Neural Head Avatars from Monocular RGB Videos

Philip-William Grassal1*
Malte Prinzler1*
Titus Leistner1
Carsten Rother1
Matthias Nießner2
Justus Thies3

1Heidelberg University 2Technical University of Munich 3Max Planck Institute for Intelligent Systems

Monocular RGB Video  Neural Head Avatar with Articulated Geometry & Photorealistic Texture

Figure 1. Given a monocular portrait video of a person, we reconstruct a Neural Head Avatar. Such a 4D avatar will be the foundation of applications like teleconferencing in VR/AR, since it enables novel-view synthesis and control over pose and expression.

Abstract

We present Neural Head Avatars, a novel neural representation that explicitly models the surface geometry and appearance of an animatable human avatar that can be used for teleconferencing in AR/VR or other applications in the movie or games industry that rely on a digital human. Our representation can be learned from a monocular RGB portrait video that features a range of different expressions and views. Specifically, we propose a hybrid representation consisting of a morphable model for the coarse shape and expressions of the face, and two feed-forward networks, predicting vertex offsets of the underlying mesh as well as a view- and expression-dependent texture. We demonstrate that this representation is able to accurately extrapolate to unseen poses and view points, and generates natural expressions while providing sharp texture details. Compared to previous works on head avatars, our method provides a disentangled shape and appearance model of the complete human head (including hair) that is compatible with the standard graphics pipeline. Moreover, it quantitatively and qualitatively outperforms current state of the art in terms of reconstruction quality and novel-view synthesis.

1. Introduction

Reconstructing and reenacting human heads has been a long studied research problem and will be a key driver for future applications in VR/AR, teleconferencing, games and the movie industry. For those applications, it is of particular interest to get an accurate 3D shape and appearance model that provides 3D consistency and strong identity preservation under novel view points, poses and expressions. Reconstructing such a model, especially from monocular input data (e.g., from a webcam), is difficult due to the complex geometry of facial dynamics and the missing 3D information [96]. Indeed, several state-of-the-art methods for talking head synthesis avoid explicit geometry reconstruction and rely on image or feature-based warping for motion control and generative networks for image synthesis [73,84,92]. These methods are generalized and deliver impressive reenactment results even with only a single input image of the subject. However, the quality of the approaches drops significantly for larger changes in pose or view point as no 3D-consistent geometry representation is used. Shape proxies such as a 3D morphable model [9,28] can be utilized to improve the 3D consistency of synthetic faces [15,44,79,81] since the facial information is embedded on the proxy surface. Besides image- or surface-based representations, volumetric representations are used [31,52]. While they show promising results without an explicit surface prior, these methods still lack a consistent full head shape reconstru-
tion from single view inputs and are not compatible with the standard rasterization pipelines.

In this work, we present Neural Head Avatars, an explicit shape and appearance representation of the complete human head (including hair) which can be used in existing graphics pipelines that use triangular meshes. Specifically, we employ coordinate-based multi-layer perceptrons (MLPs) to predict the 3D meshes and dynamic textures, depending on the facial expression and pose of real humans. These networks are embedded on the surface of the FLAME morphable head model [49] which also serves as a coarse shape and expression proxy. We show that one can optimize such an explicit head representation based on a short monocular RGB video sequence. Using color-dependent and color-independent energy terms during optimization, we disentangle the reconstruction of surface geometry and color detail. The resulting controllable 4D avatar (3D model + motion) is subject-specific and generates novel poses and expressions while preserving high photo-realism. Moreover, it demonstrates great visual quality under large view point changes and, therefore, addresses one of the main drawbacks of related approaches.

In summary, our contributions are:

- Neural Head Avatars, a novel, subject-specific representation for articulated human heads that explicitly reconstruct the full head geometry and produces photo-realistic results even under large view point changes,
- A fully differentiable optimization pipeline to optimize Neural Head Avatars from a short, monocular RGB video with color-dependent and color-independent energy terms that allow for the disentanglement of the surface shape and color detail.

2. Related Work

Reconstructing controllable 4D head or facial avatars is an actively studied field at the intersection of computer vision and computer graphics. For an extensive review of methods, we refer to the state of the art report of Zollhöfer et al. [96].

Image-based models. Image-based models synthesize the face of a subject without relying on any (explicit or implicit) representation in 3D space. These methods either utilize (learned) warping fields [6, 73] to deform an input image to match new poses or expressions, or deploy encoder-decoder architectures, where the encoder extracts an identity code from a given source image and a decoder synthesizes the output image [57, 83, 92, 94]. The decoder may be conditioned by facial landmarks [57, 92], facial contours [83], or parsing maps [94]. Even though these methods produce high-quality results and even allow for real-time synthesis [92], they suffer from artifacts for strong pose and expression changes, and lack geometric and temporal consistency. This is mainly due to the fact, that the appearance of deformations in three-dimensional space (e.g., yaw opening, head rotation) must be learned in 2D by these models.

Implicit models. Implicit models represent the geometry using implicit surface functions (e.g., signed distance functions) or by volumetric representations. A common approach is to represent the appearance of a target person in a discrete latent feature voxel grid that can be deformed to synthesize dynamic deformations [52, 84]. Motivated by their recent success in 3D scene reconstruction [77], neural radiance fields (NeRF) in combination with volumetric rendering [55] have been used to replace the discrete feature voxel grids [5, 31, 43, 50, 58–60, 63, 64, 68, 75, 90]. Articulated head avatars can be synthesized by conditioning the NeRF on low-dimensional parameters of a face model [31, 86] or audio signals [38]. Pixel-aligned Volumetric avatars [68] are generalized across subjects and can generate novel views, based on single or multiple input images. Even though solving geometric and temporal inconsistencies, the proposed methods either fail to disentangle pose and expression [84], are limited to static reconstructions [43, 68, 69, 85] or fail to generalize to unseen poses and expressions [31].

Explicit models. The majority of head reconstruction methods relies on explicit scene representations, i.e., triangular meshes [7, 8, 13, 19, 20, 22, 32–34, 39, 40, 44, 48, 78–82, 88, 89, 96]. For these methods, morphable models are used as a prior to reconstruct the face from incomplete (e.g., partially occluded) or noisy data (e.g., from depth maps). Morphable models are computed from a population of 3D head scans [28], and provide statistical information on physiologically plausible head shapes and facial movements [12, 25, 49, 62]. In addition to the geometry, these models can provide a statistical linear model for the texture [3, 9, 10, 29, 62] which can be used to reconstruct faces from RGB data only [9, 81, 96]. Recent work utilizes generative adversarial networks (GAN) [37, 56] to generate and optimize albedo and normal maps for specific subjects [36, 47]. Other approaches utilize 2D neural rendering [76, 77], to learn how to render photo-realistic imagery of a specific subject from a short training dataset [44, 78, 79]. These approaches are based on deep neural networks which can be conditioned on coarse RGB renderings based on a linear texture model [44], uv-maps [15], latent feature maps [79] or point clouds [67]. While these methods produce geometrically consistent avatars that can be easily controlled, they either are limited to craniofacial structures and do not include the synthesis of hair [36, 47] or suffer from temporal and spatial inconsistencies due to their loose bound to
the geometric backbone \cite{15, 67, 79}. Other methods model mesh displacements to reconstruct fine detail wrinkles and hair \cite{4, 18, 30}, therefore, maintaining a close bound to the underlying geometry. However, none of them produces photo-realistic outputs under novel views.

In contrast to previous work, we jointly optimize a photo-realistic dynamic texture together with subject specific geometry that includes facial detail and hair structure, stored in a coordinate-based, fully connected neural network. By deploying a texture-based approach, we tightly link appearance and underlying geometry which ensures spatial consistency and generalization to unseen poses and expressions. Articulated jaw, neck, and eyes allow for intuitive avatar control. We demonstrate that an avatar can be optimized from a short monocular RGB sequence without the need of special (multi-view) camera setups as in \cite{51, 54}.

3. Method

Given an RGB video sequence of a talking person consisting of \(N\) consecutive frames \(I_1, I_2, \ldots, I_N\), we reconstruct a 4D neural avatar based on a explicit representation that allows for pose- and expression-dependent novel viewpoint synthesis. Specifically, our model outputs a classical triangle mesh, i.e., vertices \(V = \{v_1, v_2, \ldots, v_n\}, v_i \in \mathbb{R}^3\), connecting faces \(F\) and a texture function \(T\) that assigns an RGB color value to each point on the surface defined by \(V\) and \(F\) (see Section 3.1). Thus, the standard graphics pipeline can be utilized to obtain a rendered image \(\hat{I}\) assuming a full perspective camera projection. Based on this image formation model, we optimize our avatar representation in an analysis-by-synthesis-based fashion (see Section 3.2). An overview of our method is depicted in Figure 2.

3.1. Explicit Neural Head Representation

Our explicit surface representation is embedded on the FLAME \cite{49} template surface and shares its topology \(F\). Specifically, we employ a multi-layer perceptron (MLP) \(G\) which models the pose-dependent offsets w.r.t. the template surface. To generate the view-, pose-, and expression-dependent texture of the face, we use an MLP \(T\) which predicts a color value at any surface point of the mesh.

Template Model. We deploy the parametric FLAME head model \cite{49} as a geometric backbone of our method:

\[
V_{\text{flame}} : \mathbb{R}^{300}, \mathbb{R}^{100}, \mathbb{R}^{3k} \rightarrow \mathbb{R}^{16227 \times 3} \\
\beta, \psi, \phi \mapsto V_{\text{flame}}(\beta, \psi, \phi)
\]

where \(\beta, \psi\) and \(\phi\) describe shape, expression and \(k = 4\) joint pose parameters, respectively. We perform minor adjustments to the FLAME topology, namely, we uniformly subdivide the faces (four-way subdivision), remove the faces belonging to the lower neck region and add additional faces to close the mouth cavity. This increases the original number of vertices from 5023 to 16227.

Geometry Refinement Network \(G\). To model facial detail and hair which is not represented by the FLAME head model, we introduce a pose-dependent offset function for geometry corrections:

\[
G : \mathbb{R}^{3k} \rightarrow \mathbb{R}^{16227 \times 3} \\
\phi \mapsto G(\phi).
\]

Using this offset function, the mesh geometry is given by:

\[
V(\beta, \psi, \phi) = V_{\text{flame}}(\beta, \psi, \phi) + G(\phi).
\]
Geometry Objective $E_{\text{geom}}$. To disentangle appearance and geometry, we define a geometry energy term which is independent of the actual appearance:

$$E_{\text{geom}} = w_{\text{link}} \cdot E_{\text{link}} + w_{\text{normal}} \cdot E_{\text{normal}} + w_{\text{semantic}} \cdot E_{\text{semantic}} + w_{\text{reg,geom}} \cdot E_{\text{reg,geom}}.$$  \hspace{1cm} (2)

The landmark energy $E_{\text{link}}$ measures the $\ell_1$ distance of detected 2D facial landmarks [16, 53] and the projected counterparts on the mesh surface. Besides the absolute positions of the landmarks, it also measures the relative distances of the eye landmarks at the upper and lower lid to improve the reconstruction of eye lid closure [30] (see ablation study in Figure 8 (c)).

The energy term $E_{\text{normal}}$ is based on pseudo-normal maps $N \in \mathbb{R}^{H \times W \times 3}$ using the pretrained model of Abrevaya et al. [1]. Based on these predictions, we formulate a reconstruction energy term for fine geometric detail. As our focus lies on high-frequency geometry detail, instead of minimizing the absolute difference between pseudo-normals $N$ and our predicted normals $\hat{N}$, we instead match their image-space Laplacians $\lambda(\cdot)$:

$$E_{\text{normal}} = |\lambda(\hat{N}) - \lambda(N)|_1.$$  

We employ $E_{\text{semantic}}$ to match the semantic regions $S_k$ of the input and $\hat{S}_k$ of the reconstructed mesh for facial skin and neck, eyes, ears, and hair:

$$E_{\text{semantic}} = \sum_{k=1}^{4} S_k \oplus \hat{S}_k,$$

where $\oplus$ denotes an xor on the region. The respective semantic maps $S_k$ are computed using [91, 95] (see Figure 3 (b)).

Besides the data terms, we employ a regularization term $E_{\text{reg,geom}}$ which regularizes the FLAME parameters, as well as the geometry MLP $G$:

$$E_{\text{reg,geom}} = w_{\text{reg,flame}} \cdot E_{\text{reg,flame}} + w_{\text{reg,offset}} \cdot E_{\text{reg,offset}}.$$  

Following [9, 81], $E_{\text{reg,flame}}$ uses the statistical properties of the linear shape model, and regularizes the prediction towards the canonical template head using an $\ell_2$-norm on $\beta, \psi, \phi$. The offsets are regularized using $E_{\text{reg,offset}}$ which consists of a Laplacian regularizer and a regularizer that controls the pose-consistency and distribution of the predicted offsets. We refer to Appendix A.3 for additional details.

Appearance Objective $E_{\text{app}}$. The appearance term $E_{\text{app}}$ measures the reproduction of the color image $I$. It depends on both the geometry as well as the texture parameters of our neural head model. We use dense per-pixel energy terms
Initialization and Optimization Strategy. To initialize our reconstruction method, we adopt the tracking algorithm proposed by [81] which optimizes camera, shape, expression, and pose parameters based on an analysis-by-synthesis approach, using the FLAME model [49] with a linear texture space [29]. The resulting reconstruction is in coarse alignment with the training sequence, but the FLAME model is limited to bald heads and lacks fine, subject-specific geometric detail (see Figure 2). We initialize the geometry refinement network $G$, by optimizing only for the $E_{geom}$ term defined in Equation (2). Once, we obtained an estimate of the full head geometry, we optimize for the texture MLP parameters w.r.t. $E_{app}$ (Equation (3)). Based on this initialization scheme, we optimize jointly the geometry and texture parameters to minimize $E_{joint}$ (Equation (1)). For implementation details, we refer to Appendix A.2

4. Results

We quantitatively and qualitatively evaluate the performance of our model on the tasks of geometry reconstruction as well as novel pose-, expression-, and view synthesis, and compare it to state-of-the-art methods.

4.1. Datasets

Since the current literature does not provide suitable datasets for the evaluation of dynamic full head approaches, we created two datasets.

Synthetic Data. Our synthetic dataset has been generated with the open source MakeHuman project [14] which allows to model fully animatable and texturized human models with high variance in appearance and facial geometry. We generated two female and two male subjects with different ethnicity and head geometry, and render animated sequences (200 training frames, 210 validation frames). The resulting sequences provide ground truth RGB-, normal- and semantic maps as well as landmarks and 3D meshes which is used to quantitatively evaluate the geometry reconstruction of our method. Consequently, for evaluation purposes, we will not rely on predicted pseudo-ground truth (normal and semantic maps, landmarks) for experiments with this dataset. Note that the synthetic meshes have different topologies compared to FLAME and our model.

Real Data. Our real dataset contains four sequences of humans, two with male actors, two with a female actor. The sequences capture various hair styles such as short hair, long hair and a hair bun. All sequences show the subjects during a natural conversation in front of a green screen in an environment with uniform lighting. We capture 750 training frames and 750 validation frames for each sequence and ensure that both sides of the head are visible at least once in the training partition. We complement the obtained RGB ground truth with detected facial landmarks as well as normal- and semantic maps by deploying pretrained models [1,16,95]. The resulting dataset is used to qualitatively and quantitatively evaluate our model on the task of novel pose-, expression-, and view synthesis, and to compare to state-of-the-art methods.

4.2. Geometry Reconstruction Quality

To quantify the head shape reconstruction quality, we utilize the rendered ground truth normals, as well as the meshes provided by the synthetic recordings. Those are compared with the predicted meshes and normal maps by computing their single-sided Hausdorff distance in millimeters (mesh alignment error, as in [66]) and their normal angular error on the validation sequence. The Hausdorff distance is computed once for the full head and once for the facial region only. We compare with the reconstructions by FLAME [49] obtained from our tracker as a baseline. The quantitative results are reported in Table 1. Our reconstruction and the misalignment error w.r.t. the ground truth are visualized in Figure 4. We can see that our approach reconstructs the talking head faithfully, and even regions, where only a few silhouette views are available (side of the bun or front/back of the neck) can be estimated, however, with a slightly higher reconstruction error. The reconstruction of real subjects in neutral pose is compared with multi-view stereo (MVS) recordings in Figure 18 in the Appendix.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Female 1</th>
<th>Female 2</th>
<th>Male 1</th>
<th>Male 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal: FLAME</td>
<td>15.8°</td>
<td>13.3°</td>
<td>15.0°</td>
<td>14.8°</td>
</tr>
<tr>
<td>Normal: Ours</td>
<td>14.4°</td>
<td>12.2°</td>
<td>13.7°</td>
<td>13.7°</td>
</tr>
<tr>
<td>Hausdorff: FLAME</td>
<td>1.4</td>
<td>0.9</td>
<td>1.5</td>
<td>1.1</td>
</tr>
<tr>
<td>Hausdorff: Ours</td>
<td>1.2</td>
<td>0.9</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Hausdorff: FLAME</td>
<td>5.5</td>
<td>4.8</td>
<td>6.1</td>
<td>5.7</td>
</tr>
<tr>
<td>Hausdorff: Ours</td>
<td>2.6</td>
<td>2.5</td>
<td>3.0</td>
<td>3.1</td>
</tr>
</tbody>
</table>

2Additionally, we will release the code for research purposes.
of our real dataset, and demonstrates that we outperform related approaches consistently. The qualitative comparison in Figure 5 confirms that our model synthesizes images with higher detail and better expression conservation than related methods.

4.4. Novel View Synthesis

The explicit geometry representation for each subject (shown in Figure 6) enables 3D-consistent novel viewpoint synthesis. We evaluate the novel viewpoint synthesis, by rendering frames from the validation set of the real sequences under novel yaw angles. We compare our method against NerFACE [31] and DeepVideoPortraits (DVP) [44] as these methods rely on a representation in 3D space that can be rotated accordingly. The underlying 3D representations are centered on the image plane before rendering to account for varying coordinate origin definitions. Figure 12 demonstrates that while related methods suffer from significant artifacts, our method maintains its high visual quality.

4.5. Ablation Study

We evaluate the influence of different architecture choices and optimization terms on the real dataset. Our geometry is based on the FLAME model, and extends it to capture person-specific detail. For deformations of the face, we rely on the blendshape-based expression model and the linear blend-skinning of the yaw bone. Nevertheless, in our experiments, we found that a static geometry refinement of the underlying FLAME model does not reconstruct the neck region realistically as the joint poses and global rotation of the initial FLAME estimate are highly ambiguous. To this end, we condition the geometry network $G$ with the joint poses of the FLAME model which compensates errors in the neck region, see Figure 8 (a).

Similar to the dynamic geometry network, we use a pose- and expression-dependent network $T$ to predict the surface radiance. Figure 8 (b) demonstrates that the use of a static texture, i.e. when the dynamic conditionings of $T$ are fixed to zero, results in less authentic synthesis results, especially for the highly dynamic mouth region.

Our model also supports eye blinks, which are modelled by the geometry of the FLAME model. To enforce the re-

<table>
<thead>
<tr>
<th>Method</th>
<th>L1 ↓</th>
<th>MS-SSIM ↑</th>
<th>LPIPS ↓</th>
<th>CSIM ↑</th>
<th>CPBD ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>VariTex [15]</td>
<td>0.207</td>
<td>0.640</td>
<td>0.316</td>
<td>0.347</td>
<td>0.565</td>
</tr>
<tr>
<td>Bi-Layer [92]</td>
<td>0.108</td>
<td>0.805</td>
<td>0.285</td>
<td>0.757</td>
<td>0.403</td>
</tr>
<tr>
<td>FOMM [73]</td>
<td>0.055</td>
<td>0.911</td>
<td>0.201</td>
<td>0.800</td>
<td>0.370</td>
</tr>
<tr>
<td>DVP [44]</td>
<td>0.042</td>
<td>0.904</td>
<td>0.098</td>
<td>0.815</td>
<td>0.632</td>
</tr>
<tr>
<td>NerFACE [31]</td>
<td>0.057</td>
<td>0.897</td>
<td>0.156</td>
<td>0.864</td>
<td>0.485</td>
</tr>
<tr>
<td>Ours</td>
<td>0.033</td>
<td>0.923</td>
<td>0.079</td>
<td>0.884</td>
<td>0.674</td>
</tr>
</tbody>
</table>

Table 2. Quantitative evaluation of the appearance reconstruction on the real dataset. Reported are the average scores over all validation frames for all subjects in our real dataset (see Section 4.1).
4.6. Discussion

We have demonstrated that our method produces high-quality results even for large head rotations due to our explicit geometry reconstruction of the face and hair region. As such, our method addresses one of the main issues of other learning-based approaches which suffer from significant artifacts when synthesizing novel views. Similar to all baselines, we do not address the synthesis of physical effects like floating or deforming hair. Incorporating physics into a 4D avatar is an interesting field for future work.

Moreover, our method exhibits limitations in regions where the explicit geometry is greatly unconstrained, most prominently the mouth cavity. As a consequence, the visual quality of the synthesized inner mouth region, especially teeth, may decrease if expressions and poses lie far outside of the training corpus (see Figure 9). The most natural approach to overcome this issue would be the integration of a well-aligned geometry of the inner mouth cavity. However, so far none of the publicly available parametric head models provides such geometry due to the difficulties in the acquisition of ground truth data. Similar to NerFACE [31] and DVP [44], our method is person-specific and, thus, requires optimization of the neural network for every new actor which takes 7 hours using two Nvidia A100 GPUs, when optimizing on images with a resolution of $512 \times 512$ px. The optimization time and necessary computational resources can be reduced greatly when optimizing against smaller images or image patches. Generalizing our approach is future work which can benefit from the findings of VariTex [15] and pi-GAN [21].
Ethical Considerations  With the advances in the synthesis of photo-realistic human avatars, the potential misuse (e.g., misinformation) becomes an increasingly important ethical concern. While active watermarking of generated content can be employed, there is no guarantee that this watermark can not be removed. To this end, there exists the field of multi-media forensics which analyzes methods for active and passive forgery detection. Passive forgery detection [2, 23, 24, 71, 72] is able to detect manipulated or synthetic imagery without any explicit watermarking. While these methods can be trained to detect specific manipulation methods [71, 72], generalized methods [2, 23, 24] have problems in reliably detecting fakes. Also, forgery detectors can be used in an adversarial training to improve the quality of the synthetic renderings. Thus, besides forgery detection algorithms, cryptographical approaches for signing the authenticity of video material have to be used in the future (which requires a trust network).

5. Conclusion

In this work, we presented Neural Head Avatars, a method that accurately reconstructs geometry and appearance of the human head from a monocular RGB sequence. Our approach combines a parametric head model with multi-layer perceptrons that refine geometry and synthesize a photorealistic texture. The resulting 4D avatar is robust with respect to large pose-, view- and expression changes, and we show that it outperforms state-of-the-art head avatar methods qualitatively and quantitatively.

While recent work on head avatar synthesis moved towards implicit representations of spatial geometry, our work demonstrates the benefits of an explicit geometry reconstruction in combination with a deep appearance network for dynamic surface textures in terms of photorealism and generalizability. We hope that our work inspires further research at the intersection of explicit geometry reconstruction and deep appearance representations.

Acknowledgements

This project has received funding from the DFG in the joint German-Japan-France grant agreement (RO 4804/3-1), the ERC Starting Grant Scan2CAD (804724), and a Microsoft Research Grant. We thank the recorded actors and video narrator for their help, as well as, Mohit Mendiratta and Guy Gafni for providing the synthesis results of [44] and [31] for our comparisons. We gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden.
References


[16] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, 2017. 4, 5, 15, 17


[56] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures, pages 1–12, 2018. 2, 17, 18


Our FLAME topology

G

A.2.1 Geometry MLP

networks. An overview is presented in Figure 11. This section details the model architecture of both expression dependent effects, e.g., wrinkles and reflections. The Texture MLP \( T \) synthesizes a dynamic, photo-realistic texture that is able to reproduce view- and expression dependent effects, e.g., wrinkles and reflections. This section details the model architecture of both networks. An overview is presented in Figure 11.

A.2. Network Architectures

Our method relies on two multi-layer perceptrons (MLPs). The Geometry MLP \( \mathcal{G} \) refines the mesh resulting from the linear FLAME head model and adds facial detail and hair structure. The Texture MLP \( T \) synthesizes a dynamic, photo-realistic texture that is able to reproduce view- and expression dependent effects, e.g., wrinkles and reflections. This section details the model architecture of both networks. An overview is presented in Figure 11.

A.2.1 Geometry MLP \( \mathcal{G} \)

The Geometry MLP \( \mathcal{G} \) adds facial detail and hair geometry to the mesh resulting from the FLAME head model. The inputs to the network are the 3D coordinates of the vertices on the template mesh (see Figure 10), normalized to the range \([-1, 1]\), as well as vertex-specific embedding vectors that are optimized during training. Dynamic geometry effects are enabled by also passing the FLAME pose parameters to \( \mathcal{G} \). In practice, to avoid overfitting only the 3 pose parameters of the neck joint are used.

The core of \( \mathcal{G} \) is a SIREN-based fully connected MLP \([74]\) which takes the vertex positions and embeddings as input, processes them by a sequence of FiLM-conditioned linear layers \([27, 65]\), and predicts the three-dimensional vertex offsets. The linear, FiLM-conditioned layers first perform an affine transformation on the input signal, followed by a sinusoidal activation function. The phase shifts and frequencies of the sinusoidal activation functions are generated by a mapping network which is a concatenation of linear layers with leaky ReLU activations. This architecture was greatly inspired by \[21\]. The input to the mapping network is the three-dimensional pose of the FLAME neck joint which enables the synthesis of dynamically changing geometry refinements. We ensure spatial geometry consistency by only allowing dynamic geometry refinements for the neck region. To this end, we compute the offsets twice: once conditioned on the original neck pose values and once with all conditions set to zero. We smoothly blend both offset predictions according to a fading body region mask.

Configuration details. The vertex embedding vectors have a feature dimension of 32 (for each vertex of template mesh). The mapping network comprises 3 hidden layers with 256 neurons each, followed by leaky ReLU activation functions with slope 0.2. The SIREN MLP consists of 6 consecutive linear FiLM layers with 128 neurons each. The final layer is a linear layer with 3 output neurons, followed by a tanh activation function.

A.2.2 Texture MLP \( T \)

The architecture of our texture representation \( T \) is similar to \( \mathcal{G} \). For obtaining the color value of a point on the mesh surface, its 3D coordinates on the FLAME template mesh as well as a surface embedding vector are fed into a SIREN MLP. The surface embedding vector is sampled from a discrete 2D grid in uv space. A mapping network takes features extracted from the FLAME pose and expression parameters and predicts the phase shifts and frequencies of the sinusoidal activation functions in the linear FiLM layers of the SIREN MLP to synthesize a dynamic texture. The dynamic texture is needed in the regions where the reconstructed geometry does not align well with the target surface. This is especially the case for the mouth cavity. Consequently, we process the FLAME pose and expression parameters into features that are highly correlated with the mouth articulation. More specifically, we determine the effective rotation that applies to the mouth region, i.e., combine global rotation and neck rotation, and use its axis-angle representation together with 10 pair-wise distances between vertices on the inner side of the lips as inputs to the mapping network. These conditions only cover mouth-related information and, therefore, are only used as input to the mapping network for the respective surface regions. For all other regions, we fill the conditioning vector with zeros. Effectively, this approach greatly limits the dynamic capacities of the \( T \) to ensure a close bound to the underly-
The Texture MLP shifts of the Linear FiLM layers are predicted by fully connected mapping networks which are conditioned on the FLAME parameters. For layers, periodic activation functions and FiLM conditionings [27, 65]. Inspired by [21], surface coordinates and spatial embeddings (either

embedded in uv space from a discrete feature grid with and a activations. The activations of both heads are summed up fed through a sequence of linear layers to produce 3 output

SIREN MLP and the normal encoder are concatenated and

inclusions and specular highlights). The output vectors of the

expression- and view-dependent effects (e.g., ambient oc-

local geometry configuration and enables the synthesis of

Geometry MLP

Mapping Network

Linear

Leaky ReLU

Linear

Leaky ReLU

Vertex Position \( \hat{p} \)

Vertex Embedding \( \hat{z} \)

Vertex Offsets \( \hat{\sigma} \)

Figure 11. Overview of our model architectures. Our Neural Head Avatar relies on SIREN-based MLPs [74] with fully connected linear layers, periodic activation functions and FiLM conditionings [27, 65]. Inspired by [21], surface coordinates and spatial embeddings (either vertex-wise for \( \tilde{G} \), or as an interpolatable grid in uv-space for \( T \)) are used as an input to the SIREN MLP. The dynamic frequencies and phase shifts of the Linear FiLM layers in the SIREN MLP.

Configuration details. The surface embedding is sampled in uv space from a discrete feature grid with 256 × 256 feature vectors with 64 channels each via bilinear interpolation. We use a separate uv map for the inner mouth region with 64 × 64 vectors. The mapping network has the same architecture as for the Geometry MLP \( G \). The SIREN MLP consists of 8 consecutive linear FiLM layers with 256 neurons each, except for the last which has 128 neurons. The normal encoder is designed as a fully convolutional network with 3 consecutive 2D-convolutional layers with stride 1 and kernel size 3 and with periodic activations. All layers have 128 feature channels except for the last one which has 32. The first model head contains one fully connected layer with 3 output neurons. The second model head contains one linear FiLM layer with 128 neurons and one linear layer with 3 neurons.

A.3. Optimization from Monocular RGB Data

Detailed Geometry Objective \( E_{\text{geom}} \) (Eq. 2) As defined in Eq. 2 in the main paper, the geometry energy term is:

\[
E_{\text{geom}} = w_{\text{lmk}} \cdot E_{\text{lmk}} + w_{\text{normal}} \cdot E_{\text{normal}} + w_{\text{semantic}} \cdot E_{\text{semantic}} + w_{\text{reg, geom}} \cdot E_{\text{reg, geom}}. \tag{4}
\]

\( E_{\text{normal}} \) and \( E_{\text{semantic}} \) are already defined in the main paper, for the other terms we provide additional explanation.

The landmark energy \( E_{\text{lmk}} \) measures the distance of detected 2D facial landmarks \( l_i \in \mathbb{R}^2 \) and the projected coun-
The Laplacian regularizer $\lambda_{\text{reg,lapl}}$ of the Laplace values $E_{\text{flame}}$ of the predicted mesh vertices $V$ w.r.t. the surface of the FLAME model $V_{\text{flame}}$ and controls the smoothness of the predicted offsets:

$$E_{\text{reg,lapl}} = |W_{\text{reg,lapl}} \circ (\lambda^*(V) - \lambda^*(V_{\text{flame}}))|_1.$$ 

Table 3. Geometry regularization weights. The regularization weights differ for the individual head regions. However, they do not change among target subjects, i.e., they can be used to reconstruct a wide variety of geometries such as different hair styles.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyeballs</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>Eye Surrounding</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$5$</td>
<td>$5$</td>
<td></td>
</tr>
<tr>
<td>Forehead</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0.05$</td>
<td>$0.1$</td>
<td></td>
</tr>
<tr>
<td>Face</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0.05$</td>
<td>$0.1$</td>
<td></td>
</tr>
<tr>
<td>Ears</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$25$</td>
<td>$50$</td>
<td></td>
</tr>
<tr>
<td>Scalp</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$10$</td>
<td>$10$</td>
<td>$0.05$</td>
<td>$0.1$</td>
<td></td>
</tr>
<tr>
<td>Neck</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0.1$</td>
<td>$0.2$</td>
<td></td>
</tr>
<tr>
<td>Lower Neck</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0.25$</td>
<td>$0.5$</td>
<td></td>
</tr>
<tr>
<td>Nose</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>$0$</td>
<td>$0$</td>
<td>$2.50 \times 10^{-2}$</td>
<td>$5.00 \times 10^{-2}$</td>
<td></td>
</tr>
</tbody>
</table>

Note that $\lambda^*(\cdot)$ denotes the discretized Laplace-Beltrami operator on the 1-ring neighborhood. $W_{\text{reg,lapl}} \in \mathbb{R}^V$ defines vertex specific weights which allow to control the smoothness in specific regions (e.g., the neck region has a higher regularization). $\circ$ denotes the component-wise Hadamard product. As in [17], we regularize pose-dependent offset variations by adding a pair-wise surface consistency term.

$$E_{\text{reg,surface}} = ||G(\phi_i) - G(\phi_j)||_1.$$ 

For large geometry corrections (e.g., long hair), we noticed an irregular vertex distribution over the surface which resulted in areas with large triangles and coarse shape modelling. To mitigate this issue, we regularize the length of edges $e_i$ in the scalp region if they deviate too much from the average edge length $\bar{e}$.

$$E_{\text{reg,edge}} = \sum e_i \begin{cases} e_i - \bar{e} & \text{if } e_i > 1.5 \cdot \bar{e} \\ 0 & \text{otherwise} \end{cases}$$

As the individual facial regions are subject to different requirements, the applied regularization weights differ. Table 3 presents the weights for the individual regions. Please note that while the weights differ for the individual face regions, they are the same for all avatars. No subject-specific fine-tuning of regularization parameters is required while still being able to reconstruct a wide set of structures (e.g., long and short hair).
Figure 12. Additional qualitative novel view synthesis comparisons. We observe that the advantages of our method, namely spatial consistency and high texture detail even under extreme head rotations, apply to a variety of identities. Also see Figure 1 for additional subjects.

Figure 13. Expressions $\psi$ and poses $\phi$ from the driving frames on the left are added to the neutral poses of optimized source avatars at the top. The reenacted avatars are displayed below.

**Detailed Appearance Objective $E_{\text{app}}$ (Eq. 3)** The appearance energy term is defined in Eq. 3 in the main paper as:

$$E_{\text{app}} = w_{\text{phot}} \cdot E_{\text{phot}} + w_{\text{perc}} \cdot E_{\text{perc}}.$$  

The photo-metric energy term $E_{\text{phot}}$ is defined on the intersection $\mathcal{V} = \mathcal{S} \cap \hat{\mathcal{S}}$ of the foreground segmentation of the input $I$ and the region that is generated by our head model $\hat{I}$:

$$E_{\text{phot}} = |\mathcal{V} \cdot (\hat{I} - I)|_1.$$  

To generate sharp textures [41, 46], we employ the style-based perceptual energy term $E_{\text{perc}}$ proposed by [35].

**Initialization and Optimization.** As discussed in Section 3.2 of the main paper, texture and geometry have to be initialized before starting the joint optimization against RGB sequences to avoid converging to bad local minima. To this end, the FLAME head model is aligned with the training sequence to obtain a coarse geometry initialization. Based on this coarse geometry, the weights of $\mathcal{G}$ and the framespecific parameters of the FLAME model, i.e., pose and expressions are optimized w.r.t. $E_{\text{geom}}$ (Eq. (2) in main paper). As a result of this optimization, we obtain a geometry estimate that aligns well with the target silhouette. Using this refined geometry, $T$ is trained to minimize $E_{\text{app}}$ (Eq. (3) in main paper). In a final optimization step, all components are optimized jointly to minimize $E_{\text{joint}}$ (Eq. (1) in main paper).

Given the 750 input frames of our real sequences, we initialize $\mathcal{G}$ for 150 epochs, $T$ is initialized for another 100 epochs, and we jointly optimize both components as long as...
Figure 15. Quantitative novel view synthesis comparison. We report the cosine similarity score between latent features predicted by a pretrained face recognition network [26]. The feature vectors are compared between the front facing ground truth and the predicted image under different rotation angles. We report the average scores over all validation frames from two sequences in our real dataset together with the respective $1\sigma$ regions. Our method consistently outperforms related approaches for pitch angles between $-15^\circ$ and $+15^\circ$. For extreme yaw rotations, we observe significantly reduced CSIM scores even though qualitative comparisons demonstrate high identity preservation under these conditions (see Figure 7 in main paper).

Figure 16. Qualitative comparison with paGAN [56] on the validation sequences. Note that only the facial region is synthesized by paGAN.

The perceptual loss on the holdout validation set decreases. This takes approximately 50 epochs. For longer sequences, we reduce the epoch count such that the number of iterations per optimization stage remains the same.

We deploy a standard Adam optimizer [45] for all frame-agnostic parameters and a standard SGD optimizer for frame-specific components (e.g., expression and pose parameters). Weight decay is applied to all texture-related components. The entire pipeline is implemented with Pytorch [61] and Pytorch3D [70].

B. Additional Results

B.1. Avatar Reenactment

Neural Head Avatars are controllable by the disentangled pose and expression spaces of the FLAME head model. This naturally enables the reenactment of an optimized avatar via a driving sequence. We demonstrate this capability in Figure 13. In this experiment, we utilize expressions $\psi$ and poses $\phi$ from a driver’s video and add them to the neutral expression of optimized avatars. $\psi$ and $\phi$ are extracted using our tracking algorithm discussed in the main paper. We observe that our method is able to faithfully transfer pose and expression between various subjects.

B.2. Novel View Synthesis

In Section 4.4 of the main paper, we qualitatively compare the synthesis from novel viewpoints against related methods. Figure 14 and Figure 12 provide further qualitative comparisons on different subjects and demonstrate that the advantages of our method apply to a variety of identities.

We also provide a quantitative evaluation of the novel view synthesis results in Figure 15. For a quantitative analysis, given that no ground truth is available for novel views, we evaluate the cosine similarity (CSIM) of the latent feature vectors predicted by a pretrained face recognition network [16] between front-facing ground truth and novel view prediction. We report the CSIM scores averaged over all validation frames of two sequences in our real dataset. We
find that our method outperforms related approaches consistently for pitch angles in a range of ±15°. However, we observe that for large yaw angles (≥ ±30°), the scores for all considered models decrease rapidly. Still, qualitative comparisons demonstrate that our method exhibits high identity preservation even under viewpoint changes in that range.

B.3. Comparison with paGAN [56]

In Figure 16, we compare our generated images of the real identities with the outputs of paGAN [56] provided by the authors. As their results are overlays on top of the original video, only the synthesized parts (facial region) should be considered when comparing to it.

B.4. Synthesis Results for Non-Caucasian Subjects

To validate that our method also performs well for non-caucasian subjects, we include synthesis results for a person of color in Figure 17. Also in this case, our method reconstructs the head geometry faithfully and renders visually plausible images.

B.5. Geometry Evaluation

We compare the geometry predicted by our approach on the real dataset with multi-view stereo (MVS) recordings of the respective identities, see Figure 18. The MVS data was captured separately with a handheld DSLR camera. As hair styles and face dynamics can not be reproduced reliably in separate recordings, only the neutral pose of the face region is compared.

B.6. Energy Ablation

Figure 19 demonstrates the effect of further energy terms on the synthesis results. We observe that $E_{\text{phot}}$ prevents color shifts, $E_{\text{semantic}}$ improves alignment of overlapping semantic regions within the avatar’s silhouette and $E_{\text{perc}}$ results in textures with more detail.