bases

Although the expression bases offer global spatial support, some parameters have little effect in certain local regions. As such, feeding the same expression code into each local field is redundant. We can prune these parameters by leveraging an attention mask with local spatial support to help enforce sparsity of each local field, i.e.,

$$t_l = D_l(\gamma(x_l), e_i \circ A_l, p_i, \omega_i^p) , \qquad (10)$$

where $A_l \in [0, 1]^{|e|}$ is a binary attention mask associated with the expression parameters e of each local deformation D_l and $\circ$ is the element-wise product.

We precompute A_l, $\forall l$ in a one-time preprocessing step by exploiting the spatial support of expression bases at each landmark location l of the 3DMM. We first compute a distance matrix $\mathcal{D} \in \mathbb{R}^{N_l \times |e|}$ containing absolute delta displacements of each expression basis activated independently over

all landmark locations. Each column $\mathcal{D}_{*,i}$ represents the spatial support of the i-th expression basis over all N_l landmarks. For each column $\mathcal{D}_{*,i}$ we zero out rows where the spatial support of landmark locations fall below the 25% quantile. We then binarize $\mathcal{D}$, yielding a binary cross attention mask A.

3.4. Training Objectives

RGB Loss   Our method trains in a self-supervised manner by comparing a rendered and ground truth RGB pixel

$$L_{RGB} = \frac{1}{|P|} \sum_p \|\hat{I}_p - I_p\|_2 \,, \qquad (11)$$

where P is the set of sampled training pixels, and $\hat{I}_p$ and I_p denote rendered and measured pixel values, respectively. We use an $\ell_{2,1}$-norm over the sampled pixel locations, akin to [53], to allow sharper per-pixel reconstruction while maintaining the same convergence speed.

Local Control Loss   We propose a novel loss function to enforce locality of each local field. The 3D facial landmarks share the same semantic face location for all poses and expressions. Therefore, the same 3D facial landmark across all frames should map to the same position in the canonical space. Based on this insight, we enforce that all rays cast towards a facial landmark in the observed space intersect the face geometry at same point in the canonical space.

Mathematically, given any two frames and corresponding latent codes $\delta_1 = (p_1, e_1)$ and $\delta_2 = (p_2, e_2)$, for each keypoint c_l, we then cast a ray based on the pose of each of the 2 images to the deformed keypoint location c_{l_1} and c_{l_2}.

Next, we solve a volume rendering integral to compute depth associated with the rays, by (i) sampling points along each ray, (ii) querying the local deformation field at these points, and (iii) and evaluating the density at the corresponding points in the canonical space. This depth determines the 3D points $\hat{x}_{1,l}$ and $\hat{x}_{2,l}$ in deformed space where the rays and face geometry intersect. (iv) We then deform the intersection points to the canonical space and enforce the estimated surface intersection should be the same in the canonical space, achieving semantic local control implicitly via intersection equivariance constraints, with respect to the ℓ_1-norm as follows:

$$L_{local} = \sum_l \|\hat{x}_{1,l} - \hat{x}_{2,l}\|_1 \qquad (12)$$

This loss helps increase sparsity of the local deformation field, hence allowing for better locality of expression control, as well as enabling sharper and more consistent local deformations, as shown in Sec. 5. We also observed that using L_1 loss is sufficient to remove effect of the outliers from occluded 2D keypoints.

Weak Mesh Prior   We also leverage prior geometric knowledge from the tracked 3D mesh. Specifically, for each sampled point x along the ray we first query the closest surface point x_m on the deformed mesh M to get a pseudo ground truth mesh constraints. We then supervise the predicted deformation $t(x)$ with pseudo ground truth labels for salient face regions as follows:

$$L_{mesh} = \frac{1}{|X|} \sum_x F_{surf}(x) \cdot \|t(x) - t_{mesh}(x_m)\|_2 \,, \quad (13)$$

where X is the set of sampled training points and $t_{mesh}(x_m)$ is the pseudo ground truth deformation obtained by first finding the vertex deformation of the 3 vertices of the triangle on which x_m is and then interpolating at x_m via barycentric coordinates. $F_{surf}(x)$ is a binary mask that contains non-zero entries for sampled points x that (i) represent foreground pixels as defined by a segmentation mask and (ii) are visible surface face regions, *i.e.*, rendering weight (per-point alpha * transmittance) is above a user-defined threshold $\tau = 10^{-4}$.

Deformation Field Regularization   We penalize large deformations of the reconstructed field t by enforcing ℓ_2-norm regularization of per-point deformations

$$L_{def} = \frac{1}{|X|} \sum_x lut(x) \cdot \|t(x)\|_2 \,, \qquad (14)$$

where $lut(x)$ is a look-up table that assigns a higher weight to points sampled from background pixels.

Volume and Latent Codes Regularization   Following the work in [7], we impose ℓ_1 sparsity loss L_{vol} on volume density. Furthermore, we penalize large latent code variations $\omega = (\omega^p, \omega^a)$ to avoid per-frame overfitting. Specifically, we use $L_{code} = \|\omega\|_2$, akin to [12, 35].

Overall, our final loss is represented as follows:

$$\begin{aligned} L \quad = \quad & L_{RGB} + \lambda_{local}L_{local} + \lambda_{mesh}L_{mesh} + \\ & \lambda_{def}L_{def} + \lambda_{vol}L_{vol} + \lambda_{code}L_{code} \qquad (15) \end{aligned}$$

3.5. Implementation Details

Our framework models the canonical space with TensoRF [7]. We use Adam optimizer [21] with a learning rate of $1e^{-3}$ decayed to $1e^{-4}$ at end of training. We pre-train on each sequence without deformation field and latent codes for 14k iterations to learn a rough estimate of the canonical space. We use $N_l = 34, R = 0.03, \tau = 1e-4, s = 0.02$, and 10 positional encoding frequencies linear in log space for each local field and we increase the weight on positional encoding frequencies [35] over 40k iterations. We utilize a ray batch size of 2048 and train for 400k iterations. The following weights are used in our training objective: $\lambda_{local} = 0.01$, $\lambda_{mesh} = 0.01, \lambda_{field} = 0.5, \lambda_{vol} = 2e^{-4}$ at start of training and change to $6e^{-4}$ after the first upsampling of the

Figure 3. **Novel pose and expression synthesis via face reenactment**: We track the user's expressions (top) via DECA and transfer the 3DMM parameters (middle) to the neural head model of *Subject2*. The model produces asymmetric expressions under the user's pose with a high level of details (bottom) that surpasses the linear 3DMM. Note that none of the transferred expressions were in the training set of *Subject2*.

voxel grid, and $\lambda_{code} = 0.01$. We use $lut(x) = \{1.0, 100.0\}$ for penalizing foreground and background sampled points, respectively. Dimension of w^p and w^a are 32 and 16 respectively. The final deformation from each local field is sum of the output of a MLP and deformation of the centroid c_l. All the hyperparameter values were found empirically and used in all our experiments. We apply the tracker DECA [10] to compute expression and pose parameters, 3D landmarks, and tracked meshes for each sequence. This data is used for conditioning our deformation fields, finding the deformation field centroids and intersection points, and weakly supervising deformation fields, respectively. To run TensoRF with tracked results from DECA, we first obtain the global translation of the head from the camera described in [68] and convert product of global translation and rigid head pose $[R|t]$ into camera extrinsics. We generate the face segmentation mask with FaceParsing [69] and mask out background with alpha matte from MODnet [19].

4. Results

In this section, we show the effectiveness of our local deformation fields in synthesizing novel expressions and poses on real videos. We conduct quantitative and qualitative comparisons with three state-of-the art baselines on implicit neural synthesis, and show a facial reenactment application for unseen asymmetric expressions.

4.1. Datasets

We evaluate our method on monocular video sequences of four real subjects, as shown in Fig. 4. We divide each video sequence into 6.5k images for training and 1.5k–2k images for testing, except for *Subject4* sequence, which is split into 2.7k and 1.8k training and testing images, respectively. We crop all frames around the face center and run our method at 500×500 resolution, except for *Subject4* sequence, which we run at 256×256 to match IM Avatar's resolution [68].

Please refer to the supplementary material for a detailed description of our dataset as well as a visualization of pose and expression distribution.

4.2. Pose and Expression Synthesis and Control

Fig. 1 shows our reconstruction results on test sequences of *Subject1* and *Subject2*. Overall, our approach can render heads with detailed mouth interior, eyes and skin that looks almost indistinguishable from the ground truth. As seen in column 3, the depth field reveals very detailed geometric esp. for *Subject2* not demonstrated before by previous approaches, as shown in the supp. material. In addition, we show for *Subject3*, we can perform local expression control by injecting a different expression code to a particular local field, in this case the one around the right eye, allowing the right eye to wink.

Fig. 3 illustrates reenactment results for *Subject2* to demonstrate that our approach renders high-quality avatars with localized expression control and crisp details in the eyes and mouth interior, even for highly non-symmetrical expressions, which are also out-of-distribution. Note that *Subject2* esp. column 2 can reproduce non-linear deformations as performed by the user that go beyond the representation power of the underlying 3DMM.

4.3. Qualitative Comparisons

We compare our approach with three implicit neural head synthesis baselines: IM Avatar(-)[1] [68], NerFace [12], and RigNeRF*[2] [1]. We train the baseline approaches with default settings on our training set and evaluate on every 10th frame of the test set. Since in RigNeRF* and our approach, the deformation field is conditioned on a per-frame deformation latent code, to evaluate against the testset, we first optimize for the deformation latent code on each test frame with

---

[1] IM Avatar (-) refers to the version trained without FLAME supervision.
[2] We reimplemented RigNeRF. Details are in supplementary material.

Figure 4. **Qualitative comparisons with state-of-the-art on test data.** Top to bottom: GT, IM Avatar(-), RigNeRF*, NerFace, and our approach. Overall our approach synthesizes fine-scale skin details e.g wrinkles and moles, and facial attributes at higher fidelity. Note that IM Avatar(-) is a version that synthesizes sharper details at the expense of poorer pose and expression extrapolation.

RGB loss while fixing all other parameters for our model for 10k iterations and RigNeRF* for 40k iterations. For NerFace and IM Avatar, we use the latent code of the first training frame. For reenactment, we use both the deformation latent code and appearance latent code of the first training frame. As discussed in [12], per-frame latent codes are used to adjust for small inaccuracies in pose and expressions for better convergence. See supplementary material, Tab. 4 and Fig. 8, for comparisons without test-time latent code optimization. Fig. 4 shows qualitative comparisons with state-of-the-art approaches on our test dataset. Our approach synthesizes fine-scale skin details at higher fidelity and overall attains sharper reconstructions for the eyes, mouth interior, and hair than previous approaches. Note that wrinkles, moles and teeth are reproduced by our approach at a level of detail not seen before. On *Subject4* sequence, state-of-the-art methods

are on par with our approach, stemming from the low image resolution of the dataset.

In our experiments, we observed that IM Avatar full model, *i.e.*, the model with FLAME supervision, produced overly smoothed renderings, making comparisons unfair. For this reason, we trained IM Avatar(-) with RGB and mask loss only to boost sharpness. While it may result in worse ability in generalization to novel pose and expression, it is not the purpose of our test.

4.4. Quantitative Comparisons

To assess the quality of the rendered avatars, we adopt the metrics by Gafni *et al.* [12]. Specifically, we measure the Manhattan distance (ℓ_1) in the RGB color space, PSNR, LPIPS [66], and SSIM [61]. Tab. 1 shows that our approach generates images with higher fidelity on *Subject1*, *Subject2*, and *Subject3*, and that outperforms state-of-the-art methods on all sequences based on LPIPS metric. However, ℓ_1, SSIM and PSNR metrics degrade on *Subject4*. This dataset is particularly challenging as test samples are mainly out of distribution. Unlike IM Avatar(-) and RigNeRF*, our approach trades quality more than robustness to unseen data. Still, our approach generates perceptually better renderings.

5. Analysis

We ablate on the use of attention mask, local control loss, and local fields vs global field and demonstrate qualitatively and quantitatively that each component assist in rendering unseen asymmetric expressions and reconstruct details.

Influence of Attention Mask Attention masks help prune the effect from expression bases that have small influence on a local region and enforces sparsity of the local fields, while retaining the representation power needed for face regions with large deformation such as the mouth. Fig. 5, column 2 shows that without the attention mask, multiple local fields could be learned to contribute to the deformation caused by eye opening and closing. Since the training set only includes expressions where two eyes open and close at the same time, influence of these local fields could not be fully disentangled, especially due to the global control of the expression codes.

Local Control Loss Local control loss further improves locality by encouraging each local field to produce consistent and independent influence over different pose and expressions. As demonstrated by comparing Fig. 5 columns 1 and 3, adding local control loss allows independent control of local fields, while also improves convergence of the deformation fields which allows more accurate modelling of details such as the teeth.

Global vs Local Fields We also examine the influence of using multiple local fields instead of a global field. We represent the deformation field with an MLP with roughly

	Subject1				Subject2				Subject3				Subject4 [68]			
	$\ell_1\downarrow$	SSIM ↑	LPIPS↓	PSNR ↑	$\ell_1\downarrow$	SSIM ↑	LPIPS↓	PSNR ↑	$\ell_1\downarrow$	SSIM ↑	LPIPS↓	PSNR ↑	$\ell_1\downarrow\downarrow$	SSIM ↑	LPIPS↓	PSNR ↑
Nerface	0.058	0.903	0.105	21.457	0.077	0.889	0.121	18.252	0.047	0.894	0.055	22.220	0.077	0.818	0.085	17.910
IM Avatar(-)	0.068	0.901	0.113	20.502	0.093	0.877	0.157	14.960	0.043	0.900	0.078	23.218	**0.063**	**0.870**	0.069	19.215
RigNeRF*	**0.055**	0.904	0.095	22.324	0.072	0.884	0.120	18.922	0.035	0.910	0.052	24.634	0.065	0.844	0.063	**19.253**
Ours	0.056	**0.934**	**0.0465**	**23.656**	**0.062**	**0.917**	**0.0576**	**20.4375**	**0.0206**	**0.971**	**0.0265**	**30.854**	0.081	0.830	**0.062**	19.085

Table 1. **Quantitative comparisons with state-of-the-art on test data.** Our approach achieves state of the art performance using LPIPS metric and generates better metrics on *Subject1*, *Subject2*, and *Subject3*. RigNeRF* and especially IM Avatar(-) generalize better to *Subject4*.
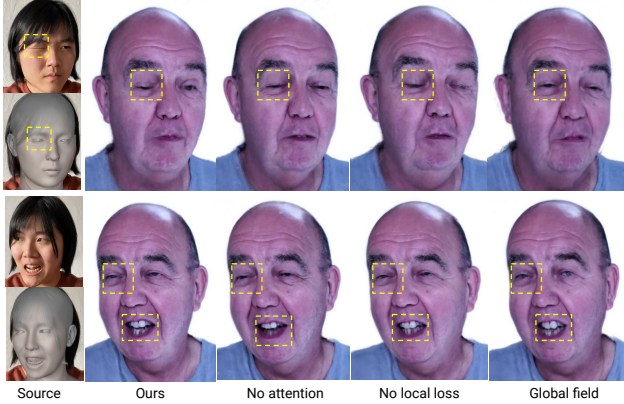

Figure 5. **Ablation tests on *Subject1* sequence.** Expression and pose reenactment with our full model, our model without attention mask, without Local Control Loss, and using global deformation field instead of local deformation fields.

Variants	Metrics			
	$\ell_1\downarrow$	SSIM ↑	LPIPS ↓	PSNR ↑
Global Field	0.0656	0.930	0.0552	22.467
No Attention mask	0.0582	0.932	0.0487	23.309
No Local	0.0577	**0.934**	0.0486	23.421
Ours	**0.0558**	**0.934**	**0.0465**	**23.656**

Table 2. **Ablation tests on *Subject1* sequence.** Image and perceptual metrics with respect to our full model, our model without attention mask, without Local Control Loss, and using global deformation field instead of local deformation fields.

the same number of parameters as the sum of parameters from all local fields, 10 number of positional encoding frequencies and supervised with mesh prior on top of RGB loss. Fig. 5, column 4 shows that while a global field with high frequencies can represent some details, it does not have sufficient representation power to model complex deformations such as mouth opening, resulting in defective teeth details. It cannot model unseen asymmetric expressions beyond the global 3DMM bases. We also evaluated each experiment against the ground truth images in the test set for *Subject1*, and computed metrics described in Sec. 4.4, listed in Tab. 2. Using all of the above mentioned components improves the reconstruct quality quantitatively.

Limitations. Our approach is designed to model fine-scale facial details. While, we can handle a varying set of pose and expressions with fine details, the reconstruction quality degrades on extreme pose and expression variations Fig. 6.

Figure 6. **Limitations.** Defects with extreme poses and expressions, and shoulder when the dataset contains big shoulder movement.

Moreover, pose and expression are not always fully disentangled in our *in-the-wild* datasets, thus, improved generalization due to extreme pose and expression variation is left as future work. Further, non-facial part of humans *e.g.* shoulders, are currently not explicitly modeled in our formulation and are not exposed to local fields and leads to noisy and sometimes blurry torso renderings.

6. Conclusion

We present an approach to model face rigs with neural radiance fields that go beyond linear 3DMM and global deformations, and model nonlinear local deformation and fine-grained facial details. A novel local control loss helps enforce the locality and consistency of local deformation fields, which are controlled via a sparse landmark set and weakly supervised via a 3DMM. Further, we introduce an attention mask to filter redundant expression parameters that have small influence on particular face regions, leading to more accurate local deformation fields, the subsequent expressions and detail reconstruction. We attain state-of-the-art statistics on perceptual radiance image quality and demonstrated that our approach is able to reconstruct controllable local details. We demonstrate that it leads to several applications, such as detail preserving mouth interior articulation, asymmetric lips and cheek movement, and eye winking, among others. Additionally, we demonstrate the limitations of our approach to in-distribution poses and expressions, and generalization improvements will be explored in the future.

7. Acknowledgments

References

[1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shecht-man, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. *CoRR*, abs/2206.06481, 2022. 1, 3, 6

[2] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason M. Saragih. Deep relightable appearance models for animatable faces. *ACM TOG*, 40(4):89:1–89:15, 2021. 1

[3] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason M. Saragih. Authentic volumetric avatars from a phone scan. *ACM TOG*, 41(4):163:1–163:19, 2022. 1, 3

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16102–16112. IEEE, 2022. 1, 3

[5] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809. Computer Vision Foundation / IEEE Computer Society, 2021. 3

[6] Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus H. Gross, Paulo F. U. Gotardo, and Derek Bradley. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM TOG*, 40(6):223:1–223:14, 2021. 2

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *CoRR*, abs/2203.09517, 2022. 5

[8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: generative radiance manifolds for 3d-aware image generation. *CoRR*, abs/2112.08867, 2021. 3

[9] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present, and future. *ACM TOG*, 39(5):157:1–157:38, 2020. 1, 2

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM TOG*, 40(4):88:1–88:13, 2021. 1, 4, 6

[11] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul E. Debevec. Driving high-resolution facial scans with video performance capture. *ACM TOG*, 34(1):8:1–8:14, 2014. 1

[12] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, pages 8649–8658, 2021. 1, 3, 5, 6, 7

[13] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM TOG*, 41(6), 2022. 3

[14] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, pages 1155–1164. Computer Vision Foundation / IEEE Computer Society, 2019. 2

[15] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. *CoRR*, abs/2112.01554, 2021. 1, 3

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*. OpenReview.net, 2022. 1, 3

[17] Kaiwen Guo, Peter Lincoln, Philip L. Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. The relightables: volumetric performance capture of humans with realistic relighting. *ACM TOG*, 38(6):217:1–217:19, 2019. 1

[18] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *CoRR*, abs/2112.05637, 2021. 1, 3

[19] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W. H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, pages 1140–1147. AAAI Press, 2022. 6

[20] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37(4):163, 2018. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[22] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *CVPR*, pages 757–766. Computer Vision Foundation / IEEE Computer Society, 2020. 1

[23] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme++: Facial shape and BRDF inference with photorealistic rendering-aware gans. *CoRR*, abs/2112.05957, 2021. 2

[24] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36(6):194:1–194:17, 2017. 3, 4

[25] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, pages 5511–5521. IEEE, 2022. 3

[26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, pages 6498–6508. Computer Vision Foundation / IEEE Computer Society, 2021. 3

[27] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *CVPR*, pages 5890–5899. Computer Vision Foundation / IEEE Computer Society, 2020. 1

[28] Stephen Lombardi, Jason M. Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM TOG*, 37(4):68, 2018. 2

[29] Shugao Ma, Tomas Simon, Jason M. Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. In *CVPR*, pages 64–73. Computer Vision Foundation / IEEE Computer Society, 2021. 1, 2

[30] Moustafa Meshry, Saksham Suri, Larry S. Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *ICCV*, pages 13809–13818. IEEE, 2021. 2

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022. 3

[32] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM TOG*, 37(6):258, 2018. 2

[33] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pages 5742–5752. IEEE, 2021. 4

[34] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13493–13503. IEEE, 2022. 1, 3

[35] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5845–5854. IEEE, 2021. 1, 3, 5

[36] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG*, 40(6):238:1–238:12, 2021. 1, 3

[37] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14294–14303. IEEE, 2021. 4

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327. Computer Vision Foundation / IEEE Computer Society, 2021. 3

[39] Mallikarjun B. R., Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Learning complete 3d morphable face models from images and videos. In *CVPR*, pages 3361–3371. Computer Vision Foundation / IEEE Computer Society, 2021. 1, 2

[40] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia Giraldez, Xavier Giró-i-Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *ICCV*, pages 5600–5609. IEEE, 2021. 3

[41] Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, Xiaokang Yang, and Stefanos Zafeiriou. Facial geometric detail recovery via implicit representation. *CoRR*, abs/2203.09692, 2022. 1

[42] Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, Xiaokang Yang, and Stefanos Zafeiriou. Facial geometric detail recovery via implicit representation. *CoRR*, abs/2203.09692, 2022. 1, 3

[43] Jérémy Riviere, Paulo F. U. Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. *ACM TOG*, 39(4):81, 2020. 1

[44] Gil Shamai, Ron Slossberg, and Ron Kimmel. Synthesizing facial photometries and corresponding geometries using generative adversarial networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(3s), 2019. 2

[45] Sahil Sharma and Vijay Kumar. 3d face reconstruction in deep learning era: A survey. *Archives of Computational Methods in Engineering*, 29(5):3475–3507, 2022. 1

[46] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Hsin-Ying Lee, Jian Ren, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. *arXiv preprint arXiv:2301.11326*, 2023. 2

[47] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *CoRR*, abs/2206.08361, 2022. 1, 3

[48] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, page 7537–7547, 2020. 3

[49] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. FML: face model learning from videos. In *CVPR*, pages 10812–10822. Computer Vision Foundation / IEEE Computer Society, 2019. 2

[50] Ayush Tewari, Mohamed Elgharib, Mallikarjun B. R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. PIE: portrait image embedding for semantic control. *ACM TOG*, 39(6):223:1–223:14, 2020. 2

[51] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason M. Saragih, Matthias Nießner, Rohit Pandey, Sean Ryan Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. State of the art on neural rendering. *Comput. Graph. Forum*, 39(2):701–727, 2020. 2

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

[53] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian

Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, pages 2549–2559. Computer Vision Foundation / IEEE Computer Society, 2018. 2, 5

[54] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, pages 1126–1135. Computer Vision Foundation / IEEE Computer Society, 2019. 2

[55] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, pages 12939–12950. IEEE, 2021. 3

[56] Javier von der Pahlen, Jorge Jimenez, Etienne Danvoye, Paul Debevec, Graham Fyffe, and Oleg Alexander. Digital Ira and beyond: Creating photoreal real-time digital characters. In *SIGGRAPH '14 ACM SIGGRAPH 2014 Courses*, pages 1–384. ACM Press, 2014. 1

[57] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo F. U. Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *SIGGRAPH*, pages 55:1–55:9. ACM, 2022. 1, 3

[58] Mengjiao Wang, Derek Bradley, Stefanos Zafeiriou, and Thabo Beeler. Facial expression synthesis using a global-local multilinear framework. *Comput. Graph. Forum*, 39(2):235–245, 2020. 1

[59] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049. Computer Vision Foundation / IEEE Computer Society, 2021. 2

[60] Ziyan Wang, Timur M. Bagautdinov, Stephen Lombardi, Tomas Simon, Jason M. Saragih, Jessica K. Hodgins, and Michael Zollhöfer. Learning compositional radiance fields of dynamic human heads. In *CVPR*, pages 5704–5713. Computer Vision Foundation / IEEE Computer Society, 2021. 3

[61] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 7

[62] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE TVCG*, 26(12):3457–3466, 2020. 2

[63] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2

[64] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM TOG*, 37(4):162, 2018. 2

[65] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, pages 9458–9467. IEEE, 2019. 2

[66] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[67] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. *CoRR*, abs/2203.14510, 2022. 1, 3, 4

[68] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Bühler, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. *CoRR*, abs/2112.07471, 2021. 1, 3, 6, 8

[69] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pages 18676–18688. IEEE, 2022. 6

[70] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. *CoRR*, abs/2203.14478, 2022. 4

[71] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186. Computer Vision Foundation / IEEE Computer Society, 2021. 2