NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads

TOBIAS KIRSCHSTEIN, Technical University of Munich, Germany
SHENHAN QIAN, Technical University of Munich, Germany
SIMON GIEBENHAIN, Technical University of Munich, Germany
TIM WALTER, Technical University of Munich, Germany
MATTHIAS NIESSNER, Technical University of Munich, Germany

Fig. 1. **NeRSemble:** Given multi-view video recordings from twelve cameras (left), our method is capable of synthesizing highly realistic novel views of human heads in complex motion. Our renderings from unseen views (right) faithfully represent static scene parts and regions undergoing highly non-rigid deformations. Along with our method, we publish our high-quality multi-view video capture data of 31.7 million frames from a total of 222 subjects.

We focus on reconstructing high-fidelity radiance fields of human heads, capturing their animations over time, and synthesizing re-renderings from novel viewpoints at arbitrary time steps. To this end, we propose a new multi-view capture setup composed of 16 calibrated machine vision cameras that record time-synchronized images at 7.1 MP resolution and 73 frames per second. With our setup, we collect a new dataset of over 4700 high-resolution, high-framerate sequences of more than 220 human heads, from which we introduce a new human head reconstruction benchmark[1]. The recorded sequences cover a wide range of facial dynamics, including head motions, natural expressions, emotions, and spoken language. In order to reconstruct high-fidelity human heads, we propose Dynamic Neural Radiance Fields using Hash Ensembles (NeRSemble). We represent scene dynamics by combining a deformation field and an ensemble of 3D multi-resolution hash encodings. The deformation field allows for precise modeling of simple scene movements, while the ensemble of hash encodings helps to represent complex dynamics. As a result, we obtain radiance field representations of human heads that capture motion over time and facilitate re-rendering of arbitrary novel viewpoints. In a series of experiments, we explore the design choices of our method and demonstrate that our approach outperforms state-of-the-art dynamic radiance field approaches by a significant margin.

[1]We will release all of our captured data, including all 4734 recordings and baseline codes, along with a new public benchmark to support further research in the area. Website: https://tobias-kirschstein.github.io/nersemble

Authors' addresses: Tobias Kirschstein, Technical University of Munich, Germany, tobias.kirschstein@tum.de; Shenhan Qian, Technical University of Munich, Germany, shenhan.qian@tum.de; Simon Giebenhain, Technical University of Munich, Germany, simon.giebenhain@tum.de; Tim Walter, Technical University of Munich, Germany, tim.michelbach@hotmail.com; Matthias Nießner, Technical University of Munich, Germany, niessner@tum.de.

CCS Concepts: • **Computing methodologies** → *Rendering*; *3D imaging*; *Volumetric models*; *Reconstruction*.

Additional Key Words and Phrases: Neural Radiance Fields, Dynamic Scene Representations, Novel View Synthesis, Multi-View Video Dataset, Human Heads

1 INTRODUCTION

In recent years, we have seen tremendous growth in the importance of digital applications that rely on photo-realistic rendering of images from captured scene representations, both in society and industry. In particular, the synthesis of novel views of dynamic human faces and heads has become the center of attention in many graphics applications ranging from computer games and movie productions to settings in virtual or augmented reality. Here, the key task is the following: given a recording of a human actor who is displaying facial expressions or talking, reconstruct a temporally-consistent 3D representation. This representation should enable the synthesis of photo-realistic re-renderings of the human face from arbitrary viewpoints and time steps.

However, reconstructing a 3D representation capable of photo-realistic novel viewpoint rendering is particularly challenging for dynamic objects. Here, we not only have to reconstruct the static appearance of a person, but we also have to simultaneously capture the motion over time and encode it in a compact scene representation. The task becomes even more challenging in the context of human faces, as fine-scale and high-fidelity detail are required for downstream applications, where the tolerance for visual artifacts

is typically very low. In particular, human heads exhibit several properties that make novel view synthesis (NVS) extremely challenging, such as the complexity of hair, differences in reflectance properties, and the elasticity of human skin that creates heavily non-rigid deformations and fine-scale wrinkles.

In the context of static scenes, we have seen neural radiance field representations (NeRFs) [Mildenhall et al. 2020] obtain compelling NVS results. The core idea of this seminal work is to leverage a volumetric rendering formulation as a reconstruction loss and encode the resulting radiance field in a neural field-based representation. Recently, there has been significant research interest in extending NeRFs to represent dynamic scenes. While some approaches rely on deformation fields to model dynamically changing scene content [Park et al. 2021a,b], others propose to replace the deformation field in favor of a time-conditioned latent code [Li et al. 2022b]. These methods have shown convincing results on short sequences with limited motion; however, faithful reconstructions of human heads with complex motion remain challenging.

In this work, we focus on addressing these challenges in the context of a newly-designed multi-view capture setup and propose NeRSemble, a novel method that combines the strengths of deformation fields and flexible latent conditioning to represent the appearance of dynamic human heads. The core idea of our approach is to store latent features in an ensemble of multi-resolution hash grids, similar to Instant NGP [Müller et al. 2022], which are blended to describe a given time step. Importantly, we utilize a deformation field before querying features from the hash grids. As a result, the deformation field represents all coarse dynamics of the scene and aligns the coordinate systems of the hash grids, which are then responsible for modeling fine details and complex movements. In order to train and evaluate our method, we design a new multi-view capture setup to record 7.1 MP videos at 73 fps with 16 machine vision cameras. With this setup, we capture a new dataset of 4734 sequences of 222 human heads with a total of 31.7 million individual frames. We evaluate our method on this newly-introduced dataset and demonstrate that we significantly outperform existing dynamic NeRF reconstruction approaches. Our dataset exceeds all comparable datasets w.r.t. resolution and number of frames per second by a large margin, and will be made publicly available. Furthermore, we will host a public benchmark on dynamic NVS of human heads, which will help to advance the field and increase comparability across methods.

To summarize, our contributions are as follows:

- A dynamic head reconstruction method based on a NeRF representation that combines a deformation field and an ensemble of multi-resolution hash encodings. This facilitates high-fidelity NVS from a sparse camera array and enables detailed representation of scenes with complex motion.
- A high-framerate and high-resolution multi-view video dataset of diverse human heads with over 4700 sequences of more than 220 subjects. The dataset will be publicly released and include a new benchmark for dynamic NVS of human heads.

Table 1. Existing multi-view video datasets of human faces. Note that for each dataset, we only count the publicly accessible recordings.

Dataset	#Subj.	#Cam.	Resolution	Fps
D3DFACS [2011]	10	6	1280 x 1024	60
BP4D-Spontaneous [2014]	41	3	1392 x 1040	25
Interdigital Light-Field [2017]	5	16	2048 x 1088	30
4DFAB [2018]	180	7	1600 x 1200	60
VOCASET [2019]	12	12	1600 x 1200	60
MEAD [2020]	48	7	1920 x 1080	30
MultiFace [2022]	13	150	2048 x 1334	30
Ours	222	16	3208 x 2200	73

2 RELATED WORK

Modeling and rendering human faces is a central topic in graphics and plays a crucial role in many applications, such as computer games, social media, telecommunication, and virtual reality.

2.1 3D Morphable Models

3D morphable models (3DMMs) have been a staple approach over the last two decades. The use of a unified mesh topology enables representing identity and expression using simple statistical tools [Blanz and Vetter 1999; Li et al. 2017]. With the additional use of texture, one can already produce compelling renderings [Blanz and Vetter 1999; Paysan et al. 2009], but mesh-based 3DMMs are inherently limited w.r.t. modeling hair or fine identity-specific details. More recently, the use of neural fields [Xie et al. 2022] has alleviated the constraint of working on topologically uniform meshes. These models are capable of modeling complete human heads, including hair [Yenamandra et al. 2021] and fine details [Giebenhain et al. 2022]. In another line of work, Zheng et al. [2022] combine ideas from neural fields and classical 3DMMs to fit monocular videos.

2.2 Neural Radiance Fields

Our work strives to achieve highly-realistic renderings of videos, including detailed hairstyles and complex deformations. Therefore, we deviate from common assumptions made in 3DMMs and focus on fitting a single multi-view video sequence to the highest degree of detail possible. Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020] have recently become state-of-the-art in NVS. While the first NeRFs were usually trained for hours or days on a single scene, recent research advances have reduced the training time to several minutes. For example, this can be achieved by grid-based optimization [Fridovich-Keil and Yu et al. 2022; Karnewar et al. 2022; Sun et al. 2022], tensor decomposition [Chen et al. 2022], or Instant NGP's [Müller et al. 2022] multi-resolution voxel hashing.

2.3 Dynamic NeRF

Extending NeRFs to time-varying, non-rigid content is another central research topic that has seen fast progress. Pumarola et al. [2020] and Park et al. [2021a; 2021b] model a single NeRF in canonical space and explicitly model backward deformations from observed frames to explain the non-rigid content of the scene. OLD: On the

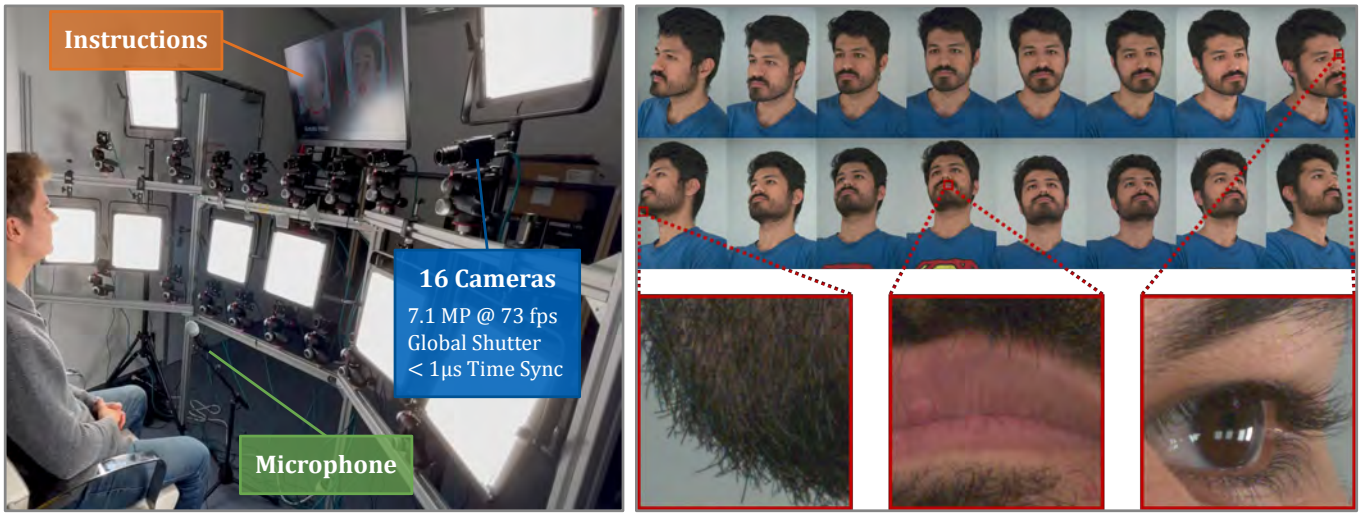

Fig. 2. Left: Our custom-built multi-view video capture setup. Right: The 16 viewpoints and the facial detail obtained from the recordings.

other hand, Li et al. [2022b] refrain from using explicit deformations and instead encode the state of the scene in a latent vector, which is directly conditioning a NeRF. Wang et al. [2022b] utilize Fourier-based compression of grid features to represent a 4D radiance field. Lombardi et al. [2019] use an image-to-volume generator in conjunction with deformation fields.

Concurrent to our work, Song et al. [2023] combine a fast NeRF backbone, i.e. TensoRF or Instant NGP, with a sliding window approach to account for temporal changes. Attal et al. [2023] combine a 4D tensor decomposition with a learned sampling method for fast dynamic NVS. In contrast to these works, we propose a hash-based decomposition in conjunction with a deformation field.

2.4 Video View Synthesis

Besides NeRF, there also exist other methods for video view synthesis that do not rely on a radiance field backbone. In an early work, Zitnick et al. [2004] use geometry-assisted image-based rendering to render novel views of dynamic scenes. More recently, Broxton et al. [2020] obtain free viewpoint videos by constructing multi-sphere images that are then transformed into a layered mesh representation for fast rendering and streaming. A different approach is pursued by Collet et al. [2015], who obtain tracked meshes of dynamic performances with a multi-view stereo system. While these mesh-based methods produce compelling video view synthesis for larger scenes, the strength of NeRFs lies in photo-realistic reconstruction of fine and complex details such as hair.

2.5 NeRF for Faces

Several works propose methods specialized to the domain of human heads. Notably, Gafni et al. [2021] use fitted 3DMM parameters to condition a NeRF, and Athar et al. [2022] extend this approach to model explicit deformations derived from the 3DMM's geometry. More recently, Zielonka et al. [2023] propose a similar approach focused on reconstruction speed and real-time rendering by utilizing a tracked 3DMM in conjunction with Instant NGP. Wang et al. [2022a] propose a generative NeRF with control over identity parameters. Hong et al. [2022] pursue a similar approach with additional expression parameters. Lombardi et al. [2021] propose a highly-optimized approach to neural rendering by explicitly storing color emission values in voxel grids that are loosely rigged to a 3DMM's surface. In this work, we propose a template-free approach as we argue that it is difficult to achieve pixel-accurate novel view synthesis with coarse geometry proxies such as FLAME [Li et al. 2017].

Similar to our method, Gao et al. [2022] recently proposed to blend features from multiple hash grids. While their approach uses parameters from a tracked 3DMM, NeRSemble jointly optimizes for blend weights and the remaining model parameters. Additionally, we show that including a deformation field before blending the hash grids brings significant improvements.

3 MULTI-VIEW VIDEO DATASET OF HUMAN FACES

We introduce a novel dataset consisting of 4734 multi-view video recordings of 222 subjects that were captured with 16 machine vision cameras. Our forward-facing capture rig covers a field of view of 93° left-to-right and 32° up-to-down. As human face motion is complex and the perceived emotion can be heavily influenced by subtle differences, we use a high resolution of 7.1 megapixels, encompassing the whole face up to the level of individual hair strands and wrinkles, as shown in Figure 2. We also ensure that no subtle movements are missed by recording at 73 frames per second. Taken together, our dataset is a unique combination of high-resolution, high frame-rate recordings of many subjects, which is currently unmatched by any other dataset (see Table 1).

3.1 Acquisition

Each recording session consists of 25 short sequences, resulting in around 3 minutes of multi-view video footage per person. We ask the participants to perform a diverse set of facial expressions in

Table 2. **Statistics of our multi-view video face dataset.**

#Participants	#Sequences	#Frames	Total Time	Disk Space
222 (157m / 65f)	4734	31.7 million	7h 30m	203 TB

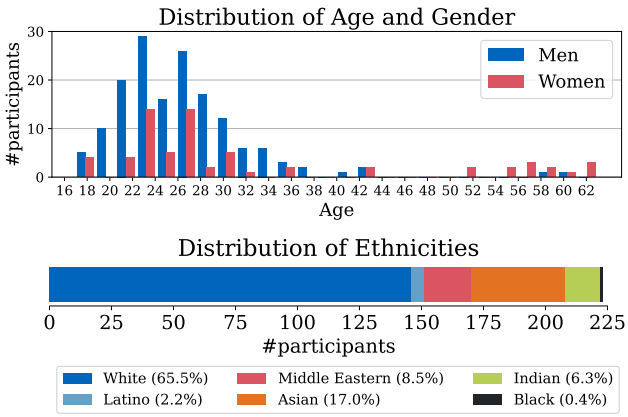

Fig. 3. **Statistics of the participants in our dataset.** Our recorded sequences feature a wide range of ages and ethnicities from both genders.

order to maximize the variety of motion. Specifically, our capture script consists of 9 expression sequences covering different facial muscle groups, 1 hair sequence with fast movements, 4 emotion sequences, 10 sentences with audio, and 1 longer sequence where subjects are free to perform arbitrary facial deformations and head motions.

To obtain high-quality video recordings, we employ a shutter speed of 3ms, which allows us to capture fast movements while avoiding motion blur. Furthermore, we use a small lens aperture to obtain sharp images everywhere in the face region. This combination yields high-quality captures but reduces the amount of incident light at the camera sensors, which requires us to illuminate our scene with 8 strong LED light panels. We further use diffusor plates on the lights to reduce specularities on the skin. Additionally, our cameras employ the precision time protocol (PTP) for accurate time synchronization. The synchronized clocks have sub-microsecond accuracy, resulting in video frames that are effectively captured simultaneously. Finally, we make use of a color checker to calibrate the white balancing factors as well as the gamma parameters of each camera. The resulting video recordings have consistent colors across viewpoints and capture fine details as shown in Figure 2.

3.2 Processing

We estimate an individual extrinsic and a shared intrinsic camera matrix by employing a fine checkerboard in combination with a bundle adjustment optimization procedure. This leads to accurate estimated camera poses, which we verified to be in the sub-millimeter regime in a synthetic setting. Furthermore, the background of our recordings is a white wall, which is captured prior to recording. From these empty backgrounds, it is later feasible to obtain high-quality foreground segmentation maps for each frame via image matting methods, e.g., using BackgroundMatting v2 [Lin et al. 2021].

Fig. 4. **Structure of our dataset.** We ask every participant to perform the same sequence of expressions.

3.3 Benchmark

Our dataset enables us to study photo-realistic human head reconstruction from multi-view videos, which is the goal of this work. Moreover, the captured data allows for use cases far beyond NVS such as generalization over human heads, immersive video conferencing, VR-ready avatar rendering, studying microexpressions, re-enacting, animating, and many more. As such, we plan to release the full dataset to the academic community. Furthermore, we will use a representative selection of recordings from our dataset to compile a benchmark for NVS on human faces. We hope that this endeavor promotes comparability across methods and ultimately advances research on high-fidelity human head reconstruction.

3.4 Data Privacy

Due to the sensitivity of the captured data, all participants in our dataset signed an agreement form compliant with GDPR requirements. Please note that GDPR compliance includes the right for every participant to request the timely deletion of their data. We will enforce these rights in the distribution of our dataset.

4 DYNAMIC NERF USING HASH ENSEMBLES

Our goal is to find a spatio-temporal representation that allows for highly realistic NVS of human heads undergoing complex non-rigid deformations. To this end, we propose a combination of a deformation field and a decomposition of the 4D scene volume into an ensemble of 3D feature volumes along the temporal dimensions in order to reconstruct the dynamics of a scene (see Figure 5).

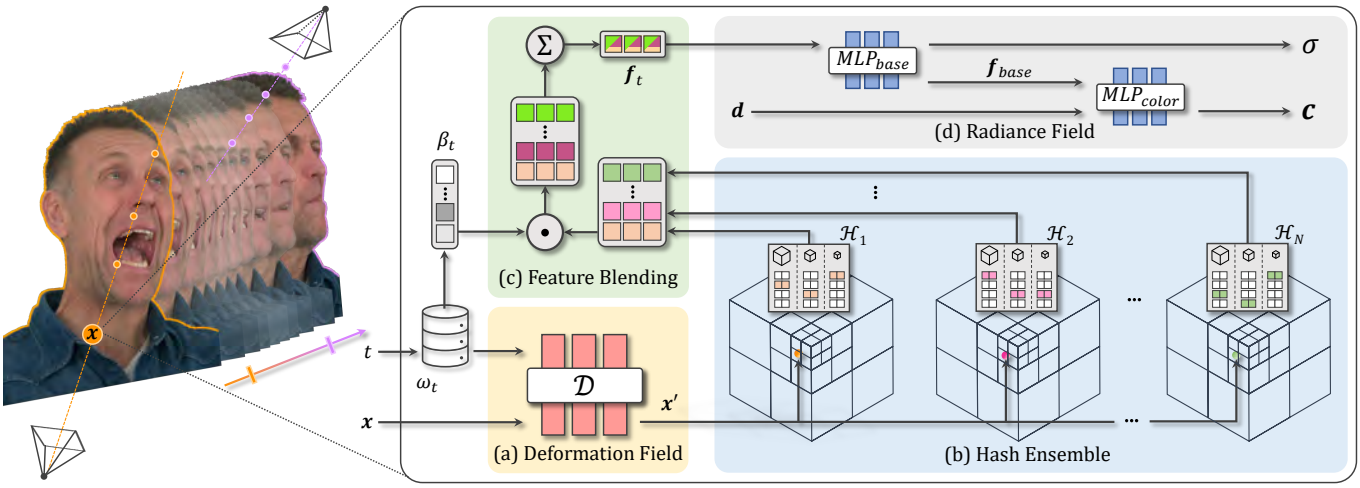

Fig. 5. **Method Overview.** NeRSemble represents a spatio-temporal radiance field for dynamic NVS using volume rendering (left). On the right side, we show how NeRSemble obtains a density $\sigma(\mathrm{x})$ and color value $\mathrm{c}(\mathrm{x}, \mathrm{d})$ for a point x on a ray at time t. (a) Given the deformation code $\boldsymbol{\omega_t}$ the point x is warped to $\mathrm{x}' = \mathcal{D}(\mathrm{x}, \boldsymbol{\omega_t})$ in the canonical space. (b) The resulting point is used to query features $\mathcal{H}_i(\mathrm{x}')$ from the i-th hash grid in our ensemble. (c) The resulting features are blended using weights β_t. Note that both $\boldsymbol{\omega_t}$ and β_t contribute to explaining temporal changes. (d) We predict density $\sigma(\mathrm{x})$ and view-dependent color $\mathrm{c}(\mathrm{x}, \mathrm{d})$ from the blended features using an efficient rendering head consisting of two small MLPs.

4.1 Preliminaries: Neural Radiance Fields

Our work builds on top of the recent success of Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020], which utilize volume rendering through a density field $\sigma(\mathrm{x})$ and view-dependent color field $\mathrm{c}(\mathrm{x}, \mathrm{d})$. Given a ray $\mathrm{r}(\tau) = \mathrm{o} + \tau\mathrm{d}$, a color value

$$C(\mathrm{r}) = \int_{\tau_n}^{\tau_f} \underbrace{T(\tau)\sigma(\mathrm{r}(\tau))}_{w(\tau)} \mathrm{c}(\mathrm{r}(\tau), \mathrm{d})\, \mathrm{d}\tau, \tag{1}$$

is obtained by integrating from near plane τ_n to far plane τ_f along the ray, where $T(\tau) = \exp\left(-\int_{\tau_n}^{\tau} \sigma(\mathrm{r}(s))\, \mathrm{d}s\right)$ denotes the accumulated transmittance up to τ.

The goal of the optimization is to solve for the optimal parameters of a multilayer perceptron (MLP) that encode the resulting radiance field. Recently, pure voxel grids [Fridovich-Keil and Yu et al. 2022] and combinations of explicit grids with MLPs [Müller et al. 2022] have been demonstrated to be effective alternatives for the radiance field representation.

Instant NGP. Our method relies on the voxel hashing scheme of Instant NGP [Müller et al. 2022], which uses multi-resolution features $\mathrm{f}(\mathrm{x})$ in combination with two small MLPs to represent the 3D fields of a NeRF:

$$[\sigma(\mathrm{x}), \mathrm{f}_{\text{base}}(\mathrm{x})] = \mathrm{MLP}_{\text{base}}(\mathrm{f}(\mathrm{x})) \tag{2}$$

$$\mathrm{c}(\mathrm{x}, \mathrm{d}) = \mathrm{MLP}_{\text{color}}(\mathrm{f}_{\text{base}}(\mathrm{x}), \mathrm{d}). \tag{3}$$

Importantly, the features are stored in a multi-resolution hash grid $\mathcal{H}$, s.t. $f(\mathrm{x}) = \mathcal{H}(\mathrm{x})$. The hash grid $\mathcal{H}$ provides a memory-efficient way to encode the 3D scene volume to a stage where a tiny MLP is powerful enough to represent even the most complex of scenes.

4.2 Multi-Resolution Hash Ensemble

Our representation of a dynamic scene is inspired by classical blend shapes [Blanz et al. 2003]. We assume that any state of the scene at time t can be expressed as a combination of feature vectors drawn from a set of multi-resolution hash grids $\{\mathcal{H}_i\}_{i=1}^{N}$, which we refer to as an ensemble of hash grids. To obtain a blended radiance field at time t, we formulate it as a linear combination of its features

$$\mathrm{f}_t(\mathrm{x}) = \sum_{i=1}^{N} \beta_{t,i} \mathcal{H}_i(\mathrm{x}), \tag{4}$$

using blend weights β, which are optimized alongside the hash ensemble, the shared $\mathrm{MLP}_{\text{base}}$ and $\mathrm{MLP}_{\text{color}}$.

This blending operation allows the model to represent complex movements since the blending takes place in feature space. Subsequently, the blended features are decoded by $\mathrm{MLP}_{\text{base}}$ and $\mathrm{MLP}_{\text{color}}$.

4.3 Spatial Alignment of Features

The blending of hash grid features is most effective if all individual elements of the ensemble are operating in a shared coordinate system. For instance, traditional blend shapes operate under perfect correspondences given by the vertex ordering of the mesh topology. Since we blend features without such a structure, we explicitly model the deformation using an $SE(3)$ field, represented by a coordinate-based MLP, following Park et al. [2021a]. More specifically, our deformation field

$$\mathcal{D} : \mathbb{R}^3 \times \mathbb{R}^{d_{\text{def}}} \to \mathbb{R}^3, (\mathrm{x}, \boldsymbol{\omega_t}) \mapsto \mathrm{x}' \tag{5}$$

maps a point x from observed space to its corresponding point x' in the canonical space, given the conditioning code $\boldsymbol{\omega_t}$ which describes the current expression. The deformation field then finds corresponding points across time steps and maps them to a shared canonical space.

Using these learned correspondences, we modify Equation 4 to operate in the canonical space:

$$\mathrm{f}^{(t)} = \sum_{i=1}^{N} \beta_{t,i} \mathcal{H}_i \left(\mathcal{D} \left(\mathrm{x}, \boldsymbol{\omega}_t \right) \right). \tag{6}$$

This way it becomes easier to blend features of the same moving point observed at two different timesteps.

4.4 Warm-Up Phase

With this combination, the hash ensemble and deformations compete to explain the dynamics of the face. Hence, the optimization is likely to result in local minima, in which $\mathcal{D}$ does not provide meaningful deformations. Therefore, we propose a warm-up phase in the optimization procedure in order to encourage $\mathcal{D}$ to learn meaningful correspondences between observed and canonical space.

During the first E_{init} epochs of optimization, we disable all but one hash grid, such that the model essentially mimics a deformable NeRF. During this stage, the deformation field $\mathcal{D}$ along with its deformation codes $\boldsymbol{\omega}_t$ are the only means to explain dynamic behavior. Thus, $\mathcal{D}$ is able to learn meaningful deformations undisturbed, which is essential to effective blending of hash table features later on.

After the warm-up, the first hash table along with our deformation field is able to explain low-frequency dynamics of the scene. We continue to add all remaining hash tables to the optimization over the course of the next E_{trans} epochs. These successively inserted tables enable us to represent fine-scale motion and detail which otherwise cannot be be represented by $\mathcal{D}$.

In order to ensure a smooth transition, we adapt the blend weights

$$\beta_{t,i}(s)^* = \alpha_i(s) \beta_{t,i} \quad (\forall i \in \{1, ..., N\}), \tag{7}$$

where i indexes the hash ensemble, $\alpha_i(s)$ is the windowing function introduced by Park et al. [2021a] and s is scheduled to linearly increase from 1 to N between epochs E_{init} and E_{trans}+E_{init}. Crucially, $\alpha_1(s) = 1$ throughout the complete optimization ensuring that the first hash table is always active.

4.5 Depth Supervision

Since our multi-view dataset provides the capabilities to compute depth maps via traditional methods, we also study the usefulness of additional depth supervision in this work. Given the depth $z^{\text{gt}}(\boldsymbol{r})$ of a ray, we compute the depth loss as

$$\mathcal{L}_{\text{depth}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\text{d}}} \left[\left(z(\boldsymbol{r}) - z^{\text{gt}}(\boldsymbol{r}) \right)^2 \right], \tag{8}$$

where the expected depth of ray $\boldsymbol{r}$ is $z(\boldsymbol{r}) = \int_{\tau_n}^{\tau_f} w(\tau) \cdot \tau \, \mathrm{d}\tau$.
Since depth observations are incomplete in practice, the depth loss is only computed on rays $\boldsymbol{r} \in \mathcal{R}_{\text{d}}$ for which the depth is known.

Additionally, we adopt the two line-of-sight priors from Urban Radiance Fields (URF) [Rematas et al. 2022] to further leverage depth constraints. First, we utilize

$$\mathcal{L}_{\text{empty}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\text{d}}} \left[\int_{\tau_n}^{z(\boldsymbol{r})-\epsilon} w(\tau)^2 d\tau \right] \tag{9}$$

to carve empty space in front of a surface, where ϵ is exponentially decayed during training as in URF. Second,

$$\mathcal{L}_{\text{near}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\text{d}}} \left[\int_{z(\boldsymbol{r})-\epsilon}^{z(\boldsymbol{r})+\epsilon} \left(w(\tau) - \mathcal{N}\left(\tau \mid z(\boldsymbol{r}), \left(\frac{\epsilon}{3} \right)^2 \right) \right)^2 d\tau \right] \tag{10}$$

encourages volumetric density in a neighborhood around the depth observation to follow a narrowing Gaussian distribution. In conjunction with $\mathcal{L}_{\text{depth}}$, these three priors form the depth supervision that is targeted at improving the geometric fidelity of the reconstruction.

4.6 Background Removal

We employ continuous-valued alpha maps $M(\boldsymbol{r})$ to discourage the model from reconstructing parts of the background. We use a sparsity enforcing L1 loss that penalizes density on rays that hit background pixels:

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\text{bg}}} \left[\left\| (1 - T(\tau_f)) - M(\boldsymbol{r}) \right\|_1 \right], \tag{11}$$

where $T(\tau_f)$ is the total transmittance of ray $\boldsymbol{r}$ and $M(\boldsymbol{r})$ is its corresponding alpha value from the precomputed alpha map.

4.7 Optimization Objective

The final loss is comprised of the following terms:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{mask}} + \underbrace{\mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{near}} + \mathcal{L}_{\text{empty}}}_{\text{depth supervision}} + \mathcal{L}_{\text{dist}} \tag{12}$$

where $\mathcal{L}_{\text{rgb}}$ is the standard MSE color loss, which we only compute on foreground rays. We also add a distortion loss $\mathcal{L}_{\text{dist}}$, which penalizes isolated islands of low density [Barron et al. 2022]. As our scenes only consist of human heads, which are roughly convex, we further compute $\mathcal{L}_{\text{dist}}$ on random rays pointing towards the center. This extends the term's regularization effect to the space behind the head, which is often occluded in our scenario.

Finally, we equip each loss term with a corresponding weight: $\lambda_{\text{depth}}, \lambda_{\text{dist}}, \lambda_{\text{near}}, \lambda_{\text{empty}} = 1\text{e}{-4}$ and $\lambda_{\text{mask}} = 1\text{e}{-2}$.

4.8 Discussion on Dynamic Scene Representations

Relation to Tensor Decomposition. Equation 4 can be interpreted as a special case of a 4D tensor decomposition, similar to the vector-matrix decomposition introduced in TensoRF [Chen et al. 2022]. A spatio-temporal tensor $\mathcal{T} \in \mathbb{R}^{D_T \times D_X \times D_Y \times D_Z}$ representing a dynamic scene can be decomposed into a sum of four vector-tensor outer-products:

$$\mathcal{T} \approx \sum_{a \in A} \sum_{i=1}^{N} v_i^a \circ M_i^{A \setminus \{a\}}, \tag{13}$$

where $\circ$ denotes the vector-tensor outer product, $A = \{X, Y, Z, T\}$ is the set of axis indices, $v_i^a \in \mathbb{R}^{D_a}$ is a vector and $M_i^{A \setminus \{a\}}$ is a 3D tensor, for example $M_i^{A \setminus \{X\}} \in \mathbb{R}^{D_T \times D_Y \times D_Z}$.

Equation 4 of our method can be seen as a special case of Equation 13, where only the term for $a = T$ is used. Instead of storing features in a dense grid $M_i^{A \setminus \{a\}}$, we employ a memory efficient hash table representation $M_i^{A \setminus \{T\}} = \mathcal{H}_i$ and the vector v_i^a corresponds to our blend weights $v_{i,t}^a = \beta_{t,i}$.

Our final method deviates from this tensor decomposition perspective by employing a deformation field $\mathcal{D}$ before querying $\mathcal{H}_i$ (see Equation 6). This effectively aligns spatial features in the hashtables across timesteps by explaining parts of the motion with the deformation field.

Another way of achieving a 4D tensor decomposition is presented in concurrent works by Attal et al. [2023]; Cao and Johnson [2023]; Fridovich-Keil et al. [2023], who combine features from the 6 possible 2D feature planes instead of 4 outer-products between 1D and 3D tensors, as in Equation 13.

Relation to HyperNeRF. HyperNeRF [Park et al. 2021b] adds so-called *ambient dimensions* to their canonical space NeRFs to resolve topological issues that cannot be modeled by a deformation field. Instead of adding continuous ambient dimensions, our method models the canonical space with multiple hash grids, essentially introducing a *discrete* ambient dimension that serves a similar purpose.

5 EXPERIMENTAL RESULTS

We evaluate our method on the task of novel view synthesis (NVS) from multi-view video recordings on 10 diverse sequences from our dataset that focus on different aspects of facial and head movements. The validation sequences contain head rotations, laughs, eye blinking, talking, hair shaking, various mouth movements as well as one free expression sequence. All videos consist of 300-500 frames at 73 fps. We choose 12 out of the 16 available viewpoints as input and evaluate the NVS task on the remaining 4. The selected evaluation views are equally distributed across the camera setup, resulting in a challenging evaluation protocol due to the presence of extreme viewing angles (see Figure 7).

5.1 Data Preparation

Before running our experiments, we exploit the controlled nature of our dataset to facilitate reconstruction of the dynamic 3D scenes from image inputs. In concrete terms, we perform the following preprocessing steps:

Depth maps generation. We employ the standard COLMAP pipeline to obtain depth maps for each of the 12 training views [Schönberger and Frahm 2016; Schönberger et al. 2016]. To remove noisy depth measurements, we discard depth values observed by fewer than 3 cameras.

Background matting. We use Background Matting v2 [Lin et al. 2021] to obtain an alpha map given a captured frame and corresponding background image. To ensure the best quality, we use their ResNet101 [He et al. 2016] version and set the error threshold to 0.01 in the refinement stage.

Image downsampling. For all of our experiments, we downsample images by a factor of two to 1604×1100 pixels, which is sufficient for all methods. Temporally, we do not downsample and conduct all experiments on the full 73 fps.

Color correction. Despite the color calibration of our cameras, there can still be slight differences in brightness across views. To address this, we first use facial segmentation masks [Yu et al. 2018] to sample pixel values from the face, the torso, and the hair region. We then align the obtained color distributions across views by solving for an affine color transformation matrix using optimal transport [Flamary et al. 2021].

5.2 Floater Removal

Grid-based scene representations generally lack the induced smoothness prior of pure MLP architectures. As a result, they tend to generate small floaters that impair the visual quality of re-renderings. Since our hash ensemble is based on Instant NGP, it inherits this tendency. To address this, we specify tight-fitting, axis-aligned bounding boxes for each sequence and only reconstruct radiance fields inside. In addition to tight scene box fitting, which we make available to all baselines, NeRSemble employs the following two techniques to suppress floaters, which are ablated in Section 5.7.

View Frustum Culling. We exclude regions in space that are seen by less than 2 train cameras and are thus especially prone to produce floaters. These regions are neither queried during training nor inference.

Occupancy Grid Filtering. Before inference, we apply a low-pass filter to the density grid that our Instant NGP backbone tracks during training and only render within the largest connected component, effectively discarding small isolated islands of density.

5.3 Implementation Details

We implement our method in PyTorch [Paszke et al. 2019] within the Nerfstudio [Tancik* et al. 2022] framework, which uses the NerfAcc [Li et al. 2022c] implementation of Instant NGP.

We train all our models for 300k iterations using a warmup schedule of $E_{\text{init}} = E_{\text{trans}} = 40\text{k}$, which takes approximately one day on a single Nvidia RTX A6000. The inference of a single frame at a resolution of 1604×1100 pixels takes roughly 25 seconds.

We use a learning rate of $1e^{-3}$ for all model components, which is decayed by a factor of 0.8 every 20k iterations. For the deformation field $\mathcal{D}$, we use a factor of 0.5 instead, such that the learning rate is sufficiently low after the warm-up phase.

Furthermore, we use $N = 32$ hash tables, each configured with the default hyperparameters of Müller et al. [2022]. For our deformation field $\mathcal{D}$, we use the default configuration of the $SE(3)$ field by [Park et al. 2021a] and 128 dimensions for the learnable deformation codes $\boldsymbol{\omega}_t$. Our blend weights $\beta \in \mathbb{R}^N$ have one weight per hash table.

5.4 Baselines

We compare our method against several state-of-the-art methods for NVS of dynamic scenes. In particular, we compare against the following methods:

5.4.1 *Dynamic NeRFs.*

Nerfies [Park et al. 2021a] serves as representative for deformable NeRFs. We use the same implementation as for HyperNeRF, but without the ambient dimensions.

HyperNeRF [Park et al. 2021b] extends Nerfies to address topological issues. Due to memory issues with their official implementation, we port their code to the Nerfstudio framework and carefully choose

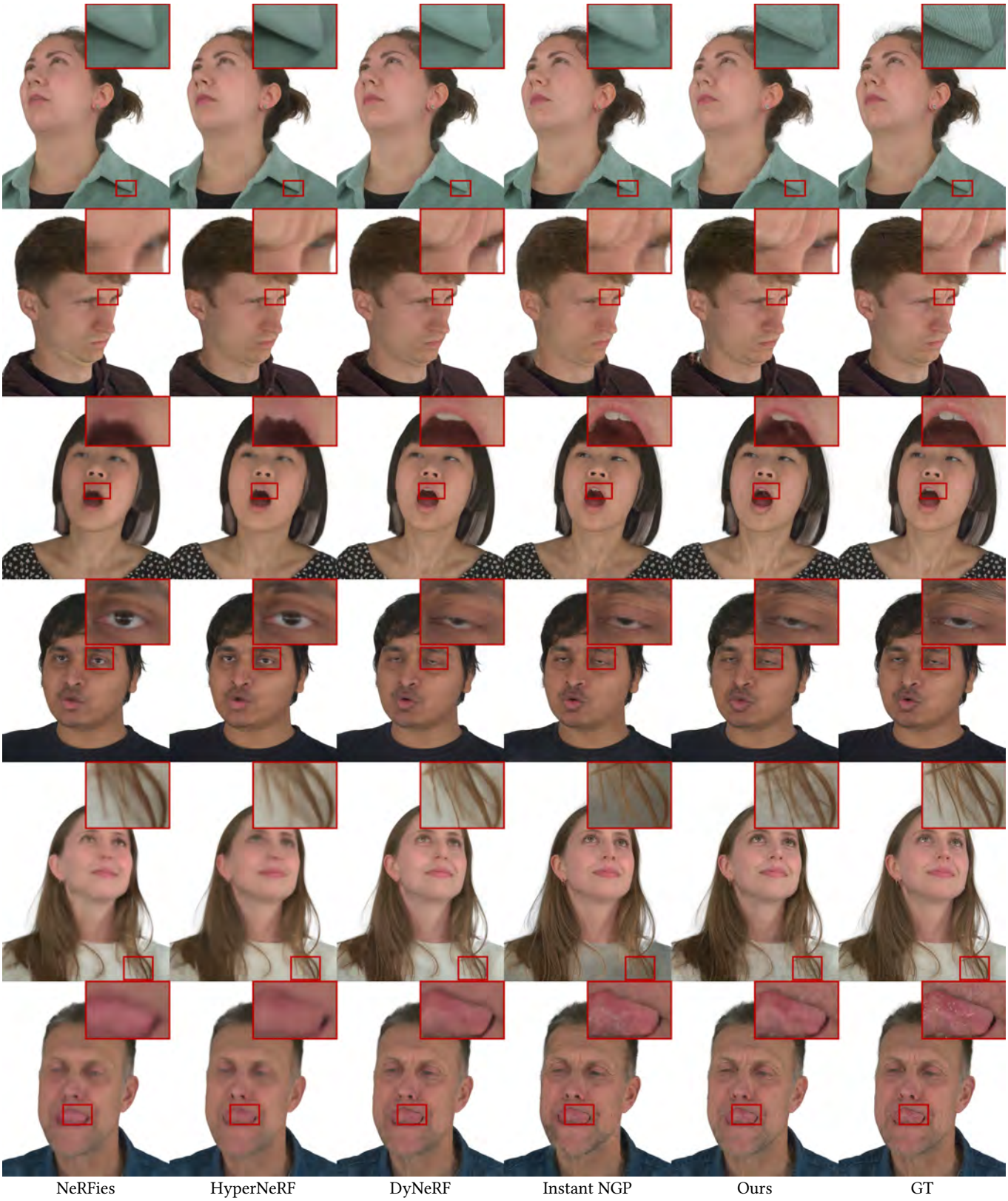

Fig. 6. **Qualitative results.** Our method reconstructs high-quality detail even for challenging expressions.

Fig. 7. **Spatial layout of our camera setup.** Marked in red are the 4 views used for evaluation. The point cloud reconstruction is obtained via COLMAP and used for additional depth supervision.

hyperparameters to match the performance of the official implementation.

DyNeRF [Li et al. 2022b], in contrast, is not constrained to represent dynamic content using a deformation field, but directly conditions a NeRF on a time-dependent latent code. Since no public code is available, we implement DyNeRF in Nerfstudio and finetune it to our data distribution.

5.4.2 Time-Agnostic Methods. Furthermore, the multi-view nature of our dataset allows for 3D reconstructions on a per-frame basis. Hence, we consider two additional baseline methods that do not consider time. First, we apply Poisson Surface Reconstruction (PSR) [Kazhdan et al. 2006] on the COLMAP point clouds. Second, we run the official Instant NGP [Müller et al. 2022] on each frame separately.

5.4.3 Face-Specific Methods. Additionally, we compare against Neural Head Avatars (NHA) [Grassal et al. 2022] and NeRFace [Gafni et al. 2021] as representatives of face-specific dynamic reconstruction methods, both rely on the geometric prior provided by tracked statistic mesh models. NHA is a mesh-based method that optimizes for vertex offsets on top of the FLAME and predicts view- and expression-dependent textures. NeRFace utilizes the 3DMM parameters directly to condition a NeRF and is thereby similar to DyNeRF. Since both methods were initially designed for monocular use-cases, we expand them to our multi-view scenario by employing a custom multi-view FLAME tracker and providing all 12 views during training. Note, that the reliance on a 3DMM provides both methods with a certain degree of reanimation ability, but potentially impairs rendering quality when provided with dense enough observations.

5.4.4 Background Modeling. For a fair comparison, we encourage all NeRF-based baselines to represent the background without density, by coloring all remaining transmittance as white. For this purpose, we use our alpha masks to set all background pixels in the ground truth images to white as well. In our experience, this simple technique allows all baselines to learn good reconstructions of the person in the foreground.

Table 3. **Quantitative evaluation.** We perform comparisons against two non-temporal baselines as well as three dynamic reconstruction methods. We evaluate unseen validation views of 10 diverse sequences from our dataset. Our method outperforms the baselines in all three metrics. The bottom two rows show ablations of our method with respect to core architectural components and the training procedure.

	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Static	PSR	12.5	0.774	0.341
	Instant NGP	28.8	0.864	0.254
Dynamic	Nerfies	29.5	0.849	0.299
	HyperNeRF	29.6	0.848	0.304
	DyNeRF	30.6	0.860	0.254
	Ours	**31.8**	**0.875**	**0.212**
Parts	NGP + Def.	30.8	0.864	0.231
	Hash Ensemble	30.5	0.857	0.257
Ablation	w/o Depth	31.5	0.873	0.217
	w/o Warmup	31.0	0.866	0.234
	only 16 tables	31.5	0.871	0.218

5.5 Evaluation Protocol

We evaluate all methods on 4 held-out camera viewpoints. Figure 7 shows the spatial arrangement of the evaluation cameras. Furthermore, in the interest of compute time, we only evaluate the prediction on 15 evenly distributed timesteps from each evaluation camera. We verified on multiple sequences that all employed image metrics differ by at most 0.02 points when evaluating only 15 timesteps instead of the full sequence.

Metrics. We report three image metrics to evaluate the visual quality of individual reconstructed frames: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [Wang et al. 2004], and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018]. All metrics are evaluated on a per-frame basis after blending predictions with the alpha masks in order to focus on the facial region. Additionally, we compute a JOD metric [Mantiuk et al. 2021] used by [Li et al. 2022b], which indicates perceptual difference to a reference video.

5.6 Comparison to State of the Art

Table 3 shows that NeRSemble quantitatively outperforms all baselines in all image metrics. In particular, our method shows strong improvements in the SSIM and LPIPS metrics that are sensitive to high-frequency details. This observation is matched by the qualitative comparison in Figure 6, where our method reconstructs better facial detail. We recommend the reader to watch the supplementary video for a more in-depth visual analysis of our method.

Evaluation of Temporal Consistency. Per-frame metrics such as PSNR, SSIM, and LPIPS do not account for temporal artifacts such as flickering. Hence, we employ the perceptual video metric JOD [Mantiuk et al. 2021] to measure visual similarity of a rendered video to its ground truth counterpart. For all major baselines, we render videos at a third of the training framerate, i.e. 24.3 fps, and average the JOD

Table 4. **Evaluation of temporal consistency** using the perceptual quality metric Just-Objectionable-Difference (JOD) [Mantiuk et al. 2021]. Higher numbers indicate less temporal flickering and a greater resemblance to the ground truth video.

Method	Inst. NGP	Nerfies	HyperNeRF	DyNeRF	Ours
JOD ↑	6.75	7.23	7.27	7.69	**7.86**

Table 5. **Evaluation against face-specific methods.** NeRSemble compares favorably to Neural Head Avatars [Grassal et al. 2022] and NeRFace [Gafni et al. 2021].

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Neural Head Avatars	31.0	0.927	0.041
NeRFace	35.2	0.956	0.047
Ours	**37.5**	**0.968**	**0.023**

Table 6. **Quantitative comparison on the Neural 3D Video Dataset.** Although NeRSemble's functionality is inspired by facial blendshapes, it can also reasonably model generic dynamic scenes.

Method	NeRFPlayer (Instant NGP)	NeRFPlayer (TensoRF)	HyperReel	Ours
PSNR ↑	30.3	30.7	**31.1**	29.9

scores over all validation views and all 10 validation sequences. The results of this temporal evaluation are given in Table 4. Note, that the Instant NGP baseline is completely time-agnostic, which leads to considerable flickering artifacts in video renderings. Figure 8 shows an example of such an artifact. In contrast, NeRSemble provides a smooth temporal experience.

Comparison to Instant NGP. The Instant NGP baseline produces compelling images on a per-frame basis as can be seen in Figure 6. However, it suffers from a strong tendency to generate floaters and scattered surfaces due to the sparse nature of our camera setup. In contrast, NeRSemble constrains the space across multiple timesteps which greatly contributes towards removing floaters. This also holds for a NeRSemble trained without any anti-floater strategies or additional losses. Such a bare-bones version of our model still outperforms Instant NGP (see the top row in Table 7). This shows that in our sparse setting, having higher expressiveness by modeling each frame independently (e.g., the Instant NGP baseline has 10-15 times more parameters than our model) does not lead to better reconstructions.

Comparison to Face-Specific Methods. To compare against NHA and NeRFace, we evaluate on 7 sequences from our dataset, excluding 3 with more complex motion where the preprocessing pipeline of NHA failed to predict facial landmarks, segmentation masks and normals. Furthermore, NHA only synthesizes the head without a torso. Therefore, we only evaluate the facial region for a fair comparison. Table 5 shows that NeRSemble outperforms both baselines despite them being specifically designed for faces.

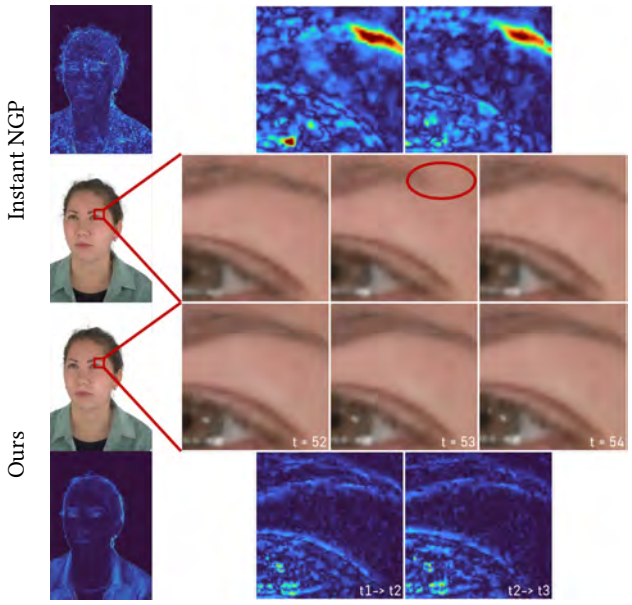

Fig. 8. **Temporal consistency.** We show a re-rendering and its temporal difference image for a novel view (left). On the right side, we demonstrate the flickering artifact of the Instant NGP baseline between three adjacent frames, where an eyebrow shrinks and grows between frames. In comparison, NeRSemble offers more temporal consistency.

Comparison on Neural 3D Video Dataset [Li et al. 2022b]. NeRSemble does not make strong assumptions on the content of a dynamic scene and is therefore applicable to more general scenarios. To demonstrate NeRSemble's generality, we evaluate on the 6 publicly available sequences of the Neural 3D Video dataset [Li et al. 2022b]. We follow the evaluation protocol of NeRFPlayer [Song et al. 2023] and HyperReel [Attal et al. 2023] that downsample the recordings to 1352 x 1014 pixel resolution, hold out the top central view for evaluation, and report metrics averaged over all 6 sequences. We further re-compute the poses with COLMAP [Schönberger et al. 2016] as the ones provided with the dataset are slightly off. Table 6 shows the quantitative results. We excluded the original DyNeRF [Li et al. 2022b] as well as StreamRF [Li et al. 2022a] from the evaluation as their numbers were only computed on 1 of the 6 available sequences and are thus not comparable to the results of NeRFPlayer and HyperReel. The evaluation shows that NeRSemble can reasonably model generic dynamic scenes despite its functionality being inspired by facial blendshapes. However, since our method relies on Instant NGP, it also inherits some of its weaknesses. In particular, it does not model light refraction as HyperReel does. As a result, NeRSemble cannot perfectly capture the effects of window panes and glass bottles which are prevalent in the Neural 3D Video dataset.

5.7 Ablations

In addition to the comparison against baselines, we conduct several experiments to validate our design choices and understand the inner workings of NeRSemble.

(a) NGP + Def. (b) Hash Ensemble (c) w/o Warmup (d) 16 Hash Tables (e) Ours (f) GT

Fig. 9. **Ablation of our model components.** Combining Instant NGP with a deformation field (a) produces sharp detail in rigidly moving areas of the scene, e.g., the torso, but struggles with more challenging motion such as mouth movements. On the other hand, employing an ensemble of hash encodings (b) can better deal with complex motions but generally produces more blurry reconstructions. Combining both components (c) leverages the strength of both architectures but still does not produce the same detail in rigidly moving areas as an Instant NGP with deformation field. By employing a warmup phase, sharp detail already returns when 16 tables are used (d) which can be further improved by increasing the number of hash encodings (e).

Contribution of Architectural Components. We ablate the effect of using a hash ensemble and the deformation field. Table 3 shows that neither a deformation field with an Instant NGP backbone (NGP + Def.) nor a plain hash ensemble matches the performance of our final model. However, both architectures are strong baselines on their own. In Figure 9, we present qualitative results, which show that the deformation-based approach generally produces sharper reconstructions, but struggles with more challenging motions that are difficult to model with deformations. On the other hand, the hash ensemble has the expressiveness to model any dynamic scene via feature blending but will typically produce more blurry results for simple movements, since it is missing the prior of a deformation field. The quantitative results in Table 3 confirm these findings, with the hash ensemble scoring a high PSNR but worse LPIPS value.

Number of Hash Tables. NeRSemble with 16 hash tables only suffers a negligible amount compared to 32 hash tables. This confirms that the ratio between the number of frames and hash tables scales well and information is shared effectively across tables.

Effect of Warm-Up Phase. Training without warm-up consistently performs worse. We attribute this to the fact that giving the model access to all hash grids right away prevents it from learning correspondences with the deformation field. As a result, the learned hash encodings are less well-aligned and cannot be blended as effectively. Visually, this manifests in slightly blurrier renderings. This insight is in line with HyperNeRF's proposal to disable the use of ambient dimensions in the beginning.

Table 7. **Ablation of floater removal techniques.** Both view frustum culling (VFC) and occupancy grid filtering (OGF) have a negligible effect on the metrics as they mostly remove floaters in areas that are omitted in our evaluation protocol. Note that a plain version of NeRSemble without any additional losses ($\mathcal{L}$) or floater removal techniques already performs competitively compared to all baselines in Table 3.

$\mathcal{L}$	VFC	OGF	PSNR ↑	SSIM ↑	LPIPS ↓
☐	☐	☐	30.4	0.868	0.230
☒	☐	☐	31.8	**0.875**	0.213
☒	☒	☐	31.8	**0.875**	0.213
☒	☐	☒	**31.9**	**0.875**	**0.212**
☒	☒	☒	31.8	**0.875**	**0.212**

Effect of Depth Supervision. Since removing the depth supervision only slightly impairs the performance, we hypothesize that the RGB information of the 12 input views already sufficiently supervises the geometry. However, exploiting depth supervision from orthogonal channels, such as a fitted 3DMM or a trained depth prediction network, could still be beneficial as it incorporates data priors from sources other than the RGB video frames.

Floater Removal Techniques. We ablate the effect of three strategies to suppress floaters. First, we isolate the effect of all additional losses, i.e. mask loss, depth supervision, and distortion loss, and note their significant impact on performance in Table 7. View frustum culling and occupancy grid filtering, on the other hand, do not affect the reported metrics but still improve visual quality when rendering novel camera trajectories.

Fig. 10. **Blend weights.** Investigating the contents of the first hash grid by setting $\beta_{t_1,i} = 0$ $(i > 1)$ reveals that the first hash grid stores some sort of average representation (left). On the right we successively set $\beta_{t_1,i} = 0.75$ for $i \in \{2, 3, 4\}$. Each table stores additional details that are exceeding the representational capacity of the deformation network. Note that we use ω_{t_1} for all shown examples and t_1 denotes the first frame.

Content of Individual Hash Grids. We analyze the contents of the individual hash grids $\mathcal{H}_i$ in Figure 10. For this purpose, we modify the learned blend weights $\beta_{t_1,i}$ $(i > 1)$ for the first frame t_1 of a sequence, while keeping $\beta_{t_1,1}$ and the deformation codes ω_{t_1} fixed. This experiment reveals that the deformation field $\mathcal{D}$ accounts for rigid movements of the scene, since modifying β_{t_1} results in well-aligned appearance changes while the head stays static. Furthermore, $\mathcal{H}_1$ seems to store a representation comparable to the mean face of the person. The remaining hash grids then behave similarly to a dynamic, volumetric texture that further adds details to the scene that are otherwise unexplained, e.g., topologically complicated deformations, expressions-dependent wrinkles, or illumination changes. We attribute the special status of the first hash grid $\mathcal{H}_1$ to the fact that it is always active during training while all other hash grids are gradually introduced during the warmup phase.

5.8 Limitations

In our experiments, we demonstrate that we can achieve convincing results with a sparse set of multi-view recordings; however, various limitations remain. Since NeRSemble models explicit correspondences across timesteps via a deformation field, it cannot perfectly capture fast hair motion (see Figure 11c). To address this, incorporating movement priors via optical flow or differentiable physics could be an interesting field for future work.
Furthermore, our method currently focuses on recovering the appearance and motion of a specific sequence by optimizing for the dynamic radiance field representation. As a result, our method is unable to learn priors that generalize across sequences. Here, we see great potential for future work on dynamic NeRFs that generalize over both identities and facial expressions. A learned prior over the distribution of realistic 4D avatars could help to further constrain the optimization procedure. This would be particularly important for monocular inputs or capturing facial regions, such as the mouth interior, that are often occluded during the majority of recording time and may thus exhibit inferior reconstruction quality (see Figure 11).

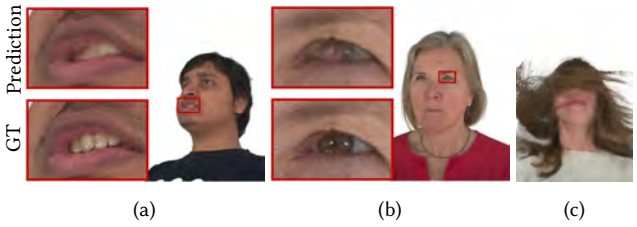

Fig. 11. **Failure cases.** The high degree of occlusion of the mouth interior can sometimes cause a hollow face illusion where teeth are falsely reconstructed at the back of the mouth (a). Specular reflections of the light sources in the eyes may cause rare eye artifacts (b). The deformation field may fail to model extremely fast hair motion, which hinders the canonical hash grids from synthesizing a sharp result for some frames (c).

6 CONCLUSION

In this work, we have proposed a new method and dataset focusing on the radiance field reconstruction of animated human heads from multi-view video inputs. To this end, we have introduced a novel multi-view video benchmark of diverse human heads containing over 220 identities with 4700 sequences. We further proposed a new method for generating photo-realistic re-renderings of arbitrary viewpoints and time steps, and hope that our dataset and accompanying benchmark will be an important contribution to the community, and facilitate future work on digital humans.

Our proposed novel representation for spatio-temporal NeRFs uses deformation fields to factor out coarse movements and an ensemble of hash grid encodings to model fine deformations and increase the temporal capacity of our model. Our experiments demonstrate that NeRSemble achieves temporally coherent and highly detailed volumetric reconstructions from multi-view video inputs, outperforming existing baselines by a significant margin, in particular when sequences contain complex motions.

ACKNOWLEDGMENTS

This work was supported by the ERC Starting Grant Scan2CAD (804724), the German Research Foundation (DFG) Grant “Making Machine Learning on Static and Dynamic 3D Data Practical”, and the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing”. We would also like to thank Maximilian Knörl for the help with data acquisition, and Angela Dai for the video voice-over.

REFERENCES

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully Controllable Neural 3D Portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20364–20373.

Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. 2023. HyperReel: High-Fidelity 6-DoF Video with Ray-Conditioned Sampling. *CVPR* (2023).

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR* (2022).

V. Blanz, C. Basso, T. Poggio, and T. Vetter. 2003. Reanimating Faces in Images and Video. (2003), 641–650.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., USA, 187–194. https://doi.org/10.1145/311535.311556

Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 86–1.

Ang Cao and Justin Johnson. 2023. HexPlane: A Fast Representation for Dynamic Scenes. *CVPR* (2023).

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *European Conference on Computer Vision (ECCV)*.

Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2018. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5117–5126.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–13.

Darren Cosker, Eva Krumhuber, and Adrian Hilton. 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *2011 international conference on computer vision*. IEEE, 2296–2303.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10101–10111.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research* 22, 78 (2021), 1–8. http://jmlr.org/papers/v22/20-451.html

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *CVPR*.

Fridovich-Keil and Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *CVPR*.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 41, 6 (2022). https://doi.org/10.1145/3550454.3555501

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2022. Learning Neural Parametric Head Models.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. 2022. Relu fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. 2006. Poisson Surface Reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing* (Cagliari, Sardinia, Italy) *(SGP '06, Vol. 256)*, Alla Sheffer and Konrad Polthier (Eds.). Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 61–70. http://dl.acm.org/citation.cfm?id=1281957.1281965

Lingzhi Li, Zhen Shen, Li Shen, Ping Tan, et al. 2022a. Streaming Radiance Fields for 3D Video Synthesis. In *Advances in Neural Information Processing Systems*.

Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. 2022c. NerfAcc: A General NeRF Accleration Toolbox. *arXiv preprint arXiv:2210.04847* (2022).

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022b. Neural 3D Video Synthesis From Multi-View Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5521–5531.

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

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (July 2019), 14 pages.

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

Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: A Visible Difference Predictor for Wide Field-of-View Video. *ACM Trans. Graph.* 40, 4, Article 49 (jul 2021), 19 pages. https://doi.org/10.1145/3450626.3459831

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. *ICCV* (2021).

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2020. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. 2022. Urban Radiance Fields. *CVPR* (2022).

Neus Sabater, Guillaume Boisson, Benoit Vandame, Paul Kerbiriou, Frederic Babon, Matthieu Hog, Tristan Langlois, Remy Gendrot, Olivier Bureller, Arno Schubert, and Valerie Allie. 2017. Dataset and Pipeline for Multi-View Light-Field Video. In *CVPR Workshops*.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. 2023. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2732–2742.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. In *CVPR*.

Matthew Tancik*, Ethan Weber*, Evonne Ng*, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. 2022. Nerfstudio: A Framework for Neural Radiance Field Development. https://github.com/nerfstudio-project/nerfstudio

Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022a. MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 55, 9 pages. https://doi.org/10.1145/3528233.3530753

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *ECCV*.

Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. 2022b. Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13524–13534.

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

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. 2022. Multiface: A Dataset for Neural Face Rendering. In *arXiv*. https://doi.org/10.48550/ARXIV.2207.11243

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. 2022. Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum* (2022). https://doi.org/10.1111/cgf.14505

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3DMM: Deep Implicit 3D Morphable Model of Human Heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12803–12813.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 325–341.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In *Computer Vision and Pattern Recognition (CVPR)*.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant Volumetric Head Avatars. *CVPR* (2023).

C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)* 23, 3 (2004), 600–608.