Video-Driven Animation of Neural Head Avatars

W. Paier[1], P. Hinzer[1] A. Hilsmann[1], P. Eisert[1,2],

[1]Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Humboldt University, Berlin, Germany

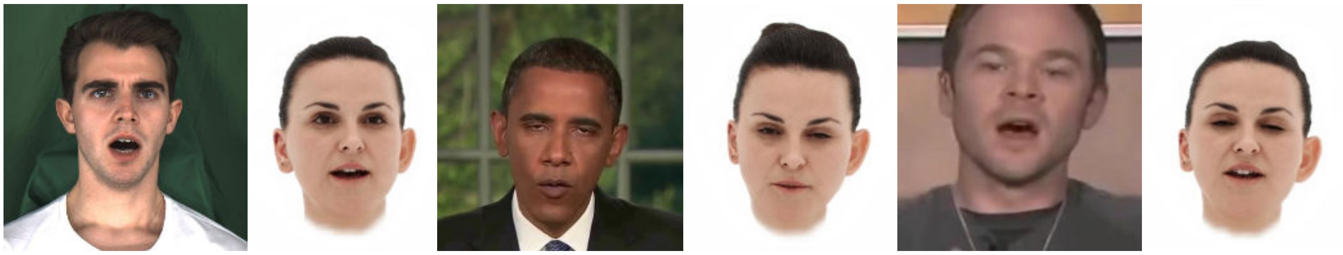

Figure 1: *Results of our video-driven animation approach trained with a single-person dataset.*

Abstract

We present a new approach for video-driven animation of high-quality neural 3D head models, addressing the challenge of person-independent animation from video input. Typically, high-quality generative models are learned for specific individuals from multi-view video footage, resulting in person-specific latent representations that drive the generation process. In order to achieve person-independent animation from video input, we introduce an LSTM-based animation network capable of translating person-independent expression features into personalized animation parameters of person-specific 3D head models. Our approach combines the advantages of personalized head models (high quality and realism) with the convenience of video-driven animation employing multi-person facial performance capture. We demonstrate the effectiveness of our approach on synthesized animations with high quality based on different source videos as well as an ablation study.

CCS Concepts

• *Computing methodologies* → *Machine learning; Computer graphics; Animation; Rendering;*

1. Introduction

Analysis and synthesis of human faces play an important role in many fields such as movie productions, game development, or virtual reality (VR). Especially, video-driven facial animation receives great attention as this technique simplifies otherwise difficult tasks such as video/photo editing, visual dubbing, or the animation of 3D human characters. While recent advances in multi-person facial re-enactment have significantly improved animation quality, challenges persist in achieving both (photo) realism and seamless integration into 3D virtual environments, crucial for creating immersive VR experiences.

On the other hand, high-quality 3D neural head avatars are often created from captured multi-view data of a single individual, resulting in person-specific latent representations driving the generation process. This presents a significant challenge when training multi-person capable video-driven animation models.

In this paper, we propose a new approach for multi-person capable video-driven animation of high-quality 3D neural head avatars, bridging the gap between realism and convenient animation for VR. Our method overcomes the limitations of existing techniques by seamlessly integrating (photo) realism of personalized head avatars into multi-person video-based animation. We employ a hybrid head representation that combines 3D mesh-based geometry, dynamic textures, and neural rendering [PHE23]. In order to drive our neural head model with video footage of an arbitrary person, we extract subject-independent expression features using the method of Feng et al. [FFBB21]. Taking into account the ambiguous mapping between source expression space (person-independent) and target expression space (animated head model), we design our animation model as a recurrent neural network (LSTM) that performs not only a frame-by-frame prediction of animation parameters but considers temporal relationships as well, resulting in more accurate animations even with unforeseen actors. To further improve the animation quality, we augmented the extracted expression features with a learned residual, which simplifies finding a good mapping between source and target expression parameters.

The remainder of this paper is structured as follows. The next section reviews related work before section 3 presents an overview of the proposed approach. Sections 4 and 5 describe the employed neural head model and the video-driven animation approach. Finally, sections 6 and 7 discuss the experimental results and draw a conclusion.

2. Related Work

Face Modelling

Learning controllable models of human faces/heads has been extensively studied over the last decades. While being comparably simple, morphable models [BV99] are still one of the most popular approaches to represent facial expressions as they can be easily extended to increase the quality of facial expressions [BV99; WBLP11; GVWT13; LYYB13; CWZ*14], incorporate additional attributes such as identity, texture, and light [VBPP05; TZN*15] or serve as a face geometry prior for various deep-learning-based methods [TZK*17; KGT*18; CCL*20; BTS*21; GTZN21; GPL*22]. However, a significant drawback of these linear models is the lack of expressiveness to correctly represent complex areas like the oral cavity, eyes, or hair. As a result, purely model-based approaches often employ ’hand crafted’ solutions (e.g. oral cavity) or simply ignore these areas [GVWT13; TZN*15], while hybrid or 2D methods tackle this problem by representing facial performances in geometry and texture space [PKHE17; DSJ*11] or directly in 2D image space [HZSX22; WML21; SLT*19; IZZE16]. In contrast, neural face models are based on deep generative architectures such as variational auto-encoder [KW13] or generative adversarial networks [GPM*14], which can synthesize detailed 3D geometry and high-resolution face textures from a latent expression vector [BLS*21; CBGB20; LBZ*20; PHE20; LSSS18]. More recently, Ma et al. [MSS*21] proposed a codec avatar that supports inference and rendering of neural head models even on mobile devices, whereas Grassal et al. [GPL*22] present an approach for creating a personalized neural head avatar even from monocular video. Apart from traditional mesh and texture-based models, several novel scene representations [MSO*19; SZW19; LSS*19; PCPM20; MST*20; TFT*20; YLT*21; RPLG21; HSM*21; GTZN21] have been presented in the recent years that allow for creating truly photo-realistic renderings of humans. For example, Thies et al. [TZN19] propose a deferred neural rendering approach to create photo-realistic renderings of 3D computer graphic models. Kim et al. [KGT*18] and Prokudin et al. [PBR21] combine neural rendering with parametric models of humans and human heads, which allows for photo-realistic rendering of animatable CG models. Volumetric representations [MSO*19; SZW19; LSS*19; PCPM20; MST*20; TFT*20] often capture the 3D structure and appearance of an object/scene with a non-linear function that depends on the 3D position in space as well as viewing direction and predicts color as well as volume density. This allows for representing fine structures such as hair fibers, objects with complex reflective properties such as glass and metal but also dynamic effects like smoke. A big drawback, however, is the high computational complexity [YLT*21; RPLG21; HSM*21; GTZN21]. While Müller et al. [MESK22] propose a new approach that supports very fast training as well as real-time rendering, they still lack semantic control, which is essential for animation.

Video-driven Facial Animation

In recent years, many methods for multi-person video-driven facial animation have been published, which can be categorized into 2D and 3D approaches.

The main advantage of 2D approaches [ZLG*23; DCK*22; WML21; HKK*20; ZPW*20; SLT*19; GSZ*18] is their simplicity of use as a single image of the target person can be animated with a driving-video of another person requiring no additional preprocessing. A big disadvantage, however, is the lack of 3D information, which results in unrealistic distortions and obvious rendering artifacts as soon as the head pose changes considerably. Several approaches have been proposed to overcome this problem. For example, Wang et al. [WML21] predict explicit features for appearance, expression, head-pose, as well as canonical 3D key points, whereas Hong et al. [HZSX22] introduce a depth-aware generative adversarial network (GAN) that predicts depth values for each face pixel. While both 2D methods can improve the visual quality of resulting videos, the lack of a consistent underlying 3D model still degrades the rendering quality under strong 3D rotations. More importantly, these methods cannot be used to animate faces for 3D applications such as games or virtual reality.

3D-aware methods [FRP*22; DBB22; FFBB21; TZN19; KGT*18] usually fit an existing 3D morphable face model to each frame of a video by predicting the corresponding model parameters (i.e. identity, expression, head pose, etc.) with a convolutional neural network. This captures the 3D facial performance of a person from a monocular video and allows for transferring the captured facial expression to a target person in a different video (e.g. by rendering the animated face model with adapted expression parameters). While this approach works well in general, the underlying linear face models often do not capture/reproduce facial expressions accurately enough and/or they cannot be easily integrated into real-time 3D environments.

Contribution

In this paper, we present a new approach for real-time video-driven animation of a personalized 3D neural head model. Unlike previous methods in multi-person video-driven animation, our approach surpasses the limitations associated with linear morphable face models that frequently exhibit insufficient accuracy in capturing and reproducing intricate facial expressions.

Instead, we employ a personalized high-quality neural head avatar that allows for photo-realistic rendering and convenient integration in 3D scenes. In order to bridge the gap between the driving video and our neural head model, we extract person-independent

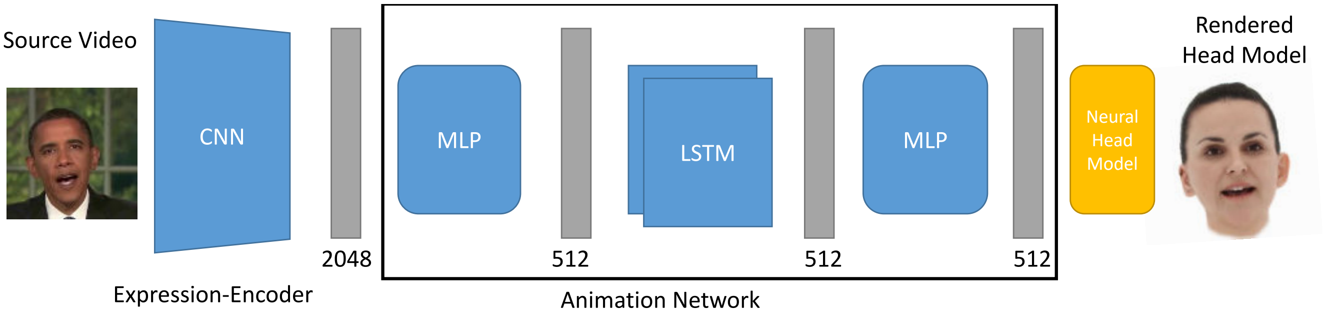

Figure 2: *High-level architecture of the proposed animation network.*

expression features that allow for transferring the facial expression from the source video to our neural head avatar using a recurrent neural network (LSTM). Our contribution thereby enhances the fidelity and realism of video-driven facial animations and therefore allows effortless integration into immersive experiences in virtual environments.

3. System Overview

In the first step, we create a personalized neural head model that represents 3D geometry, head motion, facial expression, and appearance of the captured person, Sec. 4. This model is learned from video data of an actor captured with a multi-view camera rig and can be driven with a latent expression vector. After training the VAE-based head model, the captured facial performance can be represented by a sequence of low-dimensional latent parameter vectors. In the second step, we compute person-independent expression features for each captured video frame using the approach of Feng et al. [FFBB21]. Hence, the complex task of video-driven animation is reduced to transforming a sequence of person-independent expression features into a sequence of animation parameters for our neural head model using a recurrent neural network, Sec. 5. Finally, to increase realism, we employ a neural rendering model that refines rasterization-based images of our head model, which allows for synthesizing photo-realistic videos of the animated head model.

4. Hybrid Head Representation

Our approach is based on a photo-realistic animatable 3D head model [PHE23] that is computed from multi-view video footage to ensure that it perfectly resembles the appearance of the captured person. The model-building process consists of three stages: first, a statistical head model is employed to recover pose and approximate head geometry for each captured frame based on automatically detected landmarks [KS14] and optical flow. In the second step, dynamic head textures are extracted in addition to the approximate geometry to reproduce fine details, small motions, and complex deformations (e.g. in the oral cavity). After geometry recovery and texture extraction, each captured frame is represented by rigid motion parameters $\mathbf{T}$, blend-shape weights $\mathbf{b}$, and an RGB image as texture. Based on this data, we train a deep generative face model (VAE) that reconstructs blend-shape weights $\mathbf{b}$ as well as face textures from a low-dimensional expression vector $\mathbf{z}$, thereby enabling natural, plausible, and realistic facial animation. An adversarial training strategy based on a patch-based discriminator network [IZZE16] helps to improve the texture reconstruction quality. After the model creation, all captured multi-view sequences are represented with a single parameter vector consisting of 3D head pose $\mathbf{T}$ as well as expression vectors $\mathbf{z}$.

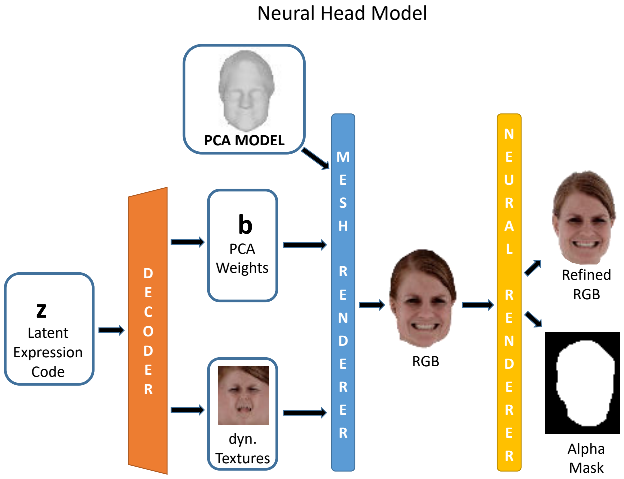

Figure 3: *High-level architecture of the employed neural head model.*

While the convenient mesh-plus-texture representation allows for correctly reconstructing the facial performance as well as most of the appearance, important details (e.g. silhouette, oral cavity, hair, etc.) cannot be reproduced, see figure 5. Therefore, a self-supervised rendering approach based on pixel-to-pixel translation is employed. The render network receives the mesh-based rendering as input and predicts a refined head image as well as weight masks that help to separate foreground from background. This simplifies the training process (i.e. no need to pre-compute foreground masks) and provides a means for integrating the rendered head model with new backgrounds or into 3D scenes.

The image formation model (1) is a convex combination of the mesh-based rendering $\mathcal{I}_{orig}$, a corrective image $\mathcal{I}_{corr}$ and the static

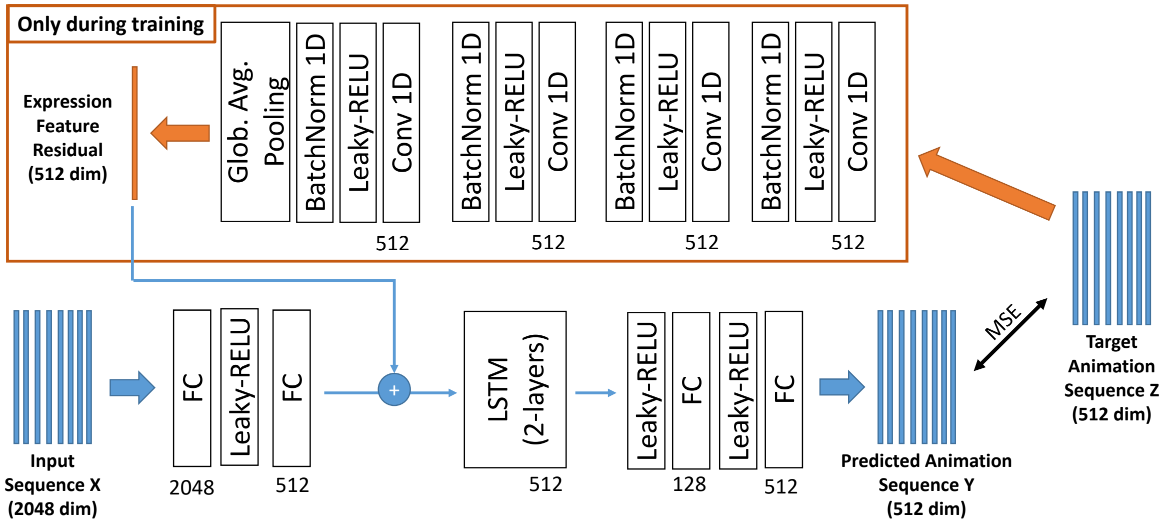

Figure 4: *Architecture of the animation network.*

background $\mathcal{I}_{backg}$, where each image contributes according to spatially varying weight maps α, β, and γ.

$$\begin{aligned}\mathcal{I}_{out} = \alpha\,\mathcal{I}_{orig} + \beta\,\mathcal{I}_{corr} + \gamma\,\mathcal{I}_{backg}\\ \alpha+\beta+\gamma = 1\end{aligned} \tag{1}$$

The real-time re-rendering network is based on a U-Net architecture [RPB15] where the input tensor contains the RGB colors of the mesh-based rendering and the output tensor consists of six channels: RGB color plus three channels that contain the weight maps α, β, and γ. Training the render model requires only the captured frame, the mesh-based rendering $\mathcal{I}_{orig}$ of the textured head model, an empty background frame (i.e. clean plate) $\mathcal{I}_{backg}$ while an alpha-mask for foreground/background segmentation is learned automatically in an unsupervised manner. For more details please refer to [PHE23].

5. Video-Based Neural Animation

After the creation of the neural head model, each frame of the captured multi-view video footage is encoded into a low dimensional latent vector **z**. This vector represents the facial expression and head pose of the actor captured in the footage. However, as the training database only consists of visual information of a single individual, training a multi-person animation approach directly poses a challenge. To overcome this limitation, we extract person-independent expression features **x** from the captured video footage using the method of Feng et al. [FFBB21], which was trained with a large number of individuals showing different facial expressions and head poses. This way, each frame of our multi-view data is labeled with a person-independent expression information **x** as well as a latent expression vector **z** that drives our neural head model. In order to animate the neural head avatar, it is essential to predict the corresponding animation parameter **z** from each expression feature **x**. Due to the inherent ambiguity of this mapping, we employ a recurrent architecture (LSTM) that performs not only a frame-wise mapping but also captures temporal relationships between sequences of input features and animation parameters, figure 4.

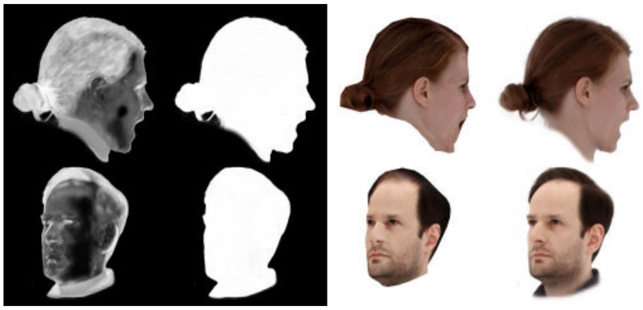

Figure 5: *This figure illustrates the quality of the employed hybrid head model. The leftmost column shows the refinement weights β followed by the foreground mask $\mathcal{F}$, the initial mesh-based rendering, and the final output $\mathcal{I}_{out}$. High intensity in the refinement mask indicates strong corrections (e.g. neck, hair, sometimes mouth), while low intensity indicates that the mesh-based rendering provides already correct pixel colors.*

However, since the predicted expression weights and jaw angles of Feng et al. [FFBB21] are optimized for the FLAME face model [LBB*17], we use the 2048-dimensional output feature vector of an earlier layer (output of the ResNet50). During our experiments, we found that computing a small expression residual feature (during training) with a CNN from the target animation parameters helps to further improve the animation quality. This CNN consists of four 1D convolutions (with kernel size 5, zero-padding of 2) followed by a Leaky-ReLU and BatchNorm. In the end, an

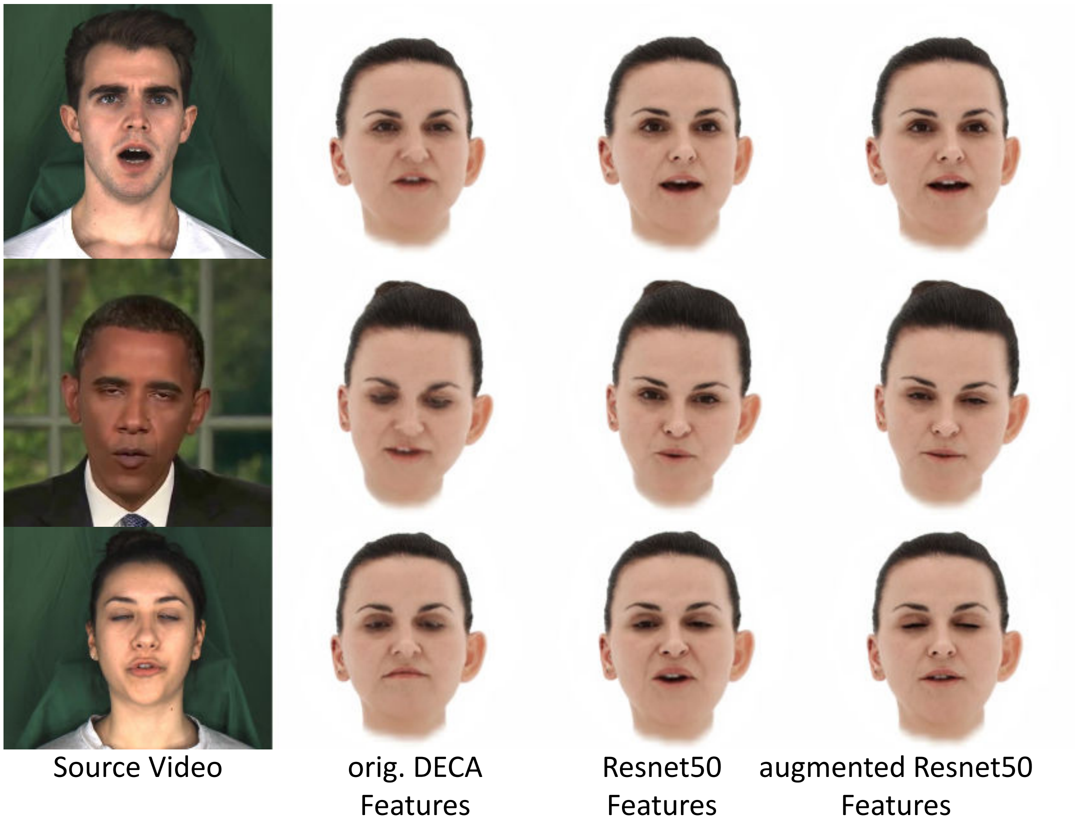

Figure 6: *Visual comparison of the facial animation based on different input features.*

average-pooling layer outputs the mean expression residual over all 8 frames, which is then injected into the animation network, see 4. Our animation network is trained batch-wise (32) with short sequences (8 frames) of **x** transformed into equally long sequences of **y** minimizing the mean squared error (MSE) between **y** and the target expression **z**. All Leaky-ReLU layers have a leakiness of $1.0e-2$. We train the network for 15000 iterations using the Adam optimizer with a learning rate of $1.0e-4$ and exponential learning rate scheduling with $\gamma = 0.96$.

6. Experimental Results and Discussion

This section presents our evaluation, and visual results generated with the proposed method as still images, while an accompanying video demonstrating dynamic effects can be found in the supplementary material. For our experiments, we captured an actress with a synchronized and calibrated multi-view camera rig consisting of three cameras (frontal, diagonally left/right) at eye level. We captured different facial expressions as well as speech (single words and monologues in English). The actress was not restricted in terms of the presented emotion. The effective capture resolution for the head is approximately 520x360 pixels. All pre-processing steps, network training, and experiments have been carried out on a regular desktop computer with 64GB Ram, 2.6 GHz CPU (14 cores with hyper-threading), and one GeForce RTX3090 graphics card. The captured data was split into four sequences with a total length of approximately 3 minutes for training and one test sequence with a total duration of approximately 30 seconds. We evaluated the proposed network architecture with different input features: the original DECA expression features (blend shape weights and jaw angles), earlier expression features produced by the Resnet50, and Resnet50 features augmented with an auxiliary feature vector that helps to disambiguate the mapping between input expression feature and target animation parameter. For inference, we use a zero residual feature vector for animation.

During development, we visually compared an MLP-based, CNN-based, and LSTM-based architecture and found that the LSTM-based performs best. Moreover, we use rather short input sequences of 8 frames as we found that more temporal context does not yield better animation results anymore.

Figure 6 illustrates the differences in the resulting animation based on the tested input features. In our experiments, we found that the generated talking head videos based on the Resnet50 features appear to be more realistic and included fewer artifacts. Especially, the augmented Resnet50 features yield livelier and more natural animations, which can be seen in the supplemental video. We evaluate our animation method against three recent approaches (DAGAN [HZSX22], LIA [WYBD22], FADM [ZLG*23]) for the generation of photo-realistic talking heads based on the driving video of an arbitrary person. Figure 7 illustrates the rendering and animation quality of all approaches based on video samples taken

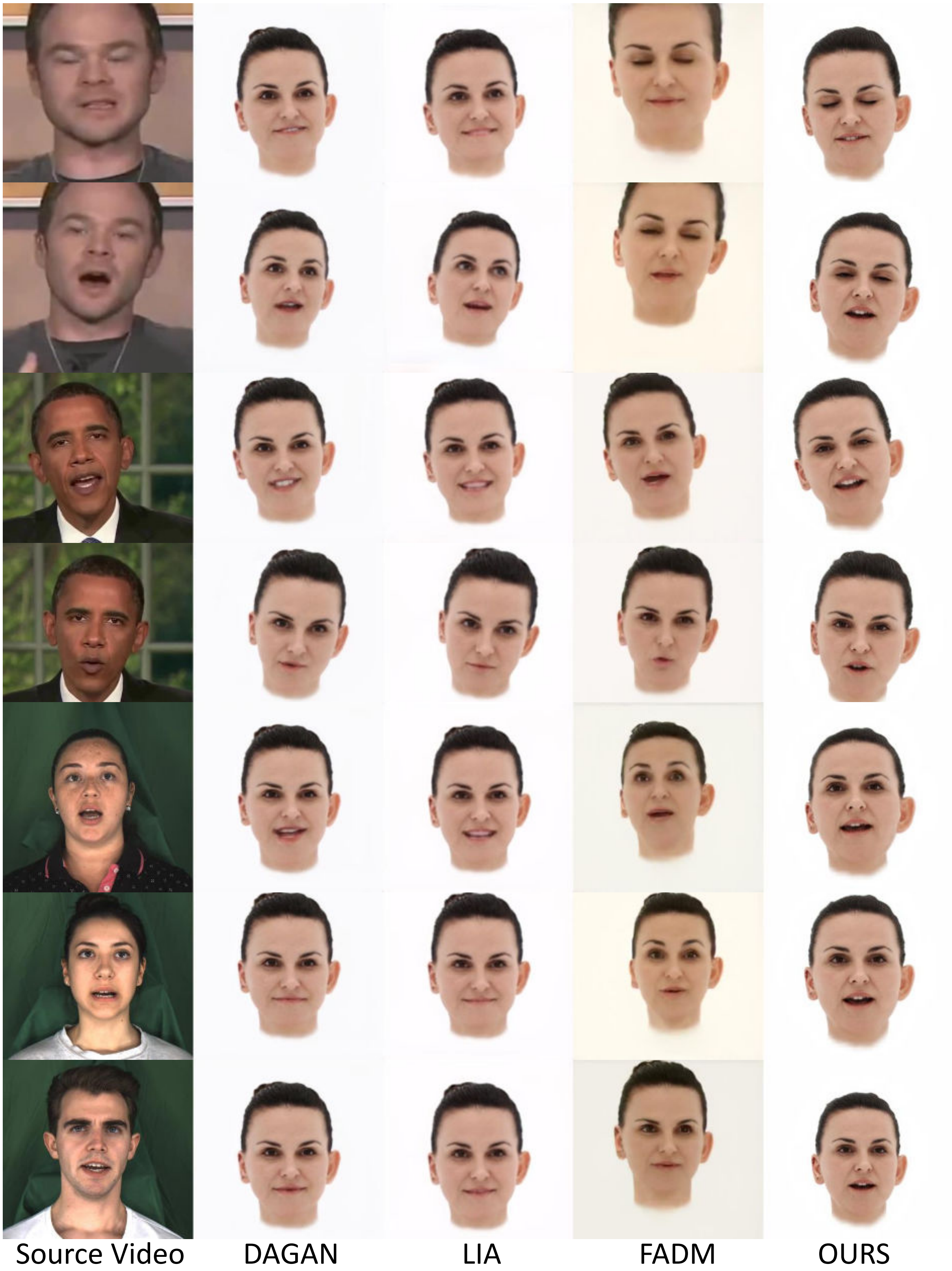

Figure 7: *Comparison of the proposed animation approach with 3 recent methods for multi-person video-driven face animation.*

from the VoxCeleb2 [CNZ18] dataset, MEAD [WWS*20] dataset, and Obama's weekly address footage [SSK17]. The main advantage of our approach is that a high-quality neural head model can be connected with a multi-person capable video-driven animation approach, which results in the higher visual quality of the synthesized videos, more visible details, fewer rendering artifacts, and more natural animations. Our intuition on why the residual features improve animation during inference is that they reduce the likelihood that the network learns spurious correlations between input expression features and target animation parameters. Expression differences that can neither be explained by the original Resnet50 features nor by the temporal context can be represented by the residual features, which are computed from the target animation parameters. Additionally, providing only an average residual feature per training sequence prevents the network from relying too much on the these artificial features.

There are also limitations: in order to achieve high visual quality a personalized neural head model is created, however this requires retraining if a new virtual character has to be integrated. Currently, the residual features are only used during training but not for inference. However, with a suitable generative model, they could enable fine-tuning of the generated animations also during inference as they capture further expression details that cannot be explained by the original input features.

7. Conclusions

We present a new method for the animation of 3D neural head models. Our method extracts person-independent expression features from monocular video and translates them successfully into realistic animation parameters for our neural head model. This allows for animating high-quality 3D head avatars by arbitrary actors even though the model is generated only from captured data of a single person. For more robust training, we augment the extracted expression features, which helps to disambiguate the mapping between source expression features and target animation space. We show that our neural head model can be successfully animated from arbitrary persons and compare our approach against recent methods for video-driven facial re-enactment demonstrating the high quality of our animation results.

Acknowledgments

This work has partly been funded by the European Union under grant agreement No 101070672 (SPIRIT), by the German Federal Ministry of Education and Research (Voluprof, grant no. 16SV8705), by the German Federal Ministry for Economic Affairs and Climate Action (ToHyVe, grant no. 01MT22002A) as well as the Fraunhofer Society in the Max Planck-Fraunhofer collaboration project NeuroHum.

References

[BLS*21] Bi, Sai, Lombardi, Stephen, Saito, Shunsuke, et al. "Deep Relightable Appearance Models for Animatable Faces". *ACM Trans. Graph.* 40.4 (July 2021). ISSN: 0730-0301 2.

[BTS*21] B R, Mallikarjun, Tewari, Ayush, Seidel, Hans-Peter, et al. "Learning Complete 3D Morphable Face Models from Images and Videos". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021 2.

[BV99] Blanz, Volker and Vetter, Thomas. "A Morphable Model for the Synthesis of 3D Faces". *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, 187–194. ISBN: 0-201-48560-5 2.

[CBGB20] Chandran, Prashanth, Bradley, Derek, Gross, Markus, and Beeler, Thabo. "Semantic Deep Face Models". *Proc. International Conference on 3D Vision (3DV)*. 2020, 345–354. DOI: 10.1109/3DV50981.2020.00044 2.

[CCL*20] Chai, Xiaoyu, Chen, Jun, Liang, Chao, et al. "Expression-Aware Face Reconstruction Via A Dual-Stream Network". *IEEE International Conference on Multimedia and Expo, ICME 2020, London, UK, July 6-10, 2020*. IEEE, 2020, 1–6 2.

[CNZ18] Chung, J. S., Nagrani, A., and Zisserman, A. "VoxCeleb2: Deep Speaker Recognition". *INTERSPEECH*. 2018 7.

[CWZ*14] Cao, Chen, Weng, Yanlin, Zhou, Shun, et al. "FaceWarehouse: A 3D Facial Expression Database for Visual Computing". *IEEE Transactions on Visualization and Computer Graphics* 20.3 (Mar. 2014), 413–425. ISSN: 1077-2626 2.

[DBB22] Danecek, Radek, Black, Michael J., and Bolkart, Timo. "EMOCA: Emotion Driven Monocular Face Capture and Animation". *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 20311–20322 2.

[DCK*22] Drobyshev, Nikita, Chelishev, Jenya, Khakhulin, Taras, et al. "MegaPortraits: One-shot Megapixel Neural Head Avatars". 2022 2.

[DSJ*11] Dale, Kevin, Sunkavalli, Kalyan, Johnson, Micah K., et al. "Video Face Replacement". *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 30 (2011) 2.

[FFBB21] Feng, Yao, Feng, Haiwen, Black, Michael J., and Bolkart, Timo. "Learning an Animatable Detailed 3D Face Model from In-The-Wild Images". Vol. 40. 8. 2021 1–4.

[FRP*22] Filntisis, Panagiotis P., Retsinas, George, Paraperas-Papantoniou, Foivos, et al. "Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos". *arXiv preprint arXiv:2207.11094* (2022) 2.

[GPL*22] Grassal, Philip-William, Prinzler, Malte, Leistner, Titus, et al. "Neural head avatars from monocular RGB videos". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 18653–18664 2.

[GPM*14] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, et al. "Generative Adversarial Nets". *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, 2672–2680 2.

[GSZ*18] Geng, Jiahao, Shao, Tianjia, Zheng, Youyi, et al. "Warp-guided GANs for Single-photo Facial Animation". *ACM Trans. Graph.* 37.6 (Dec. 2018), 231:1–231:12. ISSN: 0730-0301. DOI: 10.1145/3272127.3275043 2.

[GTZN21] Gafni, Guy, Thies, Justus, Zollöfer, Michael, and Niessner, Matthias. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction". *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021 2.

[GVWT13] Garrido, Pablo, Valgaert, Levi, Wu, Chenglei, and Theobalt, Christian. "Reconstructing Detailed Dynamic Face Geometry from Monocular Video". *ACM Trans. Graph.* 32.6 (Nov. 2013), 158:1–158:10. ISSN: 0730-0301 2.

[HKK*20] Ha, Sungjoo, Kersner, Martin, Kim, Beomsu, et al. "Marionette: Few-shot face reenactment preserving identity of unseen targets". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, 10893–10900 2.

[HSM*21] Hedman, Peter, Srinivasan, Pratul P., Mildenhall, Ben, et al. "Baking Neural Radiance Fields for Real-Time View Synthesis". *arXiv* (2021) 2.

[HZSX22] Hong, Fa-Ting, Zhang, Longhao, Shen, Li, and Xu, Dan. "Depth-Aware Generative Adversarial Network for Talking Head Video Generation". 2022 2, 5.

[IZZE16] Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. "Image-to-Image Translation with Conditional Adversarial Networks". *arxiv* (2016) 2, 3.

[KGT*18] Kim, Hyeongwoo, Garrido, Pablo, Tewari, Ayush, et al. "Deep Video Portraits". *ACM Trans. Graph.* 37.4 (July 2018). issn: 0730-0301 2.

[KS14] Kazemi, Vahid and Sullivan, Josephine. "One millisecond face alignment with an ensemble of regression trees". *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, 1867–1874 3.

[KW13] Kingma, Diederik P and Welling, Max. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML] 2.

[LBB*17] Li, Tianye, Bolkart, Timo, Black, Michael. J., et al. "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 4.

[LBZ*20] Li, Ruilong, Bladin, Karl, Zhao, Yajie, et al. "Learning Formation of Physically-Based Face Attributes". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, 3407–3416 2.

[LSS*19] Lombardi, Stephen, Simon, Tomas, Saragih, Jason, et al. "Neural Volumes: Learning Dynamic Renderable Volumes from Images". *ACM Trans. Graph.* 38.4 (July 2019). issn: 0730-0301 2.

[LSSS18] Lombardi, Stephen, Saragih, Jason M., Simon, Tomas, and Sheikh, Yaser. "Deep Appearance Models for Face Rendering". *CoRR* abs/1808.00362 (2018) 2.

[LYYB13] Li, Hao, Yu, Jihun, Ye, Yuting, and Bregler, Chris. "Realtime Facial Animation with On-the-fly Correctives". *ACM Trans. Graph.* 32.4 (July 2013), 42:1–42:10. issn: 0730-0301 2.

[MESK22] Müller, Thomas, Evans, Alex, Schied, Christoph, and Keller, Alexander. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". *ACM Trans. Graph.* 41.4 (July 2022), 102:1–102:15 2.

[MSO*19] Mildenhall, Ben, Srinivasan, Pratul P., Ortiz-Cayon, Rodrigo, et al. "Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines". *ACM Trans. Graph.* 38.4 (July 2019). issn: 0730-0301 2.

[MSS*21] Ma, Shugao, Simon, Tomas, Saragih, Jason M., et al. "Pixel Codec Avatars". *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. 2021, 64–73 2.

[MST*20] Mildenhall, Ben, Srinivasan, Pratul P., Tancik, Matthew, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". *ECCV*. 2020 2.

[PBR21] Prokudin, Sergey, Black, Michael J., and Romero, Javier. "SMPLpix: Neural Avatars from 3D Human Models". *Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2021, 1810–1819 2.

[PCPM20] Pumarola, Albert, Corona, Enric, Pons-Moll, Gerard, and Moreno-Noguer, Francesc. "D-NeRF: Neural Radiance Fields for Dynamic Scenes". *arXiv preprint arXiv:2011.13961* (2020) 2.

[PHE20] Paier, Wolfgang, Hilsmann, Anna, and Eisert, Peter. "Interactive Facial Animation with Deep Neural Networks". *IET Computer Vision, Special Issue on Computer Vision for the Creative Industries* 14.6 (Sept. 2020), 359–369 2.

[PHE23] Paier, Wolfgang, Hilsmann, Anna, and Eisert, Peter. *Unsupervised Learning of Style-Aware Facial Animation from Real Acting Performances*. 2023. arXiv: 2306.10006 [cs.CV] 1, 3, 4.

[PKHE17] Paier, Wolfgang, Kettern, Markus, Hilsmann, Anna, and Eisert, Peter. "A Hybrid Approach for Facial Performance Analysis and Editing". *IEEE Trans. on Circuits and Systems for Video Technology* 27.4 (Apr. 2017), 784–797. issn: 1051-8215 2.

[RPB15] Ronneberger, O., P.Fischer, and Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. LNCS. (available on arXiv:1505.04597 [cs.CV]). Springer, 2015, 234–241 4.

[RPLG21] Reiser, Christian, Peng, Songyou, Liao, Yiyi, and Geiger, Andreas. "KiloNeRF: Speeding up Neural Radiance Fields with Thousands of Tiny MLPs". *CoRR* abs/2103.13744 (2021) 2.

[SLT*19] Siarohin, Aliaksandr, Lathuilière, Stéphane, Tulyakov, Sergey, et al. "First Order Motion Model for Image Animation". *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2019 2.

[SSK17] Suwajanakorn, Supasorn, Seitz, Steven M., and Kemelmacher-Shlizerman, Ira. "Synthesizing Obama: Learning Lip Sync from Audio". *ACM Trans. Graph.* 36.4 (July 2017). issn: 0730-0301 7.

[SZW19] Sitzmann, Vincent, Zollhoefer, Michael, and Wetzstein, Gordon. "Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations". *Advances in Neural Information Processing Systems*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., et al. Vol. 32. Curran Associates, Inc., 2019 2.

[TFT*20] Tewari, A., Fried, O., Thies, J., et al. "State of the Art on Neural Rendering". *Computer Graphics Forum (EG STAR 2020)* (2020) 2.

[TZK*17] Tewari, Ayush, Zollöfer, Michael, Kim, Hyeongwoo, et al. "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction". *The IEEE International Conference on Computer Vision (ICCV)*. 2017 2.

[TZN*15] Thies, J., Zollhöfer, M., Niessner, M., et al. "Real-time Expression Transfer for Facial Reenactment". *ACM Transactions on Graphics (TOG)* 34.6 (2015) 2.

[TZN19] Thies, Justus, Zollhöfer, Michael, and Niessner, Matthias. "Deferred neural rendering: image synthesis using neural textures". *ACM Trans. Graph.* 38.4 (2019), 66:1–66:12 2.

[VBPP05] Vlasic, Daniel, Brand, Matthew, Pfister, Hanspeter, and Popović, Jovan. "Face Transfer with Multilinear Models". *ACM Trans. Graph.* 24.3 (July 2005), 426–433. issn: 0730-0301 2.

[WBLP11] Weise, Thibaut, Bouaziz, Sofien, Li, Hao, and Pauly, Mark. "Realtime Performance-Based Facial Animation". *ACM Trans. Graph.* 30.4 (July 2011). issn: 0730-0301 2.

[WML21] Wang, Ting-Chun, Mallya, Arun, and Liu, Ming-Yu. "One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021 2.

[WWS*20] Wang, Kaisiyuan, Wu, Qianyi, Song, Linsen, et al. "MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation". *ECCV*. Aug. 2020 7.

[WYBD22] Wang, Yaohui, Yang, Di, Bremond, Francois, and Dantcheva, Antitza. "Latent Image Animator: Learning to Animate Images via Latent Space Navigation". *International Conference on Learning Representations*. 2022 5.

[YLT*21] Yu, Alex, Li, Ruilong, Tancik, Matthew, et al. "PlenOctrees for Real-time Rendering of Neural Radiance Fields". *ICCV*. 2021 2.

[ZLG*23] Zeng, Bohan, Liu, Xuhui, Gao, Sicheng, et al. *Face Animation with an Attribute-Guided Diffusion Model*. 2023. arXiv: 2304.03199 [cs.CV] 2, 5.

[ZPW*20] Zeng, Xianfang, Pan, Yusu, Wang, Mengmeng, et al. "Realistic face reenactment via self-supervised disentangling of identity and pose". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, 12757–12764 2.