Instant Volumetric Head Avatars

Wojciech Zielonka Timo Bolkart Justus Thies
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{wojciech.zielonka, timo.bolkart, justus.thies}@tuebingen.mpg.de

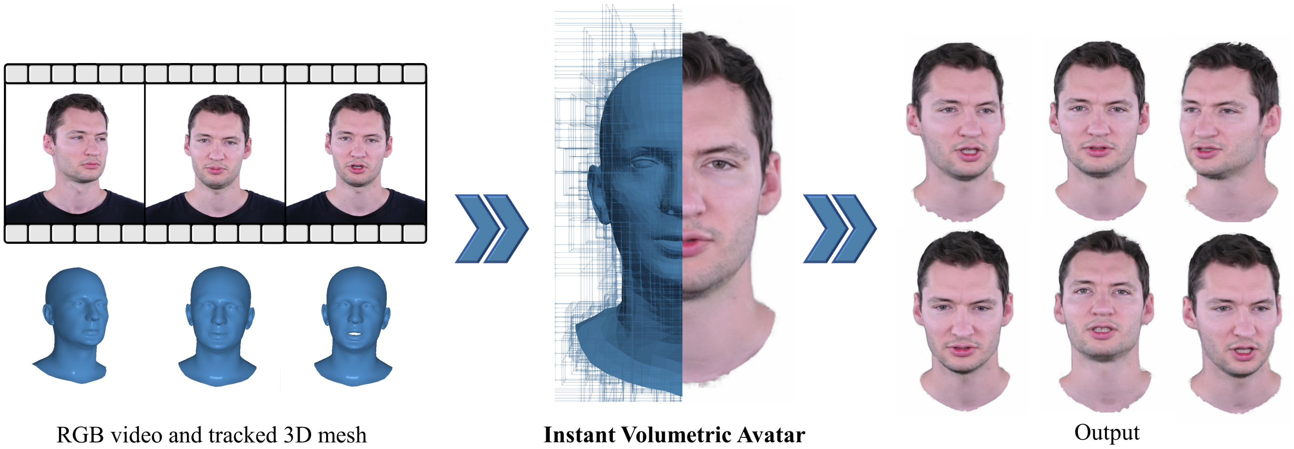

Figure 1. Given a short monocular RGB video, our method instantaneously optimizes a deformable neural radiance field to synthesize a photo-realistic animatable 3D neural head avatar. The neural radiance field is embedded in a multi-resolution grid around a 3D face model which guides the deformations. The resulting head avatar can be viewed under novel views and animated at interactive frame rates.

Abstract

We present Instant Volumetric Head Avatars (INSTA), a novel approach for reconstructing photo-realistic digital avatars instantaneously. INSTA models a dynamic neural radiance field based on neural graphics primitives embedded around a parametric face model. Our pipeline is trained on a single monocular RGB portrait video that observes the subject under different expressions and views. While state-of-the-art methods take up to several days to train an avatar, our method can reconstruct a digital avatar in less than 10 minutes on modern GPU hardware, which is orders of magnitude faster than previous solutions. In addition, it allows for the interactive rendering of novel poses and expressions. By leveraging the geometry prior of the underlying parametric face model, we demonstrate that INSTA extrapolates to unseen poses. In quantitative and qualitative studies on various subjects, INSTA outperforms state-of-the-art methods regarding rendering quality and training time. Project website: https://zielon.github.io/insta/

1. Introduction

For immersive telepresence in AR or VR, we aim for digital humans (avatars) that mimic the motions and facial expressions of the actual subjects participating in a meeting. Besides the motion, these avatars should reflect the human's shape and appearance. Instead of prerecorded, old avatars, we aim to instantaneously reconstruct the subject's look to capture the actual appearance during a meeting. To this end, we propose Instant Volumetric Head Avatars (INSTA), which enables the reconstruction of an avatar within a few minutes ($\sim$10 min) and can be driven at interactive frame rates. For easy accessibility, we rely on commodity hardware to train and capture the avatar. Specifically, we use a single RGB camera to record the input video. State-of-the-art methods that use similar input data to reconstruct a human avatar require a relatively long time to train, ranging from around one day [20] to almost a week [16,58]. Our approach uses dynamic neural radiance fields [16] based on neural graphics primitives [38], which are embedded around a parametric face model [25], allowing low training times and fast evaluation. In contrast to existing methods, we use a metrical face reconstruction [59] to ensure that the avatar has metrical dimensions such that it can be viewed in an AR/VR scenario where objects of known size are present. We employ a canonical space where the dynamic neural radiance field is constructed. Leveraging the motion estimation employing the parametric face model FLAME [25], we establish a deformation field around the surface using a bounding volume hierarchy (BVH) [12]. Using this deformation field, we map points from the deformed space into the canonical space, where we evaluate the neural radiance field. As the surface deformation of the FLAME model does not include details like wrinkles or the mouth interior, we condition the neural radiance field by the facial expression parameters. To improve the extrapolation to novel views, we further leverage the FLAME-based face reconstruction to provide a geometric prior in terms of rendered depth maps during training of the NeRF [36]. In comparison to state-of-the-art methods like NeRFace [16], IMAvatar [58], or Neural Head Avatars (NHA) [20], our method achieves a higher rendering quality while being significantly faster to train and evaluate. We quantify this improvement in a series of experiments, including an ablation study on our method.

In summary, we present Instant Volumetric Head Avatars with the following contributions:

- a surface-embedded dynamic neural radiance field based on neural graphics primitives, which allows us to reconstruct metrical avatars in a few minutes instead of hours or days,
- and a 3DMM-driven geometry regularization of the dynamic density field to improve pose extrapolation, an important aspect of AR/VR applications.

2. Related Work

INSTA is reconstructing animatable digital human avatars from monocular video data based on 3D neural rendering [48]. Current solutions are using implicit representations [8, 16, 29, 36, 40, 41] optimized via differentiable volumetric rendering, or are based on explicit models [5, 7, 20, 49] for instance, triangle or tetrahedral meshes using differentiable rasterization [10, 22, 30, 33]. For a concise overview of neural rendering methods and face reconstruction, we point the reader to the state-of-the-art reports by Zollhöfer et al. [60], and Tewari et al. [47, 48].

Static Neural Radiance Fields. Mildenhall et al. [36] and its many follow-up works [3,4,28,35,39,44,45,51,56], synthesize novel views of a complex static scene using differentiable volumetric rendering. Many methods suffer from a long training time (1-5 days). To this end, different acceleration methods have been proposed to improve the training time. Yu et al. [15] achieved 100× speedup by using a sparse voxel grid storing density and spherical harmonics coefficients at each node. The final color is the composition of tri-linearly interpolated values of each voxel intersecting with the ray. TensorRF [9] factorizes the 4D NeRF scene into multiple compact low-rank tensor components achieving high performance and compactness. The coordinate-based MLP is replaced with a voxel grid of features, and the final color is its vector-matrix outer product. Müller et al. [38] introduced a new computer graphics primitive in the form of tiny MLPs which benefit from a multi-resolution hashing encoding. The key idea is similar to Yu et al. [15]. The space is divided into an independent multi-level grid with feature vectors at the vertices of the grid. A spatial hash function [46] is used to store the voxel grid efficiently. Each point sampled on the ray is encoded by the interpolated feature vector of the corresponding grid level and passed to a tiny neural network to synthesize the final color. Our method uses this efficient architecture to model the face in a canonical space.

Some of the static NeRF methods [2, 13, 44, 52] use additional depth maps to improve alignment and quality for static scenes. The depth priors help guide the ray sampling and better estimate the transmittance, resulting in improved geometry and color recovery. While we are working with RGB images only, our method leverages the geometry prior of the 3DMM to guide the depth estimation during training, which results in an improved extrapolation ability w.r.t. view changes.

Deformable Neural Radiance Fields. After the introduction of NeRF [36] for static scenes, a natural research direction was to generalize it to dynamic, time-varying ones [14, 26, 40, 41, 43, 50]. The reconstruction problem is divided into two different spaces, the deformed scene, and the canonical space, with a neural network as the mapper between them. For human body modeling, a series of approaches have been proposed that leverage the kinematic chain of the SMPL [32] body model to condition the mapping function. Peng et al. [42] proposed to learn blend weights to estimate the linear blend skinning-based warping field between canonical and deformed space based on the body skeleton. Similarly, Neural Actor [29] uses a 3D body mesh proxy to learn pose-dependent geometric deformation and view-dependent appearance effects defined in the canonical space. Lombardi et al. [31], which defines surface-aligned neural volumes to improve the rendering speed. Garbin et al. [18] build a tetrahedral deformation graph around a radiance field based on the underlying mesh on which the deformations are defined, effectively transforming sampled points according to the current cage state. Xu et al. [53] propose surface-aligned neural radiance fields by projecting points in space to the surface of the body mesh. Our idea is based on a similar principle. However, instead of projecting points onto the mesh surface, we construct a 3D space around the head and deform points based on the deformation defined by the nearest triangles.

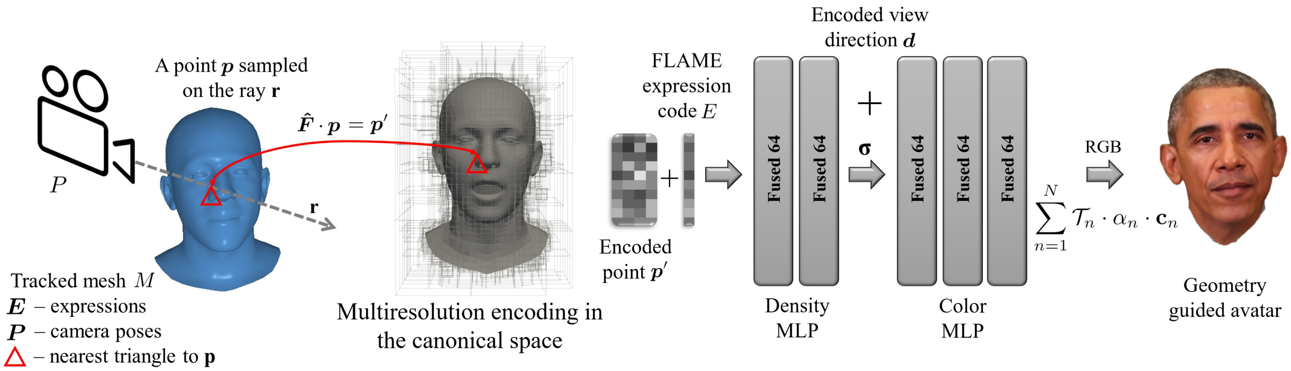
A point $\boldsymbol{p}$ sampled on the ray $\mathbf{r}$
$\hat{\boldsymbol{F}} \cdot \boldsymbol{p} = \boldsymbol{p}'$
P
$\mathbf{r}$
Tracked mesh M
$\boldsymbol{E}$ – expressions
$\boldsymbol{P}$ – camera poses
△ – nearest triangle to $\mathbf{p}$
Multiresolution encoding in the canonical space
FLAME expression code E
Encoded point $\boldsymbol{p}'$
Encoded view direction $\boldsymbol{d}$
Fused 64
Fused 64
σ
Fused 64
Fused 64
Fused 64
Density MLP
Color MLP
RGB
$\sum_{n=1}^{N} \mathcal{T}_n \cdot \alpha_n \cdot \mathbf{c}_n$
Geometry guided avatar

Figure 2. **Overview**. INSTA follows differentiable volumetric optimization introduced in [36, 38]. For each sampled point $\boldsymbol{p} \in \mathbb{R}^4$ in deformed space (in homogeneous coordinates), we are computing the nearest neighbor triangle on the mesh $T_{def} \in \boldsymbol{M}_i$ and its topological corresponding twin in the canonical space $T_{canon} \in \boldsymbol{M}^{canon}$. The deformation gradient of the triangle from deformed space to canonical space $\hat{\boldsymbol{F}} \in \mathbb{R}^{4\times4}$ defines the deformation field. Specifically, $\boldsymbol{p}$ is transformed to the canonical space by $\boldsymbol{p}' = \hat{\boldsymbol{F}} \cdot \boldsymbol{p}$. After canonicalization, the point is encoded using a multi-resolution hashing [38]. This feature is passed to fully fused multi-layer perceptrons [37] with additional conditioning on the facial expressions E_i and the encoded view direction $\boldsymbol{d}$.

In contrast to modeling the deformation explicitly, Gafni et al. [16] implicitly model the facial expressions by conditioning the NeRF MLP with the global expression code obtained from 3DMM tracking [49] and by optimizing per latent frame codes to increase the network capacity for overfitting. In our approach, we leverage the idea of dynamic neural radiance fields to improve the mouth region's rendering, which is not represented by the face model motion prior. Inspired by 3DMMs, IMAvatar [58] learns the subject-specific implicit representation of texture together with expression blendshapes and blend skinning weights. They optimize an implicit surface by incorporating ray marching from Yariv et al. [54] with root-finding of the occupancy function [11] to locate canonical correspondence of deformed points. However, we found the training time-consuming (∼5 days) and unstable (can diverge). In a concurrent work, Gao et al. [17] create personalized blendshapes using neural graphics primitives, where for each of the blendshapes, a multi-resolution grid [38] is trained.

3. Instant Deformable Neural Radiance Field

Our goal is to create instant digital avatars which can be learned in a few minutes and rendered in interactive time. For this purpose, we are using a geometry-guided deformable neural radiance field embedded into a multi-resolution hashing grid [38], exploiting differentiable volumetric rendering [36] (see Fig. 2).

For a given monocular video consisting of images $\boldsymbol{I} = \{I_i\}$ along with optimized intrinsic camera parameters $\boldsymbol{K} \in \mathbb{R}^{3\times3}$, tracked FLAME [25] meshes $\boldsymbol{M} = \{M_i\}$ with corresponding facial expression coefficients $\boldsymbol{E} = \{E_i\}$ and poses $\boldsymbol{P} = \{P_i\}$, our goal is to build a controllable head avatar represented by a neural radiance field. To this end, we employ a canonical space where the neural radiance field is constructed. To render specific facial expressions using volumetric rendering, we canonicalize the samples on a ray from the deformed space to query the radiance field in the canonical space.

Volumetric Rendering. We take advantage of the recent advances in interactive NeRF optimization and use neural graphic primitives [38] to represent the radiance field. The representation of the avatar is optimized using the differentiable volumetric rendering equation:

$$\hat{\boldsymbol{C}} = \int_0^D \mathcal{T}(t) \cdot \sigma(t) \cdot \mathbf{c}(t)\, dt \; + \; \mathcal{T}(D) \cdot \mathbf{c}_{\text{bg}}, \tag{1}$$

where $\mathcal{T}(t_n) = \exp\left(-\int_0^{t_n} \sigma(t)\, dt\right)$ is the transmittance which indicates the probability of a ray traveling from $[0, t_n)$ without interaction with any other particles [36], $\sigma(t)$ is the density and $\mathbf{c}(t)$ is the radiance at position $\boldsymbol{p}_t$. Note that the sample points $\boldsymbol{p}_t$ are canonicalized to access the actual radiance field. Following NeRFace [16], we condition every sample $\boldsymbol{p}_t$ on the ray with the 3DMM facial expression code $E_i \in \mathbb{R}^{16}$ of video frame i. Please note that in contrast to NeRFace [16] and IMAvatar [58], we do not use additional per-frame learnable codes. The viewing vector $\boldsymbol{v} \in \mathbb{R}^3$ is encoded using spherical harmonics projection on four basis functions [1, 38] resulting in the final viewing vector encoding $\boldsymbol{d} \in \mathbb{R}^{16}$ which is concatenated with density MLP output. While the viewing conditioning is applied on the entire avatar, the conditioning on facial expressions is bounded to the dynamically changing mouth region and is set to a constant vector $E_i = \mathbf{1}$ for the other regions.

Canonicalization. We define a mapping function $\Phi(\boldsymbol{p}, M_i)$ that projects a point $\boldsymbol{p} \in \mathbb{R}^4$ from the time-varying deformed space (where the volumetric rendering is performed) to the canonical space. The mapping function leverages the time-varying surface approximation M_i and a predefined mesh in canonical space M^{canon}. We employ a nearest triangle search in deformed space to compute the deformation gradient $\boldsymbol{F} \in \mathbb{R}^{4\times4}$ which is used to map point $\boldsymbol{p}$ to the canonical counterpart $\boldsymbol{p'}$. The deformation gradient $\boldsymbol{F}$ is computed via the known Frenet frames of the deformed triangle $T_{def} \in M_i$ and the canonical triangle $T_{canon} \in M^{canon}$. Specifically, we compute the rotation matrices $\{\boldsymbol{R}_{canon}, \boldsymbol{R}_{def}\} \in \mathbb{R}^{3\times3}$ based on the corresponding tangent, bitangent, and normal vectors of a triangle. With the translations $\{\boldsymbol{t}_{canon}, \boldsymbol{t}_{def}\} \in \mathbb{R}^3$ defined by a vertex of the triangle, they form the Frenet coordinate system frames $\boldsymbol{L}_{canon}$ and $\boldsymbol{L}_{def} \in \mathbb{R}^{4\times4}$:

$$\boldsymbol{L}_{def} = \begin{bmatrix} \boldsymbol{R}_{def} & \boldsymbol{t}_{def} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, \quad \boldsymbol{L}_{canon} = \begin{bmatrix} \boldsymbol{R}_{canon} & \boldsymbol{t}_{canon} \\ \boldsymbol{0}^T & 1 \end{bmatrix}. \tag{2}$$

To account for any potential triangle size change between deformed and canonical spaces, we compute an isotropic scaling factor $\lambda \in \mathbb{R}$ via the relative surface area change of the given triangle w.r.t. its canonical twin $\lambda = \frac{a_{def}}{a_{canon}}$. The deformation gradient F is defined as:

$$\boldsymbol{F} = \boldsymbol{L}_{canon} \cdot \Lambda \cdot \boldsymbol{L}_{def}^{-1}, \quad \Lambda = \begin{bmatrix} \lambda \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix}. \tag{3}$$

To avoid transformation discontinuity, which arises from the local coordinate system of each triangle, we additionally perform exponentially weighted averaging of the transformations of the adjacent faces of the triangle's edges:

$$\hat{\boldsymbol{F}} = \frac{1}{\sum_{f\in A} \omega_f} \cdot \sum_{f\in A} \omega_f \boldsymbol{F}_f, \tag{4}$$

where $\omega_f = \exp(-\beta||\boldsymbol{c}_f - \boldsymbol{p}||_2)$, $\beta = 4$ and A is the set of adjacent faces to T (including T with $\beta = 1$) with corresponding centroids $\boldsymbol{c}_f$. Please note that all vertex positions are defined in meters (FLAME metrical space).

To achieve interactive rendering as well as instantaneous optimization of the neural radiance field, we leverage a classical bounding volume hierarchy (BVH) [12] to significantly increase the nearest triangle search speed for the sampled points $\boldsymbol{p}_t$ on the ray. Note that methods like IMAvatar [58] perform computation-heavy root-finding procedures to calculate surface points iteratively [11]. Our method builds a BVH based on the corresponding deformed mesh M_i of frame i to establish the mapping function to the canonical mesh. Our BVH is implemented on GPU to utilize massively parallel nearest triangle search [23]. To alleviate the triangle search for highly tessellated FLAME regions, we simplified the eyeballs and the eye region [19]. Moreover, an additional set of triangles in the mouth region is used to serve as a deformation proxy (see sup. mat.).

3.1. Training Objectives

The optimization of the neural radiance field is based on a color reproduction objective and a geometry prior based on the 3DMM. Following NeRF [36], we redefine the volumetric rendering Equation (1) with piece-wise constant density and color, and rewrite it in terms of alpha-compositing:

$$\hat{\boldsymbol{C}}(t_{N+1}) = \sum_{n=1}^{N} \mathcal{T}_n \cdot \alpha_n \cdot \mathbf{c}_n, \tag{5}$$

where $\mathcal{T}_n = \prod_{n=1}^{N-1}(1 - \alpha_n)$ weight α_n is defined as $\alpha_n \equiv 1 - \exp(-\sigma_n \delta_n)$ and δ_n is a step size equal $\frac{\sqrt{3}}{1024}$. To measure the photometric error, we use a Huber loss [21] with $\rho = 0.1$:

$$\mathscr{L}_{color} = \begin{cases} \frac{1}{2}(\boldsymbol{C} - \hat{\boldsymbol{C}})^2 & if \left|(\boldsymbol{C} - \hat{\boldsymbol{C}}\right| < \rho \\ \rho((\boldsymbol{C} - \hat{\boldsymbol{C}}) - \frac{1}{2}\rho) & otherwise \end{cases} \tag{6}$$

We enforce a depth loss to leverage the geometry prior of the reconstructed face based on the 3DMM FLAME. Specifically, we rasterize the depth of the tracking mesh M_i and apply an L1 distance between this map and the ray termination of the volumetric rendering. As the FLAME model does not contain details like hair, we restrict the geometry prior to the face region:

$$\mathscr{L}_{geom} = \sum_{\boldsymbol{r}} |\mathbb{1}_{face}\{(z(\boldsymbol{r}) - \hat{z}(\boldsymbol{r}))\}|, \tag{7}$$

where $\hat{z} = \sum_{n=1}^{N} \mathcal{T}_n \cdot \alpha_n \cdot t_n$, and t_n is the current sample position, and $\mathbb{1}_{face}\{\}$ is a segmentation indicator function which enables the loss for the face region. The $\mathbb{1}_{face}$ function uses face parsing information [55] to decide a given pixel membership. The total loss $\mathscr{L}$ is defined as:

$$\mathscr{L} = \sum_{\boldsymbol{r}} \lambda_{color}(\boldsymbol{r})\mathscr{L}_{color}(\boldsymbol{r}) + \lambda_{geom}\mathscr{L}_{geom}(\boldsymbol{r}), \tag{8}$$

where $\lambda_{geom} = 1.25$ controls the influence of the geometry prior and $\lambda_{color}(\boldsymbol{r})$ weights the color loss contribution based on a face parsing mask. Specifically, we weight the color loss higher for the mouth region with $\lambda_{color} = 40$ and $\lambda_{color} = 1$ otherwise.

We implemented our animatable dynamic radiance field using the Nvidia NGP C++ framework [37]. We use two fully fused MLPs [37], each with 64 neurons, for color and density predictions. The density MLP outputs feature values vector $\boldsymbol{\sigma} \in \mathbb{R}^{16}$ where the first value is the log-space density. The vector $\boldsymbol{\sigma}$ is later concatenated with the encoded viewing vector $\boldsymbol{d}$ to be the input of the color network. For optimization, we used Adam [24] with an exponential moving average on the weights and fixed learning rate $\eta = 2.5\mathrm{e}{-3}$. In our experiments, we train the network for 32k optimization steps. We randomly sample 1700 frames from the whole dataset during the training and load them into the processing buffer. Every 1500 steps, we repeat the procedure and resample the dataset.

4. Dataset

Our method takes a single video as input to generate the volumetric avatar of the depicted subject. For our experiments, we recorded multiple actors with a Nikon Z6 II Camera as well as used sequences from Youtube, resulting in a set of twelve actors. For the in-house recordings, we captured around 2-3min of monocular RGB Full HD videos, which later were cropped, sub-sampled to 25fps, and resized to 512^2 resolution. We additionally use background foreground segmentation using robust matting [27] and an off-the-shelf face parsing framework [55] for image segmentation and clothes removal.

Dataset Tracking Generation. An essential part of this project is temporally stable face tracking of the monocular input data. To this end, we use the analysis-by-synthesis-based face tracker from MICA [59], based on Face2Face [49] using a sampling-based differentiable rendering. We refer to the original paper [49] for more details. We extend the optimization with two extra blendshapes for eyelids and iris tracking using Mediapipe [34]. In contrast to MICA, we also optimize for FLAME shape parameters, with regularization towards MICA shape prediction instead of the average face shape as in Face2Face [49]. Note that for our prototype, we implemented the tracking in PyTorch, which is significantly slower than the original Face2Face implementation, which can track faces in real-time.

5. Results

In this section, we evaluate the quality of the synthesized digital human avatars generated by our method INSTA in comparison to state-of-the-art. For this purpose, we use the test sequences from our dataset, which consist of the last 350 frames of each video.

5.1. Image Quality Evaluation

To evaluate our method in terms of the image quality and novel view extrapolation, we make a comparison to NeRFace [16], IMAvatar [58], and Neural Head Avatars (NHA) [20]. For this comparison, we use the original implementations of the authors. Note that for IMAvatar, we use the most recent version of the author's code, which contains additional semantic information for mouth interior and FLAME geometry supervision which is different from the original paper. Figure 4 depicts qualitative results evaluated on the test sequences. To evaluate the image quality of the results quantitatively, we use several pixel-wise metrics; mean squared error, SSIM, PSNR, and the perceptual metric LPIPS [57] (see Table 1). Note that IMAvatar is trained at a resolution of 256^2 due to its computational complexity; for the comparison, we upsample the results to 512^2.

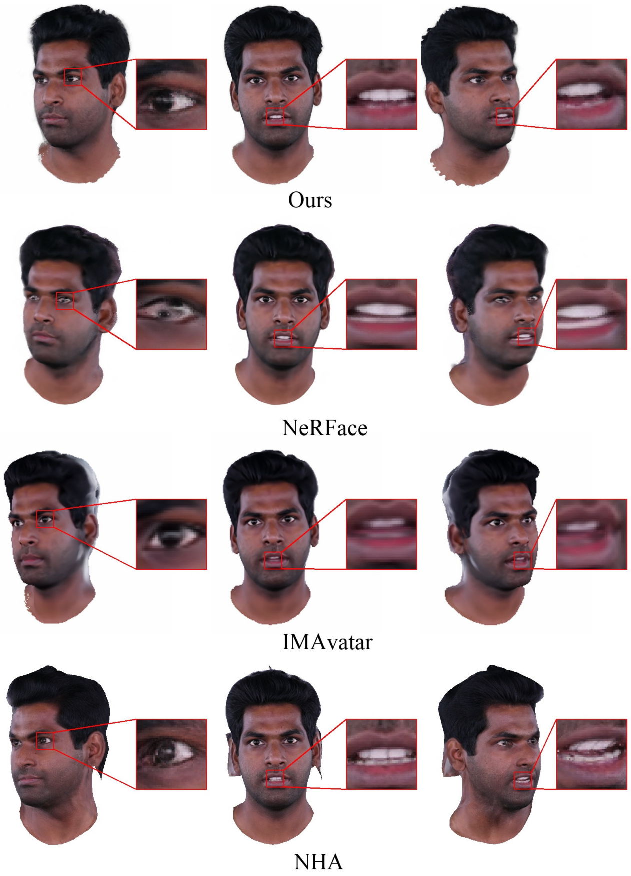

Figure 3. Qualitative comparison for novel view extrapolation. As can be seen, our method can better handle image synthesis under novel poses. NHA [20] suffers from degenerated geometry with many artifacts at the ear region. NeRFace [16] lacks high-frequency details for eyes and teeth, and IMAvatar [58] shows silhouette artifacts at gracing angles.

All methods produce sharp and photo-realistic images which are hard to distinguish from the ground truth. However, the most noticeable artifacts, especially for the ear regions, were generated by NHA. Moreover, IMAvatar, for some of the videos, had problems with convergence and stability, leading to diverging optimization and premature termination of the training. Compared to these methods,

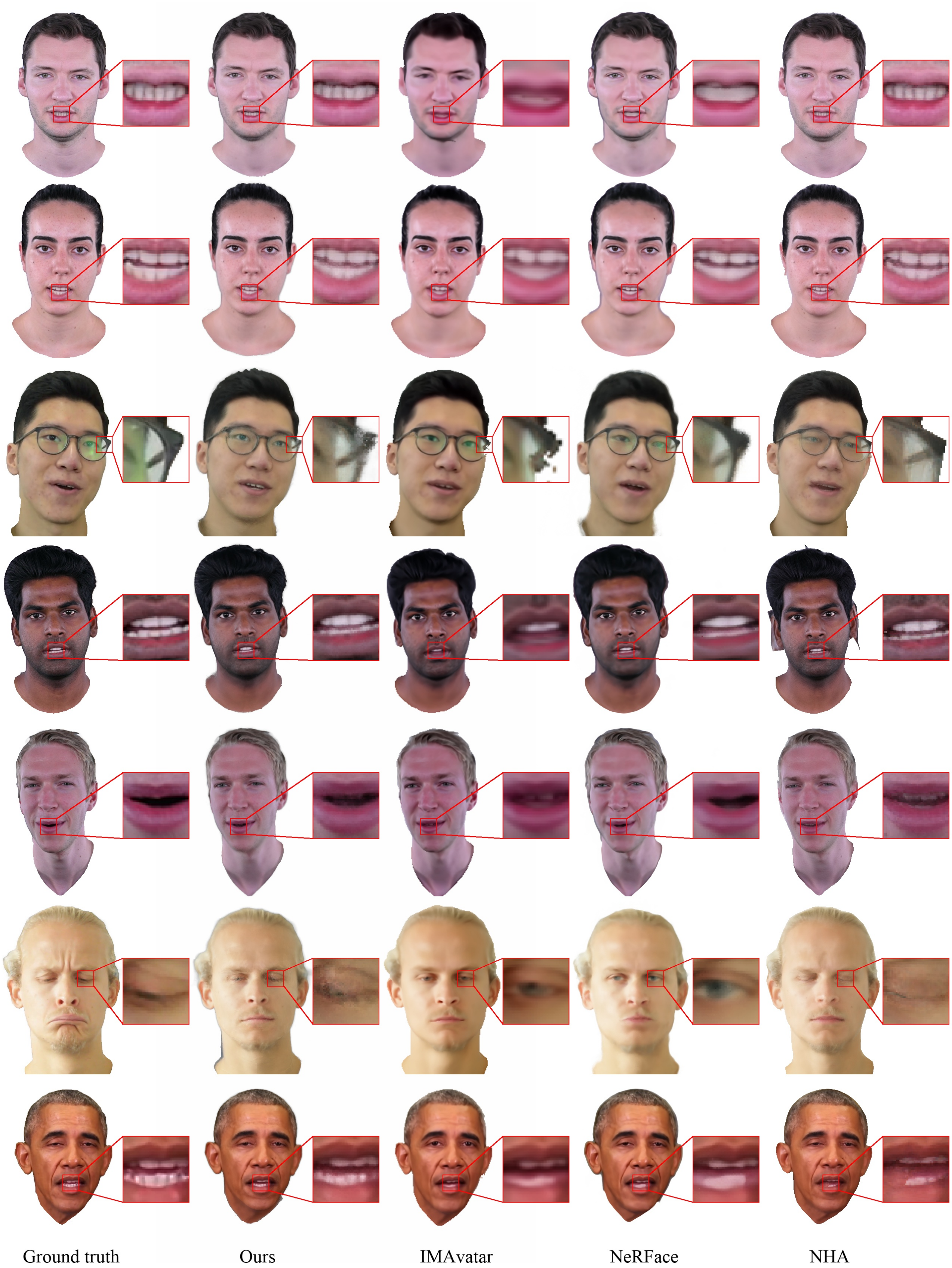

Figure 4. Qualitative comparisons show that our method produces high-quality facial avatars which beat the state-of-the-art methods in terms of image quality (e.g., capturing fine details like lips and teeth) while being significantly faster to obtain.

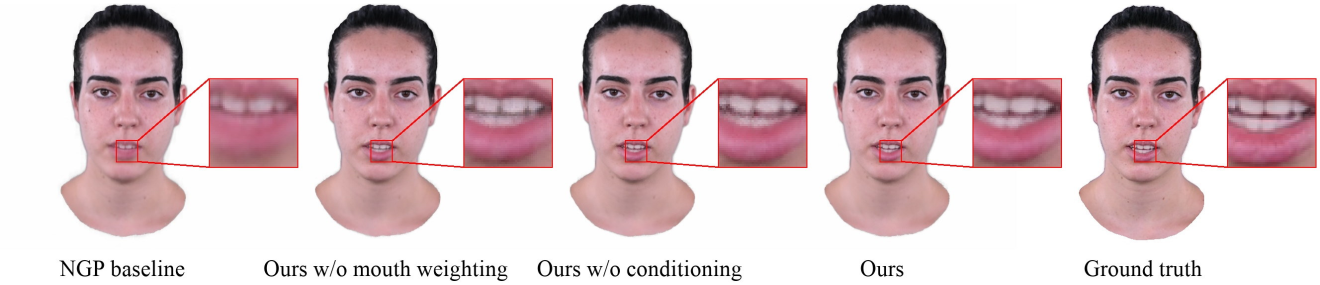

Figure 5. Embedding the neural radiance field around the deformable face model allows us to model dynamic sequences in contrast to the static radiance field of NGP [38]. The expression conditioning and face-parsing-based weighting leads to sharper teeth reconstructions.

Method	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓
NHA [20]	0.0022	27.71	**0.95**	**0.04**	0.63
IMAvatar [58]	0.0023	27.62	0.94	0.06	12.34
NeRFace [16]	**0.0018**	**29.28**	**0.95**	0.07	9.68
Ours	**0.0018**	28.97	**0.95**	0.05	**0.05**

Table 1. Average photometric errors over 19 videos from our dataset, NHA, IMAvatar, and NeRFace datasets (see Fig. 4). The average rendering time of a single frame in seconds is denoted as *Time* in the rightmost column. Our method is on par with NeRFace of Gafni et al. w.r.t. the pixel-wise error metrics. Additionally, our approach achieves low perceptual error in comparison to all methods while being significantly faster to train and evaluate.

our approach can achieve the best image quality while being significantly faster to train (see sup. mat.).

Extrapolation to novel views is an essential aspect of 3D digital avatars that are used in AR or VR applications. In Figure 3, we depict a viewpoint extrapolation comparison with the baseline methods. We can observe that NeRFace [16] produces blurry results in the area of eyes and teeth. IMAvatar [58] exhibits artifacts at gracing angles at the silhouette, and NHA [20] suffers from degenerated geometry with strong artifacts at the ears. In contrast to these methods, our method can robustly generate photo-realistic images under novel poses and achieves high visual quality, especially in the skin and mouth region.

5.2. Ablation Studies

We conducted a series of ablation studies to analyze the different components of our pipeline. Specifically, we are interested in the influence of localized expression conditioning for teeth quality (Figure 5), the effect of the geometric prior (Figure 8), especially for the novel view synthesis, and the importance of the deformation field (Figure 6).

Deformation Field. Figure 6 shows the impact of the deformation field and the conditioning on the quality of the renderings. We conducted two experiments where we used **a)** a global conditioning instead of the local one and **b)** global conditioning with per-frame learnable codes and without the deformation field (similar to NeRFace). As can be seen, local conditioning and the mesh-based deformation field helps to avoid overfitting to the short training sequences.

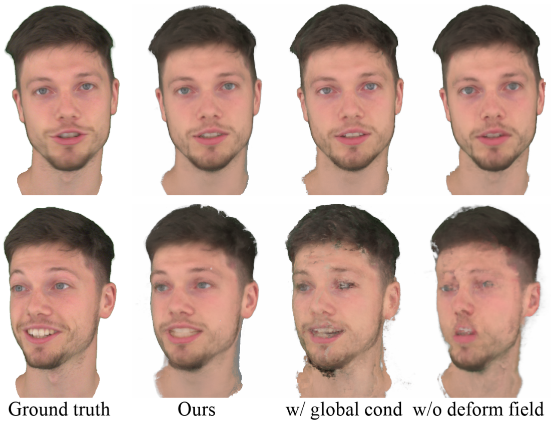

Figure 6. Ablation study w.r.t. the conditioning and deformation field. From left to right: ground truth, ours, ours with global conditioning, and ours without deformation field but with per-frame learnable codes (NeRFace).

Geometric Prior. We leverage the geometric prior of the 3DMM FLAME [25] to regularize the depth estimations of our volumetric rendering method. During training, we render depth maps of the per-frame 3DMM reconstructions and measure a loss between the estimated ray termination and the depth of the rendered face model. In Figure 8, we show an ablation study w.r.t. this geometric prior. The generated digital avatar is shown from an unseen profile view, an extreme extrapolation from the training data which observed views in a range of ±40°. Using the additional geometric prior improves the stability and quality of the results.

Expression Conditioning. Most publicly available 3DMMs [6,25] do not explicitly model teeth. However, this region is especially challenging for the reconstruction of 3D facial avatars due to highly dynamic lips, which can occlude the teeth depending on the given expressions. To compen-

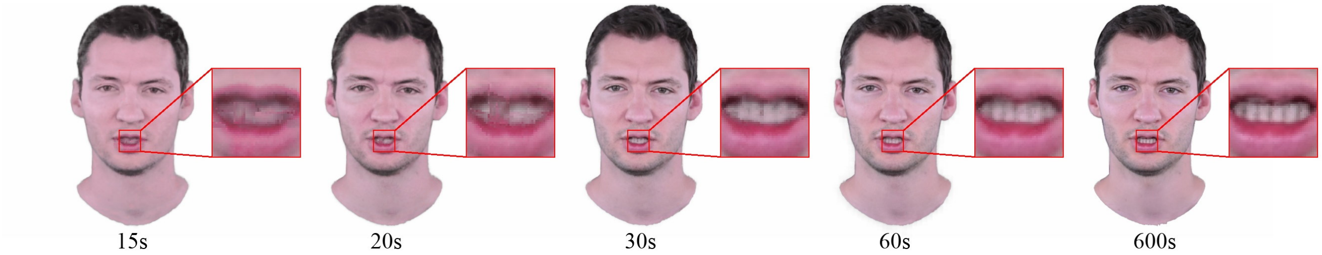

Figure 7. INSTA allows training personalized volumetric avatars from RGB videos within a couple of seconds. Already after 30 seconds of optimization, we achieve good results where the geometry and appearance match the input. To improve the reconstruction of high-frequency details like teeth, the method needs to train approximately 10 min.

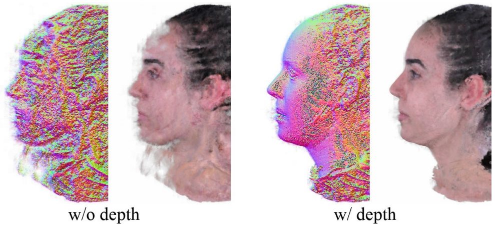

Figure 8. The geometric prior of the 3DMM helps for extrapolation to extreme novel views, in this case, 90°.

sate for the missing geometry, we condition this region on FLAME expression coefficients. In Figure 5, we show that using this additional information helps to improve the synthesis of the mouth interior. Furthermore, we demonstrate that a higher color term weight on the mouth region (Equation (8)) improves the visual quality.

6. Discussion

While our method INSTA shows better quality and speed compared to state-of-the-art RGB-video-based avatar generation techniques, there are still several challenges that need to be addressed in future work. Our model handles the dynamically changing facial expressions but does not capture dynamically changing hairs. Thus, the hair quality is not on par with the face interior and still needs improvements in the level of detail. Furthermore, the used 3DMM does not model teeth geometry. A better approximation of the mouth region would increase the viewpoint extrapolation with improved quality of teeth. While our method achieves real-time frame rates for rendering at a resolution of 512^2, the rendering speed needs to be improved to enable high-quality video conferences in AR or VR, especially when a higher resolution is required. With additional engineering, the training process of our method could be moved to a background process that would continuously refine our canonical avatar after an initial warm-up stage. For example, regions initially not visible could be captured during the conversation, and the avatar would be updated accordingly.

7. Limitations

An important quality factor of our method is face tracking, as misalignments of the geometry and the images will be propagated to the final avatar. Another limiting aspect is the mouth interior quality due to the lack of geometry in that region, as can be seen in Figure 9.

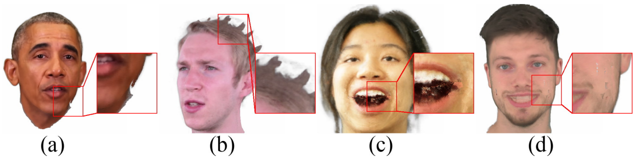

Figure 9. Failure cases: **(a)** and **(b)** exhibits outline artifacts at the chin and hair which stem from geometry misalignment of the tracker, **(c)** extreme expressions can cause artifacts in the mouth region, and **(d)** extrapolation of expressions can lead to artifacts.

8. Conclusion

Instant Volumetric Head Avatars (INSTA) is a novel approach that instantaneously optimizes geometry-guided 3D digital avatars. Our method takes a monocular RGB video as input and optimizes a subject's dynamic neural radiance field in less than 10 minutes using neural graphics primitives embedded around a 3DMM. In comparisons and ablation studies, we demonstrate the capabilities of INSTA, which enable us to instantaneously create avatars that reflect reality and not a prerecorded appearance that might deviate from the current look of the person. We believe this paradigm change to adaptable online avatars is a stepping stone toward immersive telepresence applications.

Acknowledgement. The authors thank all participants of the study and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting WZ. While TB is a part-time employee of Amazon, his research was performed solely at and exclusively funded by MPI. JT is supported by Microsoft Research gift funds.

References

[1] S. Axler, P. Bourdon, and R. Wade. *Harmonic Function Theory*. Graduate Texts in Mathematics. Springer, 2001. 3

[2] Dejan Azinovic, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6280–6291. IEEE, 2022. 2

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5835–5844. IEEE, 2021. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5460–5469. IEEE, 2022. 2

[5] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. volume 23, pages 669–676, 2004. 2

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. pages 187–194, 1999. 7

[7] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4):40:1–40:10, 2013. 2

[8] 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 Trans. Graph.*, 41(4):163:1–163:19, 2022. 2

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. arXiv, 2022. 2

[10] Wenzheng Chen, Huan Ling, Jun Gao, Edward J. Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9605–9616, 2019. 2

[11] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. SNARF: differentiable forward skinning for animating non-rigid neural implicit shapes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11574–11584. IEEE, 2021. 3, 4

[12] James H. Clark. Hierarchical geometric models for visible-surface algorithms. page 267, 1976. 2, 4

[13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12872–12881. IEEE, 2022. 2

[14] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14304–14314. IEEE, 2021. 2

[15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5491–5500. IEEE, 2022. 2

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8649–8658. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 7

[17] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. volume abs/2210.06108, 2022. 3

[18] Stephan J. Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations. arXiv, 2022. 2

[19] Michael Garland and Paul S. Heckbert. Surface simplification using quadric error metrics. pages 209–216, 1997. 4

[20] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18632–18643. IEEE, 2022. 1, 2, 5, 7

[21] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. 4

[22] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 2

[23] Tero Karras. Thinking parallel. `https://developer.nvidia.com/blog/thinking-parallel-part-i-collision-detection-gpu/`, 2012. 4

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[25] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017. 1, 2, 3, 7

[26] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6498–6508. Computer Vision Foundation / IEEE, 2021. 2

[27] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 3132–3141. IEEE, 2022. 5

[28] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 40(6):219:1–219:16, 2021. 2

[30] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7707–7716. IEEE, 2019. 2

[31] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4):59:1–59:13, 2021. 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. volume 34, pages 248:1–248:16, 2015. 2

[33] Matthew M. Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, volume 8695 of *Lecture Notes in Computer Science*, pages 154–169. Springer, 2014. 2

[34] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. volume abs/1906.08172, 2019. 5

[35] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16169–16178. IEEE, 2022. 2

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020. 2, 3, 4

[37] Thomas Müller. tiny-cuda-nn, 4 2021. 3, 4

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2, 3, 7

[39] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. pages 5470–5480, 2022. 2

[40] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5845–5854. IEEE, 2021. 2

[41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021. 2

[42] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. volume abs/2105.02872, 2021. 2

[43] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10318–10327. Computer Vision Foundation / IEEE, 2021. 2

[44] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. pages 12882–12891, 2022. 2

[45] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. pages 8238–8248, 2022. 2

[46] Matthias Teschner, Bruno Heidelberger, Matthias Müller, Danat Pomerantes, and Markus H. Gross. Optimized spatial hashing for collision detection of deformable objects. In Thomas Ertl, editor, *8th International Fall Workshop on Vision, Modeling, and Visualization, VMV 2003, München, Germany, November 19-21, 2003*, pages 47–54. Aka GmbH, 2003. 2

[47] 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

[48] 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. volume 41, pages 703–735, 2022. 2

[49] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2387–2395. IEEE Computer Society, 2016. 2, 3, 5

[50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12939–12950. IEEE, 2021. 2

[51] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. pages 5481–5490, 2022. 2

[52] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5590–5599. IEEE, 2021. 2

[53] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15862–15871. IEEE, 2022. 2

[54] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[55] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation. volume 129, pages 3051–3068, 2021. 4, 5

[56] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. arXiv, 2020. 2

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 5

[58] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13535–13545. IEEE, 2022. 1, 2, 3, 4, 5, 7

[59] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, volume 13673 of *Lecture Notes in Computer Science*, pages 250–269. Springer, 2022. 1, 5

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