Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction

Guy Gafni[1]   Justus Thies[1]   Michael Zollhöfer[2]   Matthias Nießner[1]
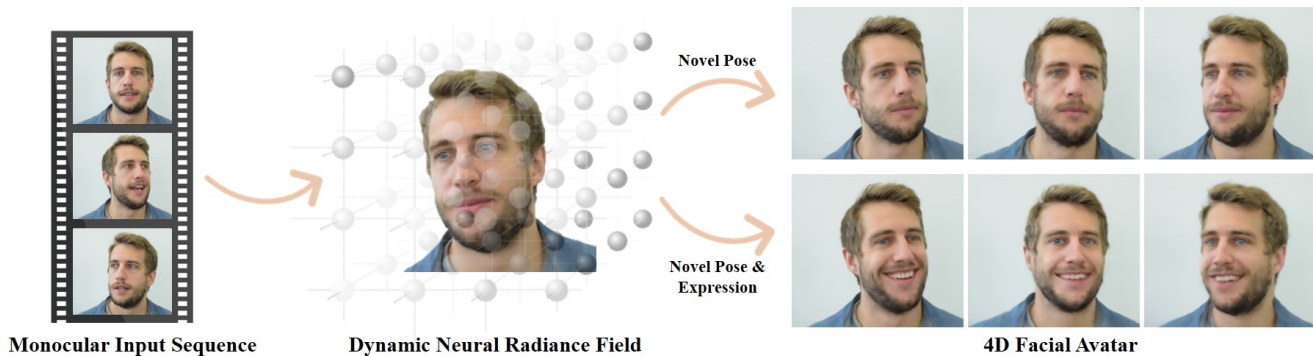[1]Technical University of Munich   [2]Facebook Reality Labs Research

Figure 1: Given a monocular portrait video sequence of a person, we reconstruct a dynamic neural radiance field representing a 4D facial avatar, which allows us to synthesize novel head poses as well as changes in facial expressions.

Abstract

We present dynamic neural radiance fields for modeling the appearance and dynamics of a human face[1]. Digitally modeling and reconstructing a talking human is a key building-block for a variety of applications. Especially, for telepresence applications in AR or VR, a faithful reproduction of the appearance including novel viewpoint or head-poses is required. In contrast to state-of-the-art approaches that model the geometry and material properties explicitly, or are purely image-based, we introduce an implicit representation of the head based on scene representation networks. To handle the dynamics of the face, we combine our scene representation network with a low-dimensional morphable model which provides explicit control over pose and expressions. We use volumetric rendering to generate images from this hybrid representation and demonstrate that such a dynamic neural scene representation can be learned from monocular input data only, without the need of a specialized capture setup. In our experiments, we show that this learned volumetric representation allows for photorealistic image generation that surpasses the quality of state-of-the-art video-based reenactment methods.

---

[1]gafniguy.github.io/4D-Facial-Avatars

1. Introduction

Reconstructing 4D models of humans and, especially, the human face, is an ongoing research problem in the field of computer vision and computer graphics. 4D avatars are essential for augmented reality (AR) and virtual reality (VR) telepresence applications as well as for video editing, such as visual dubbing in movie productions. These applications need a faithful reconstruction of the human's appearance, as well as the ability to change the viewpoint or head pose (especially, in VR) and the expressions (e.g., for visual dubbing). Representing a human head with explicit geometry and material properties (e.g., albedo, reflectance) is challenging; the skin has effects like subsurface scattering, the eyes are highly reflective and the hair has a complex geometry with fine scale details. While the explicit reconstruction of high quality geometry of the skin surface in a multi-view studio setup is tractable [4, 16, 52], hair is often approximated by retrieval and refinement of hair styles [18, 50], which leads to an unrealistic visual reproduction.

To handle the material properties and complex geometry of a 4D facial avatar, we introduce *dynamic neural radiance fields*. Our approach is a neural rendering method combining classical volume rendering with a novel neural

scene representation network to achieve novel head pose and expression synthesis. In contrast to related work on learned scene representations that focuses on static objects and multi-view input data, we are able to represent the dynamically changing surface of a human's face only based on *monocular* camera recordings. The representation is a stepping stone towards reconstruction of 4D facial avatars using commodity hardware, allowing for novel viewpoint synthesis of the head in a virtual reality setting, pose changes in videos or even facial reenactment where the expressions of one person are transferred to another person (represented by our scene representation network). The learned scene representation is a volumetric representation which is key to capturing hair, but also the mouth interior where classical methods struggle because of missing 3D geometry. The implicit representation of the geometry and appearance defines a continuous function in space that does not suffer from discretization artifacts of voxel grids (e.g., limited resolution) and is optimized to represent the head as good as possible w.r.t. the final re-renderings and the underlying network architecture. In contrast to state-of-the-art facial reenactment and video editing approaches [21, 39], our volumetric approach is able to synthesize 3D-consistent content with large head pose changes. Large head pose changes (or view changes) are required for VR or AR applications, but can also be used for face frontalization or to dampen the variance of motion. The semantically meaningful conditioning used in our method also allows for user-driven edits of a video in a post-processing scenario.

Specifically, our method is based on a short portrait video sequence of a person. To represent the expressions of the face, we leverage a low-dimensional morphable model [7, 41]. Given the pose of the model and the expression parameters of a specific frame of a sequence that has to be synthesized, we dispatch rays in a canonical space where our neural scene representation network is embedded. Along the rays, we perform volumetric integration of density and color values predicted by our scene representation network that is inspired by the work of Mildenhall *et al.* [28], which focuses on high quality multi-view reconstruction of a static scene. Note that the scene representation network is not only conditioned on the sample point locations but also on the expressions of the morphable model which allows for the dynamically changing content that has to be stored in the neural network. During test time, this conditioning allows us to apply novel head poses as well as expressions to synthesize a new image. We demonstrate that our technique is able to faithfully represent a 4D facial avatar and show photorealistic results that surpass state-of-the-art facial reenactment methods.

To summarize, we show that neural scene representation networks can be used to store and represent the dynamically changing surface of a human head in a controllable manner.

Our contributions are:

- Dynamic Neural Radiance Fields to represent 4D facial avatars based on a low dimensional morphable model.

- An efficient end-to-end learnable approach that uses a single camera to reconstruct such a radiance field.

2. Related Work

Our approach is a neural rendering method to represent and generate images of a human head. It is related to recent approaches on neural scene representation networks, as well as neural rendering methods for human portrait video synthesis and facial avatar reconstruction. In the following, we discuss the most related literature in the two fields in detail.

Face Reconstruction based on a Morphable Model For a summary of facial reconstruction methods, we refer to the state-of-the-art report of Zollhöfer *et al.* [52]. Our method is built upon a low-dimensional morphable model [7, 41] which is a building block of numerous facial reconstruction and animation approaches [14, 15, 46, 45, 40, 41, 5, 6, 21, 39, 38]. In contrast to these methods, we are not relying on the coarse representation of the surface of the face. Some methods [9, 23, 8, 17] also focused on corrective shapes [8], dynamically adapting the blendshape basis [23] or applied non-rigid mesh deformation [9] to compensate for the coarse geometry of the morphable model. In our approach, we are not relying on a template mesh or an explicit surface representation. Instead, we represent the geometry and appearance implicitly using a deep neural network and use volumetric rendering to generate new images.

Human Avatar Reconstruction The goal of our approach is the photorealistic reproduction of the head of a human observed from a monocular input stream. Multiple methods exist that reconstruct personalized face rigs based on hand-held monocular input. Ichim *et al.* [20] assume a static pose and expression to reconstruct the head via multi-view stereo. Hu *et al.* [19] combine face digitization and hair reconstruction to estimate the head geometry and appearance from a single image. Our implicit function represents the face region as well as the hair in a single formulation, also recovering the volumetric effects of the hair.

Human Portrait Video Synthesis There is a wide range of human portrait video synthesis and editing approaches. Classical computer graphics approaches use a morphable model reconstruction and forward rendering with optimized textures and a texture atlas for different mouth interiors (since the morphable model is too coarse to model the mouth cavity) [15, 14, 41, 43, 42]. Image warping is
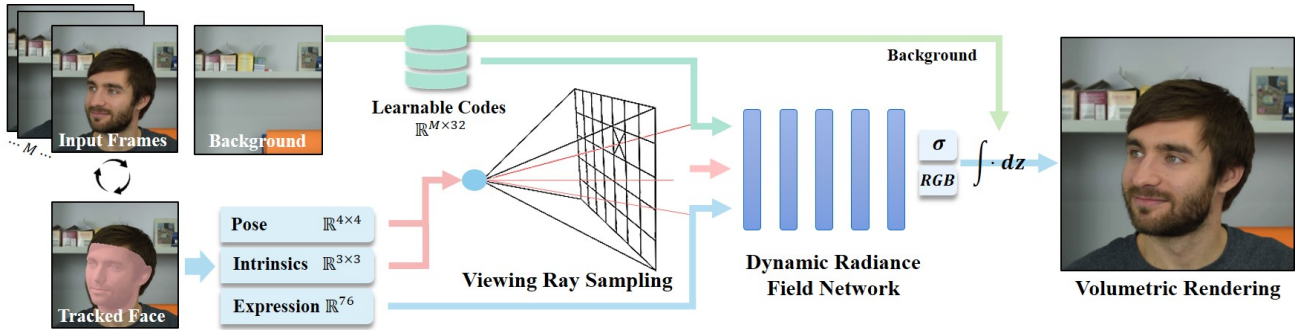
Figure 2: Overview of our 4D facial avatar reconstruction pipeline. Given a portrait video and an image without the person (background image) as input, we apply facial expression tracking using a 3D morphable model. Based on the estimated pose and expression, we use volumetric rendering to synthesize the image of the face. The samples along the viewing rays are input to our dynamic radiance field, which is additionally conditioned on a learnable per-frame latent code. Since the background is static, we set the color of the last sample point of each ray to the corresponding value of the background image.

used in Averbuch-Elor *et al.* [3]. In contrast, the most recent approaches are hybrids between classical rendering and learned image synthesis. Deep Video Portraits [21] is one of the first methods that uses rendered correspondence maps together with an image-to-image translation network to output photorealistic imagery. Deferred Neural Rendering [39, 38] extends this idea, by introducing neural feature descriptors that are embedded on the surface of a coarse reconstructed face mesh. Instead of this dense conditioning input or rendered feature maps, there are also methods that work on rendered facial landmarks [48, 10, 47]. These approaches can also be applied to single images. First Order Motion Model [32] is a data-driven approach that decouples appearance and motion in a video of a specific class (e.g., human faces) and allows application of the motion in a source video to a target image.

Neural Scene Representation Networks Neural scene representation networks are building blocks of several neural rendering and neural reconstruction approaches. A summary of neural rendering approaches is given in the state-of-the-art report of Tewari *et al.* [37]. Sitzmann *et al.* [35] introduced neural scene representation networks (SRNs). The geometry and appearance of an object is represented as a neural network that can be sampled at points in space. A ray marching approach is used to sample from the neural network to render the reconstructed surface. On synthetic data, they show the capabilities of such an implicit representation. A neural scene representation network is a compact representation that does not suffer from limited resolution as for example, discrete grid structures that store learnable features, e.g., Deep Voxels [34] or Neural Volumes [26]. Mildenhall *et al.* [28] extend this idea to store radiance fields in a neural network. They assume a static object and multi-view data. A key contribution is the volumetric in-

tegration and the usage of positional encoding for higher detailed reconstructions. Follow-up work extends this idea by using different positional encodings [36] and in-the-wild training data including appearance interpolation [27]. Concurrent work of Sitzmann *et al.* [33] proposes the usage of sinusoidal activation functions for the scene representation network. Neural Sparse Voxel Fields [25] employ an Octree to cull empty space and speed up rendering. While these methods have a focus on static objects, we are dealing with a dynamically changing surface of a face. We use a similar volumetric integration scheme to [28] with an additional layer for the static background. The dynamic neural scene representation is not only conditioned on the sample position and view direction, but also on the facial deformations.

3. Method

Our approach enables 4D reconstruction of a facial avatar based on a single portrait video of a person (see Fig. 2). The geometry and appearance of the human head is represented implicitly by a neural scene representation network. Specifically, the neural scene representation network stores a dynamic neural radiance field which is used during volumetric rendering. The dynamics of the human face, i.e., the facial expressions, are first captured with a state-of-the-art face tracking approach [41]. The resulting low dimensional expression parameters of the morphable model are used as conditioning for the neural scene representation network. Note that the expression parameters have semantic meaning allowing us to change specific expressions (see Fig. 3) or to apply the expressions of a different recorded person (see Fig. 7). In addition, we employ the pose parameters (rotation, translation) of the face tracking to transform the rays into a canonical space that is shared by all frames.

3.1. Dynamic Neural Radiance Fields

We represent the dynamic radiance field of a talking human head using a multi-layer perceptron (MLP) $\mathcal{D}_\theta$ that is embedded in a canonical space. As the dynamic radiance field is a function of position $\mathbf{p}$, view $\vec{v}$ and dynamics in terms of facial expressions δ, we provide these inputs to the MLP which outputs color as well as density values for volumetric rendering:

$$\mathcal{D}_\theta(\mathbf{p}, \vec{v}, \delta, \gamma) = (RGB, \sigma) \tag{1}$$

Note, to compensate for errors in the facial expression and pose estimation, we also provide a per-frame learnable latent code γ to the MLP. Instead of directly inputting the canonical position $\mathbf{p}$ and viewing direction $\vec{v}$, we use positional encoding as introduced by Mildenhall *et al.* [28]. In our experiments, we use 10 frequencies for the position $\mathbf{p}$ and 4 frequencies for the viewing direction $\vec{v}$.

Dynamics Conditioning A key component of the dynamic neural radiance fields is the conditioning on the dynamically changing facial expressions. The facial expressions δ are represented by coefficients of a low dimensional delta-blendshape basis of a morphable model ($\delta \in \mathbb{R}^{76}$). To estimate the per-frame expressions δ_i, we use an optimization-based facial reconstruction and tracking pipeline [41]. Note that these expression vectors only model the coarse geometric surface changes and do not model changes of for example the eye orientation. Besides expression parameters, we also store the rigid pose $P_i \in \mathbb{R}^{4 \times 4}$ of the face which allows us to transform camera space points to points in the canonical space of the head.

To compensate for missing information of the expression vectors, we introduce learnable latent codes γ_i (one for each frame). In the experiments, we are using $\gamma_i \in \mathbb{R}^{32}$ and regularize them via an ℓ_2 loss using weight decay ($\lambda = 0.05$). In Fig. 4, we show that the latent code improves the overall sharpness of the reconstruction. Evaluating the Learned Perceptual Image Patch Similarity (LPIPS) [49] metric for the generated images with and without latent codes results in 0.059 and 0.068, respectively.

3.2. Volumetric Rendering of Portrait Videos

In our experiments, we assume a static camera, and a static background. The moving and talking human in the training portrait video is represented with a dynamic neural radiance field as introduced in the previous section. To render images of this implicit geometry and appearance representation, we use volumetric rendering. We cast rays through each individual pixel of a frame, and accumulate the sampled density and RGB values along the rays to compute the final output color. Using the tracking information P of the morphable model, we transform the ray sample
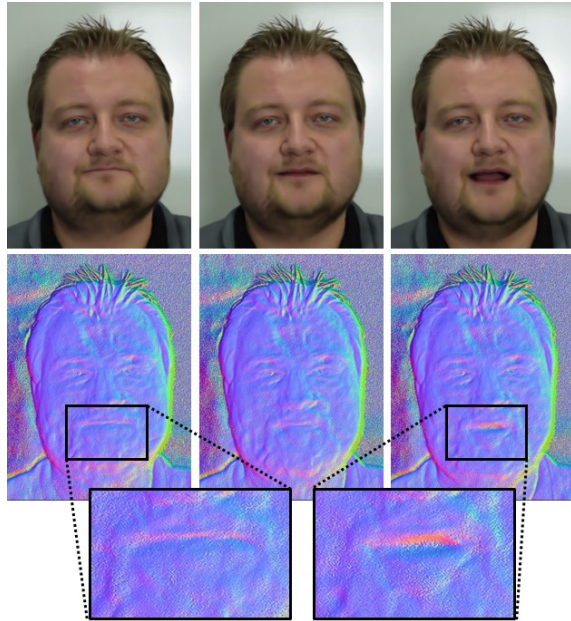

Figure 3: Our dynamic radiance field allows for manual editing via the expression vector δ. In the middle we show the reconstruction of the original expression. On the left and right we show the results of modifying the blendshape coefficient of the mouth opening (left -0.4, right $+0.4$). The bottom row shows the corresponding normal maps computed via the predicted depth. As can be seen, the dynamic radiance field adapts not only the appearance, but also the geometry according to the input expression.

points to the canonical space of the head model and evaluate the dynamic neural radiance field at these locations. Note that this transformation matrix P gives us the control over the head pose during test time.

We use a similar two-stage volumetric integration approach to Mildenhall *et al.* [28]. Specifically, we have two instances of the dynamic neural radiance field network, a coarse and a fine one. The densities predicted by the coarse network are used for importance sampling of the query points for the fine network, such that areas of high density are sampled more. The expected color $\mathcal{C}$ of a camera ray $\mathbf{r}(t) = \mathbf{c} + t\vec{\mathbf{d}}$ with camera center $\mathbf{c}$, viewing direction $\vec{\mathbf{d}}$ and near z_{near} and far bounds z_{far} is evaluated as:

$$\mathcal{C}(\mathbf{r}; \theta, P, \delta, \gamma) = \int_{z_{\text{near}}}^{z_{\text{far}}} \sigma_\theta\left(\mathbf{r}(t)\right) \cdot \text{RGB}_\theta\left(\mathbf{r}(t), \vec{\mathbf{d}}\right) \cdot T(t)dt, \tag{2}$$

where $RGB_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are computed via the neural scene representation network $\mathcal{D}_\theta$ at a certain point on the ray with head pose P, expressions δ and learnable latent code γ. $T(t)$ is the accumulated transmittance along the ray
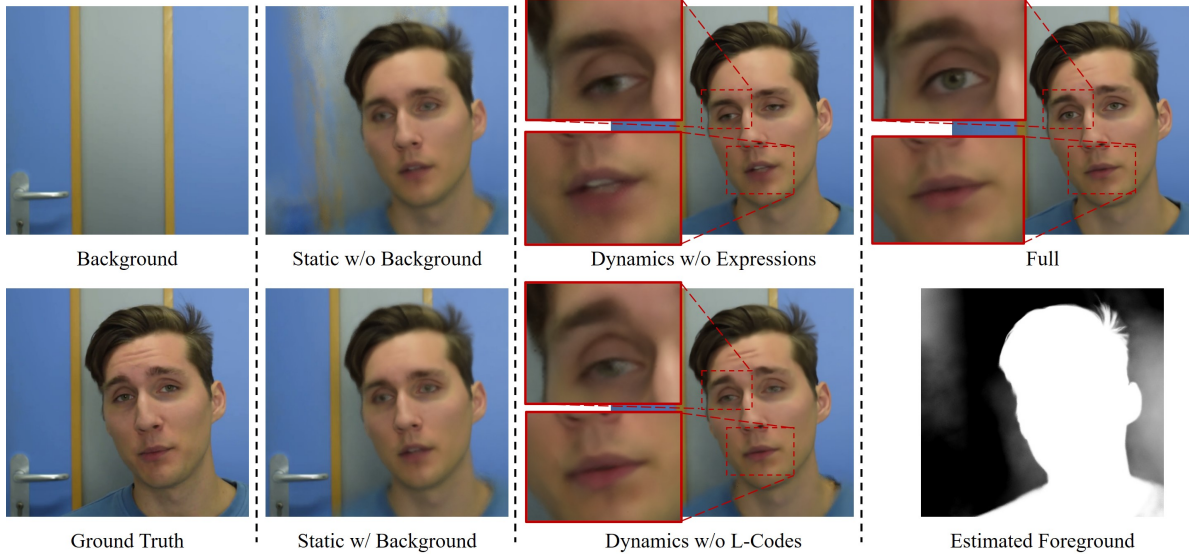
Figure 4: The background image enables us to faithfully reproduce the entire image. While the dynamics are mainly conditioned on the facial expressions, the learnable latent codes improve the sharpness of the image significantly. Our method also implicitly gives access to a foreground segmentation. Note that the shown images are from the test set (latent code is taken from the first frame of training set).

| Background | Static w/o Background | Dynamics w/o Expressions | Full |
| Ground Truth | Static w/ Background | Dynamics w/o L-Codes | Estimated Foreground |

from z_{near} to t:

$$T(t) = \exp\left(-\int_{z_{\text{near}}}^{t} \sigma_\theta\left(\mathbf{r}\left(s\right)\right) ds\right). \quad (3)$$

Note that the expected color is evaluated for both the coarse and the fine networks (with learnable weights θ_{coarse} and θ_{fine}, respectively) to compute corresponding reconstruction losses at train time (see Eq. 4).

We decouple the static background and the dynamically changing foreground by leveraging a single capture of the background $\mathcal{B}$ (i.e., without the person). The last sample on the ray $\mathbf{r}$ is assumed to lie on the background with a fixed color, namely, the color of the pixel corresponding to the ray, from the background image. Since the volumetric rendering is fully differentiable, the network picks up on this signal, and learns to predict low density values for the foreground samples if the ray is passing through a background pixel, and vice versa - for foreground pixels, i.e., pixels that correspond to torso and head geometry, the networks predict higher densities, effectively ignoring the background image. This way, the network learns a foreground-background decomposition in a self-supervised manner (see Fig. 4).

3.3. Network Architecture and Training

As mentioned above, the dynamic neural radiance field is represented as an MLP. Specifically, we pass the positional encoding of the query points through a backbone of 8 fully-connected layers, each 256 neurons-wide, followed by ReLu activation functions. Past the backbone, the activations are fed through a single layer to predict the density

value, as well as a 4-layer, 128 neuron-wide branch, conditioned on the viewing direction, to predict the final color value of the query point.

We optimize the network weights of both the coarse and the fine networks based on a photometric reconstruction error metric over the training images I_i ($i \in [1, M]$):

$$L_{total} = \sum_{i=1}^{M} L_i(\theta_{coarse}) + L_i(\theta_{fine}) \quad (4)$$

with

$$L_i(\theta) = \sum_{j \in \text{pixels}} \left\|\mathcal{C}\left(\mathbf{r}_j; \theta, P_i, \delta_i, \gamma_i\right) - I_i[j]\right\|^2. \quad (5)$$

For each training image I_i and training iteration, we sample a batch of 2048 viewing rays through the image pixels. We use a bounding box of the head (given by the morphable model) to sample the rays such that 95% of them correspond to pixels within the bounding box and, thus allowing us to reconstruct the face with a high fidelity. Stratified sampling is used to sample 64 points along each ray, which are fed into the coarse network $\mathcal{D}_{\theta_{coarse}}$. Based on the density distribution along the ray, we re-sample 64 points and evaluate the color integration (see Eq. 2) using the fine network $\mathcal{D}_{\theta_{fine}}$. Our method is implemented in PyTorch [29]. Both networks and the learnable codes γ_i are optimized using Adam [22] ($lr = 0.0005$). In our experiments, we use 512×512 images and train each model for $400k$ iterations.
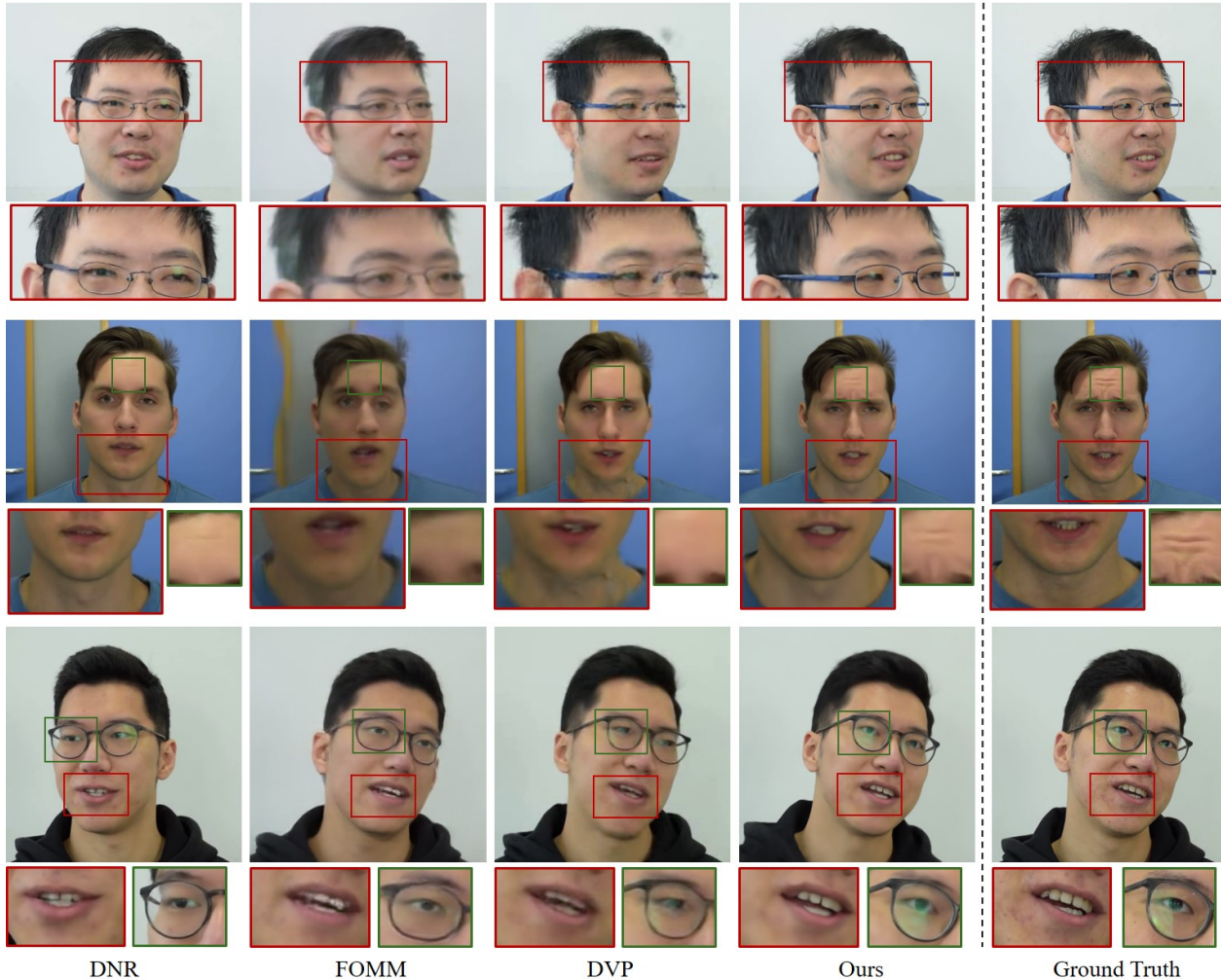
Figure 5: Comparison to state-of-the-art facial reenactment methods based on self-reenactment. From left to right: Deferred Neural Rendering (DNR) [39], First Order Motion Models (FOMM) [32], Deep Video Portraits (DVP) [21], Ours and the ground truth image. Note that DNR does not provide control over the pose parameters and only changes the facial expressions. As can be seen, our approach faithfully reconstructs the expression and appearance of the faces, and can also represent the geometry of the glasses including the view-dependent effects (see last row).

4. Results

Our approach allows the reconstruction of a 4D facial avatar based on monocular video sequences (see Sec. 4.3). In the following, we analyze our method qualitatively and quantitatively on real data (Sec. 4.1). Specifically, we show comparisons to state-of-the-art facial reenactment methods (Sec. 4.2) and discuss the conducted ablation studies (Sec. 4.4). The advantages of our approach is best seen in the supplemental video, especially, the 3D consistency of pose changes and the faithful reproduction of appearance.

4.1. Monocular Training Data

Our method uses short monocular RGB video sequences. We captured various human subjects with a Nikon D5300

DSLR camera at a resolution of 1920×1080 pixels with a framerate of 50 frames per second. The images are cropped to 1080×1080 and scaled to 512×512. The sequences have a length of about 2 min (6000 frames). We hold out the last 20 seconds (1000 frames) to serve as a test sequence for each reconstruction. The subjects were asked to engage in normal conversation, including expressions like smiling.

4.2. Comparison to the State of the Art

From the application stand-point, our method competes with state-of-the-art facial reenactment methods that allow to apply pose and expression changes. Specifically, we compare our method with Deep Video Portrait of Kim *et al*. [21], Deferred Neural Rendering of Thies *et al*. [39] and First-Order Motion Models of Siarohin *et al*. [32]. In

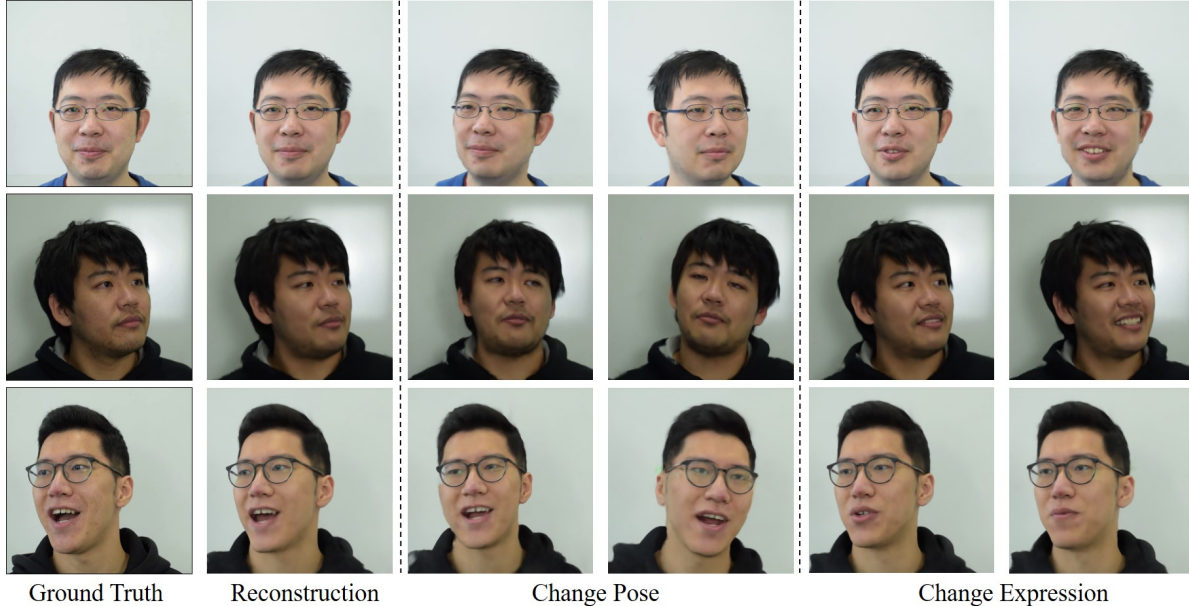Ground Truth   Reconstruction   Change Pose   Change Expression

Figure 6: We demonstrate the manual controllability of pose and expression using our 4D facial avatars reconstructed from monocular video inputs. Specifically, we demonstrate 3D consistent novel head pose synthesis and expression changes (by changing the 'open mouth' blendshape coefficient).

Fig. 5 we show qualitative results of the above-mentioned and our own method in a self-reenactment scenario. As can be seen, our method is able to reproduce the photorealistic appearance of the subjects. In contrast to the other methods, our approach generates 3D consistent results including view-dependent effects like the reflections on the glasses. Especially, synthesizing new head rotations is challenging for the baseline methods. Note that the approach of Thies *et al.* [39] only controls the facial expressions and not the pose. To quantitatively evaluate our method and the other two approaches, we compute the mean L_1-distance, Peak Signal-to-Noise Ratio (PSNR), and Structure Similarity Index (SSIM) [51], as well as the Learned Perceptual Image Patch Similarity (LPIPS) [49], see Tab. 1.

4.3. Novel Pose and Expression Synthesis

The goal of our method is the reconstruction of a 4D facial avatar with explicit control over pose and expressions. We show several reconstructed avatars in Fig. 6 including synthesized images with modified facial expressions and rigid pose. The results are best seen in the supplemental video, which shows that our dynamic neural scene representation can effectively store the appearance and geometry of a talking head. In addition to manual expression and pose edits, we demonstrate facial reenactment where we transfer the facial expressions of one person to another, see Fig. 7.

4.4. Ablation Studies

Our method assumes a static background and receives a background image as input. This background image helps

Method	$L_1 \downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FOMM [32]	0.036	23.77	0.91	0.16
DVP [21]	0.021	25.67	0.93	0.10
Ours (no BG)	0.035	23.52	0.90	0.18
Ours (no dyn.)	0.024	26.65	0.93	0.11
Ours (full)	**0.019**	**26.85**	**0.95**	**0.06**

Table 1: Quantitative evaluation of our method in comparison to state-of-the-art facial reenactment methods based on self-reenactment (see. Fig. 5). Ours (no dyn.) refers to our method without conditioning on dynamics. Ours (no BG) is our method without background image input.

to disentangle the foreground (4D facial avatar) and the background (see Fig. 4). The conditioning on the facial dynamics in form of the per-frame facial expression coefficients and learnable latent codes is one of the key components of our approach. Note that during test time we always employ the latent code from the first frame of the training set. Besides qualitative results, we also list a quantitative evaluation in Tab. 1. As can be seen, all components of our approach improve the quality of the results.

While static neural radiance fields can achieve satisfactory quality with as few as 100 posed images [28], our method requires more training data. In our setting the dynamic radiance field is required to generalize over the space of expression vectors. To quantify the need of a large training corpus, we conducted experiments by only training on the first halves and quarters of the training sequences, such that a lower variety of expressions and poses is seen during

8655

Figure 7: Our 4D facial avatars allow for facial reenactment, where the expressions of a source person are transferred to a target actor which we represent with our dynamic neural radiance field. Note that for facial reenactment we only need to train a model for the target actor; the expressions and pose changes from the source actor can be obtained in real-time [41].

Method	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
Ours (25%)	0.029	24.22	0.93	0.09
Ours (50%)	0.024	25.47	0.94	0.07
Ours (full)	**0.019**	**26.85**	**0.95**	**0.06**

Table 2: Ablation study w.r.t. training corpus size. All metrics significantly benefit from a larger training corpus.

training. The measured degradation in quality as we train with less data is shown in Tab. 2.

5. Limitations

In comparison to the state-of-the-art methods, our volumetric 4D representation of the head shows significantly better reconstruction abilities both quantitatively and qualitatively. Nevertheless, our approach still has limitations which we want to discuss in the following. The morphable model we use [7, 41] does not model eye blinks and eye movements, thus, these deformations can not explicitly be controlled in our approach. However, eye blinks are implicitly correlated with other expression parameters, and consequently modelled by our method. Our method is not restricted to this morphable model and could also be used with more sophisticated models that include these additional control handles.

The focus of our work is the reconstruction of the human head; we are currently not modelling the dynamics of the upper body. In future work, our approach can be extended to these regions (given a consistent tracking of the torso).

6. Ethical Considerations

While the described method was developed with the best of intentions and has the potential to positively influence the world via better content creation tools, light-weight facial editing methods have the potential for being misused. Our method currently requires an image of the background without the human subject, which makes it non-trivial to apply to in-the-wild videos. Moreover, the detection of manipulated videos is an active area of research [31, 12, 1, 24, 30, 13, 11, 2, 44] and we foresee significant improvements in this domain.

7. Conclusion

We have presented a novel method for learning and rendering controllable 4D facial avatars based on dynamic neural radiance fields. Using volumetric rendering, we are able to capture arbitrary geometry and topology such as hair, eyeware, hats etc., which typically is not supported by morphable model based methods. In contrast to other volumetric approaches which require an expensive calibrated multi-view rig, our method requires only a single view from a fixed camera, such as a webcam. This makes our method suitable for capturing avatars of end users at home, using only 2 minutes of their time. The reconstructed avatars can be rendered photorealistically under novel poses and expressions. The achieved quality beats state-of-the-art facial reenactment methods both quantitatively and qualitatively.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[2] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection, 2020.

[3] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. Bringing portraits to life. *ACM Trans. Graph.*, 36(6), Nov. 2017.

[4] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.*, 30:75:1–75:10, August 2011.

[5] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Proc. EUROGRAPHICS*, volume 22, pages 641–650, 2003.

[6] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. *CGF*, 23(3):669–676, 2004.

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, pages 187–194, 1999.

[8] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. In *ACM TOG*, volume 32, pages 40:1–40:10, 2013.

[9] Yen-Lin Chen, Hsiang-Tao Wu, Fuhao Shi, Xin Tong, and Jinxiang Chai. Accurate and robust 3d facial capture using a single rgbd camera. *Proc. ICCV*, pages 3615–3622, 2013.

[10] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection, 2020.

[12] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv*, 2018.

[13] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[14] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormaehlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proc. CVPR*, 2014.

[15] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. VDub - modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer Graphics Forum (Proceedings of EUROGRAPHICS)*, 2015.

[16] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph.*, 37(6), Dec. 2018.

[17] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. Single-view hair modeling using a hairstyle database. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 34(4), 2015.

[19] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6), Nov. 2017.

[20] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM TOG*, 34(4):45:1–45:14, 2015.

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

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[23] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. In *ACM TOG*, volume 32, 2013.

[24] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[25] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020.

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019.

[27] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *arXiv*, 2020.

[28] 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 *ECCV*, 2020.

[29] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[31] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces, 2018.

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for

image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[33] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

[34] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[35] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019.

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

[37] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. State of the art on neural rendering. *EG*, 2020.

[38] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020.

[39] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)*, 2019.

[40] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM TOG*, 34(6), 2015.

[41] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.

[42] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. FaceVR: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. on Graph.*, 2018.

[43] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. HeadOn: Real-time reenactment of human portrait videos. *ACM Trans. on Graph. (Proceedings of SIGGRAPH)*, 2018.

[44] Luisa Verdoliva. Media forensics and deepfakes: an overview, 2020.

[45] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. In *ACM TOG*, volume 30, 2011.

[46] Thibaut Weise, Hao Li, Luc J. Van Gool, and Mark Pauly. Face/Off: live facial puppetry. In *Proc. SCA*, pages 7–16, 2009.

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

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[50] Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 249–265, Cham, 2018. Springer International Publishing.

[51] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[52] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum (Eurographics State of the Art Reports)*, 37(2), 2018.