Facial Reconstruction and Rendering for Expression Conditioned Animation

Rajath Jayashankar[*]

Rutgers University - CS 549
Spring 2022

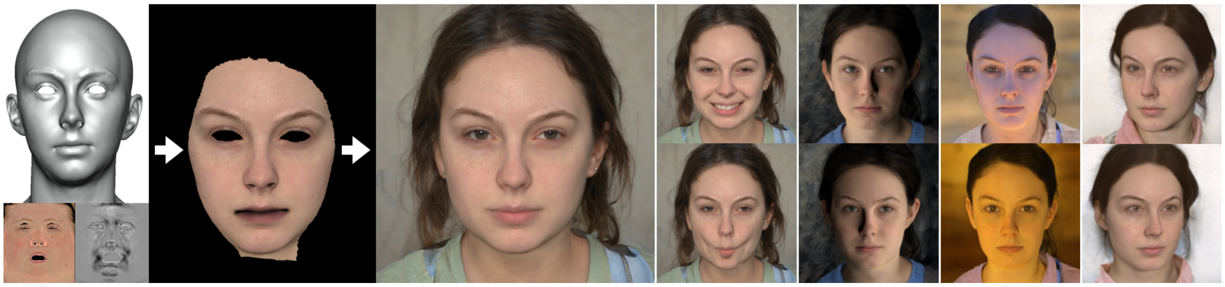


Figure 1: Results from Rendering with Style [6], a combination of traditional and neural rendering approaches to create high quality face renders. Given a 3D face geometry and ray-traces skin as prior, projection is done on neural image generator to in-paint non-skin pixels. We can results of facial expression parameter changes, lighting changes, environment maps and viewpoint in the above results.

Abstract

Facial reconstruction has been a longstanding research problem in the fields of Computer Vision and graphics. There are currently many methods that allow us to reconstruct monocular 3D faces with fine geometric details. However, many of these methods do not provide methods to realistically animate or transfer expressions to these models, let alone in real-time. In this report, I examine possible improvements in three SOTA methods through a Multi-stage joint optimization of expressive 3D Facial Reconstruction and Neural Rendering methods and how they could be combined to provide better results.

First, I train DECA (Detailed Expression Capture and Animation) [10] with an expression labeled dataset to allow for better expression transfer during animation. The dataset is Multi-view Emotional Audio-visual Dataset (MEAD) [26] DECA produces a UV displacement map from a low-dimensional representation which consists of parameters that are person-specific and generic parameters while disentangling these person-specific parameters and expression dependant winkles. Second, I try to combine the FLAME (Faces Learned with an Articulated Model and Expressions) model with current methods of neural rendering of faces, such as Rendering with Style [6]. Finally, I combine the model from the first approach with the current SOTA Neural Rendering technique, NerFace [11].

Index Terms: Computing methodologies—Mesh model—;

1 Introduction

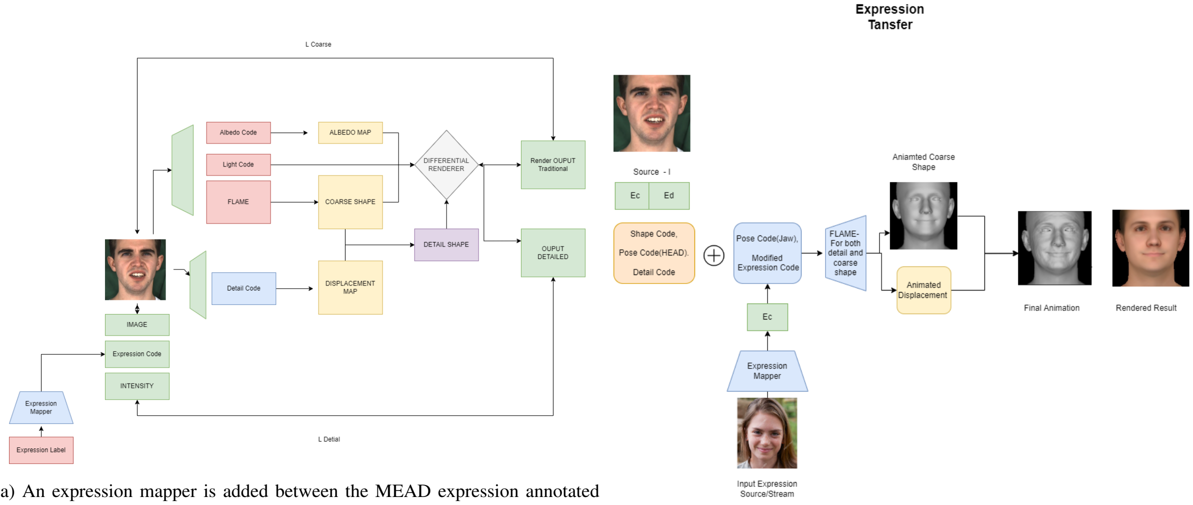
It has been two decades since the pioneering work of Vetter and Blanz [25] that first showed a method to reconstruct the 3D shape and surface texture of a face from a single image. With the advent of immersive virtual environments, AR/VR, and the current push for virtual telepresence, there has never been a greater need for methods that can capture and render realistic digital humans efficiently while requiring little artist effort. Creating a 3D facial avatar [15], either from a studio capture setup or from facial reconstruction, is arduous work and involves manual fine-tuning. 3D face reconstruction methods have rapidly advanced, enabling 3D avatar creation for mixed reality applications, video editing, image synthesis [12], face recognition [4], or speech-driven facial animation [9]. To make the process manageable, most existing methods [1, 7] need a prior either about the geometry or the appearance for the models to work, leveraging pre-computed 3D face models [5]. These models reconstruct the coarse face structure but fail to capture the fine geometric details such as expression-dependent wrinkles, which are essential for realism and support human emotion analysis.

The current mesh based methods require high-quality 3D training scans to recover detailed facial geometry or lack detail when occluded [13]. These methods do not decouple the expression dependant features such as wrinkled and hence lack realism during animation. Previous methods that have been trained on expression conditioned models, do not generalize well across in the wild images or model the expression dependant details as part of the appearance map instead of the geometry, hence preventing realistic mesh relighting.

Representing a human face with material properties and features such as skin texture [3], eyes which tend to be reflective, and complex hair structure [14] is challenging problem and commonly overlooked by mesh based methods. To handle material properties and complex geometry in facial avatars, dynamic neural radiance fields [11] has been a new approach. This is a neural rendering approach combining classic volume rendering with a neural scene representation network to produce head poses and expressions. This can be achieved using only monocular images from a video. The learned scene representation which is key to capturing hair, interior mouth regions, which is the key difference from the before mentioned methods.

In this paper I have tried to combine the above mentioned

[*]E-mail: rajath.jay@rutgers.edu



(a) An expression mapper is added between the MEAD expression annotated dataset to train DECA coarse details pipeline. It learns an expression-conditioned displacement model by leveraging detail consistency information from multiple images of the same individual with different expressions.

(b) Once trained, DECA animates a face by combining the reconstructed source identity's shape, head pose, and detail code, with the reconstructed source expression's jaw pose and the same expression mapper can be used either as parameter values or an image to get the animated coarse shape and displacement map.

Figure 2: Expression condition DECA training and animation pipelines.

methods to improve upon existing SOTA implementations. The first method is an modification on DECA [10] by using emotion annotated MEAD data set to create a expression-conditioned detail model. Here I have use the expression parameters from MEAD and mapped it to the FLAME model expression components for animation and detail model training. The second approach involved modifying the stylerenderer pipeline to use the expression conditioned model from before and then do a FLAME to BFM [21] conversion. This model was then ray-traces from inital texture map from input image to get ray-traces prior required for StlyeRenderer. However, the ray-traces results weren't detailed enough and couldn't be used further down the pipeline.

The final method involves using DECA and landmark fitting to extract per-frame camera, shape, and expression parameters. Next, these parameters are used to render the FLAME model geometry. This model is then used as a spatial prior on ray sampling where only points that lie on rays that intersect the model are affected by the expression parameters. Finally, given the i-th frame, we shoot rays, we sample points along them and input these points to the deformable NeRF.

2 Related Work

I have tried both a mesh based approach and neural rendering to represent and generate images of human faces. These are related to recent approaches of generic light weight head models and approaches on neural scene representation respectively. Im the following I will discuss some of the most related literature.

Face Reconstruction based on a Morphable Model: A summary of facial reconstruction methods can be found to the state of the art report of Zollhofer [27]. My approach for neural rendering is based on the FLAME model which acts the morphable model which is the building block for many facial reconstruction and animation approaches. There have also been approaches that decouple the coarse structure like DECA [10] and learn facial features like wrinkles that are added onto the morphable model separately to get more realistic deformations. Some methods are based on dynamically adapting the blendshape basis [17] compensate for coarse geometry by non-rigid mesh deformation [8]. In my final approach I represent the geometry and appearance using deep neural network and use volumetric rendering.

Human Avatar Reconstruction: The main goal of facial reconstruction and animation is to be able to reproduce photo-realistic images of a head of a person. This is essential in avatar creation where monocular input stream must be reproduced. There are multiple methods the reconstruct faces based on much monocular input streams. One methods is based on static pose and expression to reconstruct the head via multi-view stereo [16]. Another approach is to combine digitizes faces with reconstruction of hair to estimate the head geometry and appearance from a single image [15].

Neural Scene Representation Networks: Current neural rendering techniques have a neural scene representation network ask their building block for neural reconstruction. A summary of the state of the art can be found in the report by Tiwari [24]. Scene representation networks were introduced by Sitzmann [23]. A neural network represent the geometry and appearance of an object as sampled points in space. To sample from the neural network, we use ray marching. The compact nature of neuarl scene represenatation allows it to be immune to limited resolution like discrete grid structures such as Deep Voxels [22] and Neural Volumes [18]. This idea was extended by Mildenhall [20] to store fields in neural networks. Key contributions are the inclusion of positional encoding for higher detailed reconstructions and volumetric integration. Most of the techniques have static objects that involve Neural Sparse Voxel Fields, use of positional encoding [2] and in the wild training including appearance interpolation [19]. However, these methods assume a static object, while faces are dynamic and have a continuous changing surface.

3 Method

In the three approaches, I have used FLAME, a statistical 3D head model that combines separate linear identity shape and expression spaces pose dependant corrective blendshapes and lines blend skinning (LBS) to articulate jaw, eyes and neck. The parameters for facial

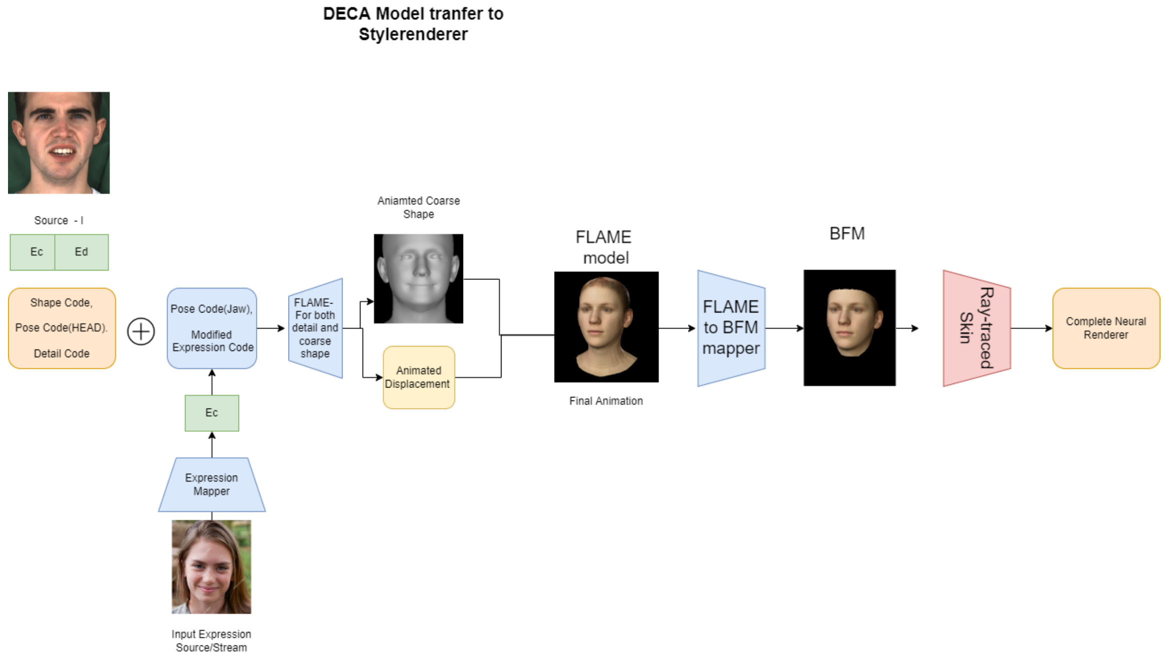


Figure 3: The above pipeline shows how the Rendering with style pipeline is modified to use DECA model, to create an animated FLAME model, that is mapped to the required a priori .

identity is give by $\beta \in R^{\|\beta\|}$, expression $\phi \in R^{\|\phi\|}$, pose $\theta \in R^{3k+3}$, gives a mesh with $n = 5023$ vertices. The model is defined as:

$$M(\beta,\theta,\phi) = W(T_p(\beta,\theta,\phi_{)},J(\beta),\theta,\Psi) \tag{1}$$

The blend skinning function is given by $W(T,J,\theta,\Psi)$, which rotates the vertices T around the joints J, smoothed by beldnweights ϕ.

FLAME does not have an appearance model and hence we need to use Basel Face Model's linear albedo in the FLAME UV layout. FLAME is able to generate face geometry with various poses, expression and shapes as input. This is however a low mesh representation, and miss some mid-frequency surface details.

3.1 Expression Conditioned DECA

The key idea of DECA lies in the fact that an individuals face shows different details depending on facials expressions but many other properties remain unchanged. Therefore it proposes to disentangle the static and dynamic facial detail training. Therefore DECA learns and expression-conditioned detail model to infer facial details from both the expression space and the detial latent space.

The coarse reconstruction pipeline was unchanged and is learned based on the FLAME model space in an analysis-by-synthesis manner. Given a 2D input image I, it's encoded to latent code. A encoder is trained which consists of a ResNet50 network followed by a fully connected layer, to regress a low-dimensional latent code. This code consists of the FLAME parameters (β,θ,ϕ) (which represent the coarse geometry). This coarse geometry uses the firs 100 FLAME shape parameters (β), 50 albedo parameters (α) and 50 expression parameters (ϕ).

In the detail reconstruction, I augment the coarse FLAME geometry with a detailed UV displacement map. Similar to the coarse reconstruction, and encoder is trained here using the expression labelled MEAD dataset. These latent code represent the subject specific facial expression and features. This is then mapped with FLAME expression ϕ and jaw pose parameters.

3.2 DECA model on Rendering with Style

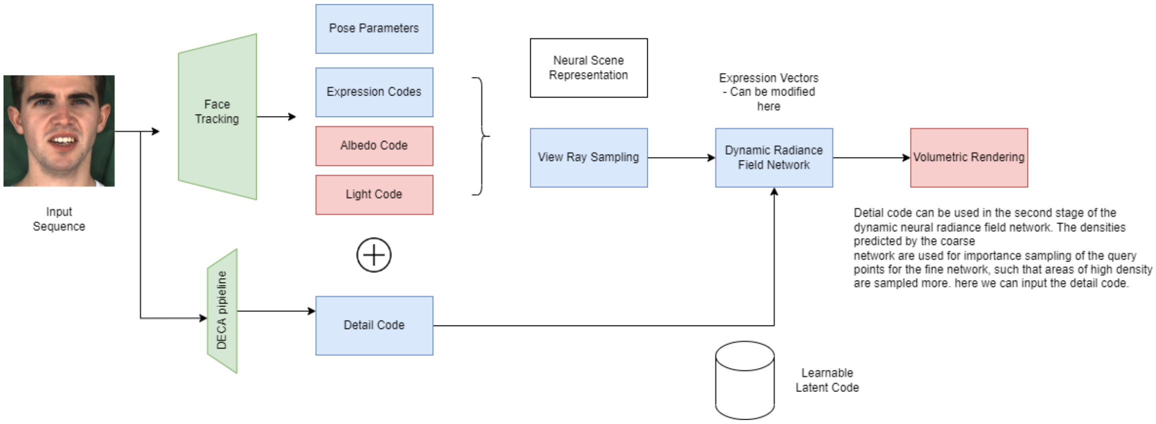
Rendering with style requires as a prior the incomplete facial scans and inpaints using a pre-trained neural face model. Therefore, to create the prior I used the DECA model to get the pose code and modified expression code. These code were used on the previous model to create the animated coarse shape and displacement map. These were combined to get the final shape that were ray-traced to produce the incomplete facial scans.

However, the ray-traced results from transformed Basel Face Model was not able to produce needed outputs as a FLAME to BFM mapping was missing parameters such as shape variance and average texture. There were additional issues with the transformations on BFM model. Rendering with Style considers high-quality 3D facial geometry and appearance maps as a priori either through capture methods or synthesis. The synthesised appearance maps were not of sufficient quality to produce reasonable outputs.

3.3 Neural Control of Radiance Fields with FLAME

Recent advances in neural rendering methods for controllable synthesis has been though Neural rendering techniques. I first model a dynamic scene and provide arbitrary control of facial expressions. The dynamic scene is modelled using a Neural Radiance Field with per-point deformation to allow for head movement, The dynamics of human face, are captures using a state-of-the-art face tracking methods. This gives us the low-dimensional expression parameters to produce the FLAME model. These parameters can be modified to give different expressions per -frame. In order to ensure disentanglement between the expression and view parameters, I have used a spatial prior on ray sampling during training.

A neural radiance field (NeRF) is a continuous function , that, given a 3D point of a scene and the direction in which the rays hit (the viewing direction) gives us the color $c = (r,g,b)$ and density σ.



Detial code can be used in the second stage of the dynamic neural radiance field network. 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. here we can input the detail code.

Figure 4: Given an input sequence video and an image without the person is segmented out as input, we apply facial expression tracking using a 3D morphable model. Based on the input or estimated expression and pose we volumetrically render to synthesize the image of the face. The samples along the camera view is given as input to the dynamic radiance field.

Here, I have used a multi layer perceptron (MLP) and positional encoding defined as m frequency bands.

Expression control in deformable neural radiance fields is achieved through changing the expression parameters provided by FLAME. For each frame we may modify the expression parameter β_i, and provide it to the deformation function D, where ω_i and ϕ_i are the deformation and appearance code.

4 Results

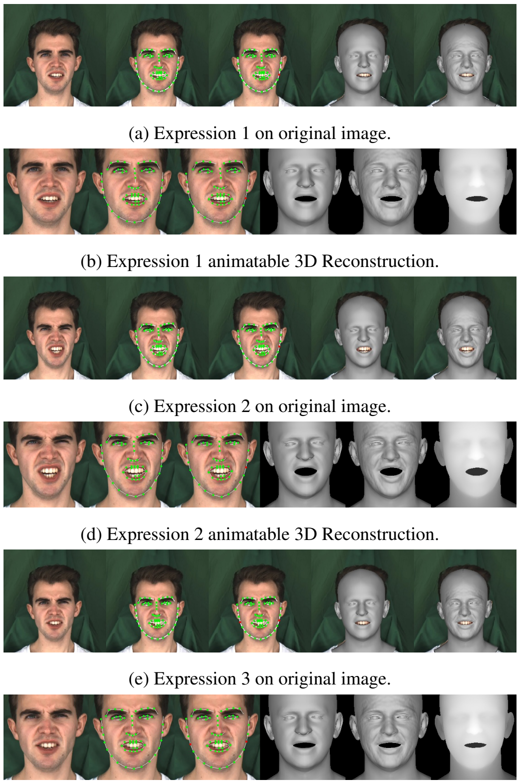
Given a single input image as input, the first methods reconstructs a 3D face shape with fine geometric details and maintains mid-frequency details. 5 shows the results on the MEAD dataset, showing both the coarse and detail structure. I use qualititave analysis to test the results as there is no ground truth for the result 3D reconstrutions and hence only a qualitative visual analysis is possible. It can be observed that even in extreme facial expression associated with the MEAD dataset, the trained model performs well, and mouth and facial wrinkles generalize well. The learned DECA model that is expression conditioned doesn't show any significant improvement over the baseline DECA model. However we are able to map in expression parameters externally due the mapper function which can be helpful in creating a animatable expression based model.

Fig. 7 shows the results from the initial training of the neural renderer. Currently I am working on improving this output by the above mentioned methods. We can observe that the baseline results are itself really good, reflection such as the ones on the glasses, skin texture etc, are all reconstructed well.

5 Conclusion and Future Work

In this paper, I have briefly discussed the three approaches on facial reconstruction that tries to improve upon the current state of the art methods. I first create an expression parameterised DECA model that was used further for other methods based on a mesh based model. I then move on to Neural rendering and Neural Radiance Fields methods for monocular facial reconstruction. I prioritised decoupling the expression parameters to be able to animate and keep intricate facial details required frame-by–frame. This method is suitable for avatar creation and enables us to create videos of avatars. The reconstructed avatars can be rendered photo-realistically under novel poses and expressions.

Future work would be to include a prior to the network that



(a) Expression 1 on original image.

(b) Expression 1 animatable 3D Reconstruction.

(c) Expression 2 on original image.

(d) Expression 2 animatable 3D Reconstruction.

(e) Expression 3 on original image.

(f) Expression 3 animatable 3D Reconstruction.

Figure 5: (Left to right) Single image input, next two images are the landmarks, corase reconstruction and the detail structure. In the detail structure we can see artifacts that make the mesh more realistic.

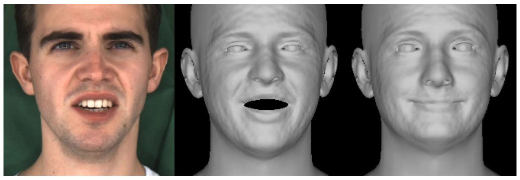


Figure 6: A single input image (Left). The resulting detail 3D reconstrcution (Middle). Animation result on changing expression parameter to a smile. (Right)

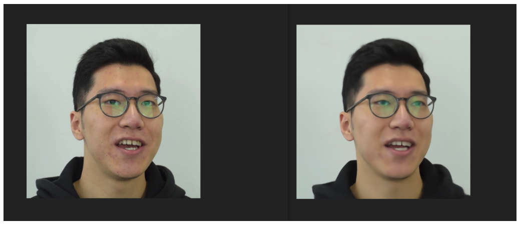


Figure 7: Ground Truth (Left). Neural Rendering Result (Right)

could disentable the static scene features and the dynamic facial expression controls. This methods would be able to take full advantage of the FLAME model and help give better animation results, potentially from various forms of input.

6 Acknowledgement

This work is part of an ongoing research project at the Intelligent Visual Interfaces Lab at Rutgers University under Professor Mubassir Kapadia, under the guidance of Jerry Chang.

References

[1] V. F. Abrevaya, A. Boukhayma, P. H. Torr, and E. Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2020.

[2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[3] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011.

[4] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Proceedings of fifth IEEE international conference on automatic face gesture recognition*, pages 202–207. IEEE, 2002.

[5] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer. Review of statistical shape spaces for 3d data with comparative analysis for human faces. *Computer Vision and Image Understanding*, 128:1–17, 2014.

[6] P. Chandran, S. Winberg, G. Zoss, J. Riviere, M. Gross, P. Gotardo, and D. Bradley. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[7] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019.

[8] Y.-L. Chen, H.-T. Wu, F. Shi, X. Tong, and J. Chai. Accurate and robust 3d facial capture using a single rgbd camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3615–3622, 2013.

[9] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019.

[10] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[11] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.

[12] P. Ghosh, P. S. Gupta, R. Uziel, A. Ranjan, M. J. Black, and T. Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020.

[13] Y. Guo, J. Cai, B. Jiang, J. Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1294–1307, 2018.

[14] L. Hu, C. Ma, L. Luo, and H. Li. Single-view hair modeling using a hairstyle database. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015.

[15] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017.

[16] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015.

[17] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1, 2013.

[18] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

[19] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[21] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[22] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[23] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020.

[25] T. Vetter and V. Blanz. Estimating coloured 3d face models from single images: An example based approach. In *European conference on computer vision*, pages 499–513. Springer, 1998.

[26] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[27] 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. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018.