Dynamic Neural Portraits

Michail Christos Doukas[1,2] Stylianos Ploumpis[2] Stefanos Zafeiriou[1,2]
[1]Imperial College London, UK [2]Huawei Technologies, London, UK
{michail.christos.doukas, stylianos.ploumpis, stefanos.zafeiriou1}@huawei.com

Abstract

We present Dynamic Neural Portraits, a novel approach to the problem of full-head reenactment. Our method generates photo-realistic video portraits by explicitly controlling head pose, facial expressions and eye gaze. Our proposed architecture is different from existing methods that rely on GAN-based image-to-image translation networks for transforming renderings of 3D faces into photo-realistic images. Instead, we build our system upon a 2D coordinate-based MLP with controllable dynamics. Our intuition to adopt a 2D-based representation, as opposed to recent 3D NeRF-like systems, stems from the fact that video portraits are captured by monocular stationary cameras, therefore, only a single viewpoint of the scene is available. Primarily, we condition our generative model on expression blendshapes, nonetheless, we show that our system can be successfully driven by audio features as well. Our experiments demonstrate that the proposed method is 270 times faster than recent NeRF-based reenactment methods, with our networks achieving speeds of 24 fps for resolutions up to 1024×1024*, while outperforming prior works in terms of visual quality.*

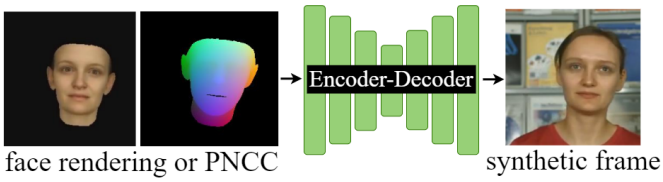
1. Introduction

Controllable video portrait synthesis is an interesting research topic that captures the attention of both Computer Graphics and Computer Vision communities. A video portrait is defined as a sequence of frames that depict a single individual performing diverse head movements and facial expressions. The person's entire head is contained within the frame's borders, along with a small part of the upper body, i.e. neck and torso, while the subject usually stands in front of a static background. Recent attempts to video portraits synthesis using neural networks have shown very promising results, based either on Generative Adversarial Networks (GANs) [20] or Neural Radiance Fields (NeRFs) [31]. The applications of such systems are numerous, ranging from video editing and movie dubbing to teleconference, virtual assistance, social media, VR and games.

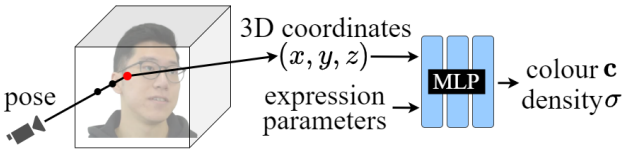
A number of learning-based solutions for generating video portraits rely on GAN-based image-to-image translation models, with an encoder-decoder architecture. For instance, Deep Video Portraits (DVP) [26] employ a network that learns a mapping from coloured renderings of 3D faces to realistic portraits. Similarly, Head2Head [27] translates images of Projected Normalized Coordinate Codes (PNCCs) [53] into photo-realistic frames, using a video-to-video framework. The renderings of 3D faces that serve as conditional input to generative neural networks depend on expression blendshapes that are obtained after fitting 3D Morphable Models (3DMMs) [5, 35, 7, 8, 29] to videos. The fitting step is followed by a physically-based rendering process, which creates the 2D renderings. On the contrary, we propose a multi-layer perceptron (MLP) that conditions synthesis directly on non-spatial data (e.g. expression parameters), and thus does not require renderings of 3D faces.

More recently, NerFACE [16] capitalised on the photo-realism achieved by NeRFs [31] and produced synthetic video portraits of higher quality, in larger resolutions, such as 512×512. Nonetheless, due to ray casting and volume sampling, image rendering with NerFACE requires several seconds per frame. Moreover, the authors make the assumption that the tracked head pose parameters coincide with the camera viewpoints of the scene, which causes significant inconsistencies in torso synthesis, as in practice the camera is static and therefore only a single viewpoint exists for the scene. AD-NeRF [22] proposes an audio-driven method for portrait synthesis and solves the camera problem by taking advantage of face segmentation maps [28]. They split the portrait into head, torso and background and devise two separate NeRF models: one for the head, which uses head poses as camera viewpoints, and another one for the torso, that treats head poses as simple inputs to the MLP, while considering camera viewpoints fixed. However, AD-NeRF is even slower during inference, as it evaluates two MLPs.

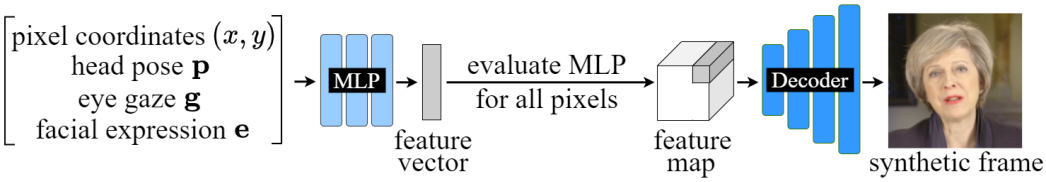
In this paper, we take a completely different approach and present *Dynamic Neural Portraits*, a fast and efficient framework for reenacting human faces. Our method draws inspiration from implicit neural representations (INR), as we leverage neural networks to parameterise video portraits. We follow practices from both conditionally inde-



(a) Image-based rendering of video portraits [26, 27]



(b) NeRF-based video portrait rendering [16, 22]



(c) Our paradigm for video portrait rendering

Figure 1. In contrast to traditional GAN-based image-to-image translation approaches to full-head reenactment, such as DVP [26] and Head2Head [27] (a), or recent NeRF-based video portrait rendering methods, namely NerFACE [16] and AD-NeRF [22] (b), we propose a novel paradigm for controllable video portrait synthesis, composed of an MLP and a CNN-based decoder (c).

pendent pixel synthesis (CIPS) [2] and neural rendering with convolutional networks. To be more specific, we perform video portrait synthesis using a 2D coordinate-based MLP with controllable dynamics. That is, we condition an MLP network on pixel coordinates, expression, pose and gaze parameters, without relying on renderings of 3D faces. Nonetheless, instead of directly predicting pixel colours, we adopt practices from 3D aware GANs [32, 21, 52, 9] and propose an MLP that produces feature vectors, which are computed across all 2D spatial locations and up-sampled with a CNN-based decoder. We optimise our "hybrid" MLP-CNN architecture jointly on the task of video portrait reconstruction. To the best of our knowledge, we are the first to condition a 2D coordinate-based MLP on expression, pose and gaze parameters for explicitly controlling video portraits, without relying on renderings of 3D faces, GANs, or NeRFs. Moreover, unlike previous methods that focus exclusively either on expression blendshapes or audio signals as driving data, our work demonstrates how to leverage both modalities using the same architecture. Our approach combines high-quality samples with unparalleled execution performance, 270 times faster than recent NeRF-based state-of-the-art reenactment methods [16, 22]. The contributions of this paper can be summarised as follows:

- We propose a new approach to full-head reenactment, with a generator that consists of a 2D coordinate-based MLP with controllable dynamics and a CNN decoder.
- We show that our architecture can be driven either by expression blendshapes or audio-based features.
- Our comprehensive experiments demonstrate that our method outperforms related state-of-the-art systems, both in terms of execution speed and image quality.

2. Related work

3D Face Modeling and Face Reenactment. Since their introduction, 3DMMs [5, 35, 7, 8, 29] have been widely used to represent human faces, as they are strong statistical models that enable explicit control over the shape and texture of 3D facial meshes. Various 3D face reconstruction methods depend on 3DMMs to recover 3D faces from visual data. Such methods are classified either as optimisation-based [46, 6], as they fit 3DMMs to visual data and estimate parameters in an analysis-by-synthesis fashion, or learning-based [18, 15] that rely on neural networks to reconstruct 3D faces. Apart from capturing coarse meshes, the later have shown very promising results in extracting even local fine details. Recovering facial shape and expression information from video data has been proven very useful for numerous face reenactment methods [17, 42, 43]. The work of Garrido *et al.* [17] is one of the first attempts to reenact faces while relying on 3D face modeling. A succeeding graphics-based method, namely Face2Face [43], performs real-time expression transfer from a driving sequence of frames to a target identity, by re-writing the interior facial region of the target video. A more recent approach, HeadOn [44], achieves full-head reenactment that includes pose and gaze transfer, based on RGB-D video data.

Learning-based Talking Head Synthesis. In contrary to the graphics-based systems described above, most of the latest face re-targeting approaches are learning-based [39, 26, 27]. Suwajanakorn *et al.* [39] are among the first to propose an audio-driven neural network for mapping acoustic signals to photo-realistic frames with accurate lip motions. Following up, Neural Voice Puppetry [40] employs a module for translating audio features into expression blendshapes, as an intermediate representation. Regarding video-

driven methods, Deep Video Portraits (DVP) [26] is one of the earliest GAN-based approaches to full-head reenactment. It relies on an image-to-image translation network that receives as inputs synthetic face renderings of a parametric 3D face model and generates images of the target subject. In a similar direction, Head2Head [27, 12] translates PNCCs [53] into photo-realistic frames, with the help of a video-based GAN for better temporal stability. Deferred Neural Rendering [41] takes a different step, by combining traditional graphics with learnable neural textures, embedded on the 3D facial mesh. Apart from the aforementioned person-specific approaches, there is a plethora of person-generic methods, which require only a few frames of the target identity [48, 37, 38, 4, 19, 50, 49, 23, 13].

Neural Representations for Scenes and Faces. With the introduction of NeRFs [31], much research has been focused on neural scene representations [30, 36], with various attempts to model human faces. More specifically, Nerfies [33] and HyperNeRF [34] have shown incredible results for reconstructing non-rigid scenes of moving heads. Despite their impressive generative ability, such systems are not able to control neither head poses nor facial movements. In a different direction, NerFACE [16] proposes to control a dynamic NeRF with the assistance of expression blendshapes. In a similar line of work, AD-NeRF [22] proposes an audio-driven NeRF-like model, based on acoustic features extracted with DeepSpeech [24, 1].

3. Method

In this section, we first describe our baseline approach to the problem of full-head reenactment (Sec. 3.1), that is a 2D coordinate-based MLP with controllable dynamics. Then, we couple the MLP with a CNN-based decoder to build our full model for video portrait synthesis (Sec. 3.2), namely *Dynamic Neural Portraits*. Finally, we show an extension of our system that supports audio-driven synthesis (Sec. 3.3).

3.1. 2D MLP with Controllable Dynamics

Let $\mathbf{I}$ be an image of a fixed resolution $H \times W$. We can represent $\mathbf{I}$ with a neural network by training a fully-connected MLP to reconstruct the image from its 2D coordinates [14]. For the synthesis of each pixel, we pass its coordinates $\mathbf{x} = (x, y)$ through the MLP network, which returns the colour of the pixel $\mathbf{c}$. We optimise the network by penalising the distance between the predicted and true colour. In order to compute the entire image, the MLP is evaluated at each position (x, y) of the coordinate grid.

The model described above is quite limited, as it learns only to reconstruct a single static image from its pixel coordinates. We extend the 2D coordinate-based MLP for handling temporally varying data, such as video portraits of human faces. Here, we focus on RGB videos captured by a monocular and stationary camera. A time-varying representation can be obtained by conditioning the MLP network on facial information that changes between frames. Let $\mathbf{I}_{1:T}$ be a sequence of frames and $\mathbf{p}_{1:T}$, $\mathbf{e}_{1:T}$ be the corresponding head pose and facial expression parameters respectively, which have been recovered with a face tracking system. An intuitive solution for modeling the video portrait with a neural network would be to estimate the pixel's RGB value from the i-th video frame with the MLP, using the pixel's coordinates $\mathbf{x}$ and tracked parameters $\mathbf{p}_i, \mathbf{e}_i$, as

$$\mathbf{c} = C(\mathbf{x}, \mathbf{p}_i, \mathbf{e}_i). \tag{1}$$

Here, $\mathbf{c}$ is the colour at 2D location $\mathbf{x} \in [0, 1]^2$, $\mathbf{p}_i \in \mathbb{R}^6$ is the head pose that is given by rotation (Euler angles) and translation parameters, and $\mathbf{e}_i \in \mathbb{R}^{n_{exp}}$ are the expression parameters, which correspond to non-rigid facial deformations. Please note that the pose parameters $\mathbf{p}_i$ describe the rigid motions of the face with respect to the camera and do not include information regarding torso movements. The position of the camera remains fixed throughout the frames.

3.2. Dynamic Neural Portraits

Even though the 2D coordinate-based MLP with controllable dynamics is a straightforward approach for modeling video portraits with a single MLP network, in practice we found that the photo-realism of generated samples and the rendering speed significantly degrade as we increase the resolution of videos. As shown in our experiments, we obtain superior results in terms of visual quality by combing the MLP with a convolutional decoder network. Following the paradigm of recent 3D aware GANs [32, 21, 52, 9], instead of predicting RGB colour values we propose an MLP that maps its input to a visual feature vector $\mathbf{f} \in \mathbb{R}^{n_f}$. Given the 2D spatial location $\mathbf{x}$ along with pose $\mathbf{p}_i$, expression $\mathbf{e}_i$ and gaze information $\mathbf{g}_i$, our MLP now predicts a feature vector

$$\mathbf{f} = F(\gamma(\mathbf{x}), \gamma(\mathbf{p}_i), \gamma(\mathbf{g}_i), \mathbf{e}_i, \mathbf{v}_i). \tag{2}$$

We observed that we acquire more accurate eye movements by adding eye gaze angles $\mathbf{g}_i \in \mathbb{R}^2$ as an extra input to our MLP network. Moreover, we noticed that introducing per-image learnable latent variables $\mathbf{v}_i$ as inputs, a technique which has been previously adopted in NeRF methods [30, 33, 34, 16], improves the stability of our network, since it enables the MLP to learn variations among frames that are not modeled by the pose and expression parameters (e.g. torso movements, illumination changes, small background motions). Furthermore, following established practices from NeRF-based systems and CIPS [2], we adopt positional encoding for the MLP inputs, and more specifically on the position $\mathbf{x}$, pose $\mathbf{p}_i$ and gaze $\mathbf{g}_i$ vectors, which are low-dimensional. The standard encoding function

$$\begin{aligned}\gamma(x) = [&x, \sin(2\pi x), \cos(2\pi x), \dots, \\ &\sin(2^{N-1}\pi x), \cos(2^{N-1}\pi x)]^\top\end{aligned} \tag{3}$$

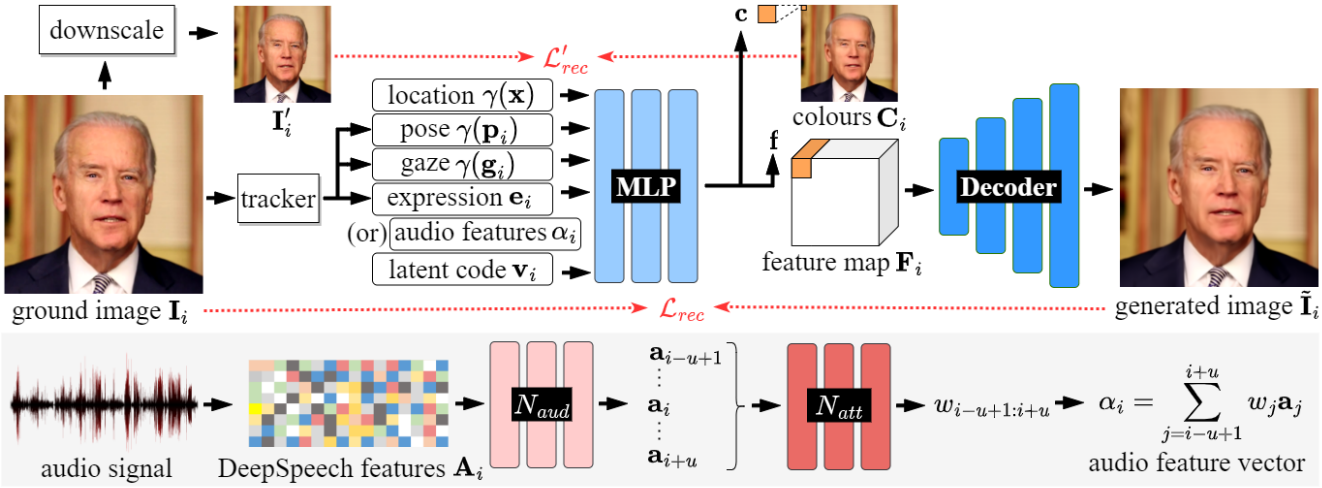


Figure 2. An overview of *Dynamic Neural Portraits* framework during training. As opposed to previous full-head reenactment methods that rely on image-to-image translation networks, our model is made up of an MLP encoder and a CNN decoder. Instead of facial expression parameters, we can drive synthesis using audio features, recovered from acoustic signals.

proposed in [31] is applied to all values of $\mathbf{x}$, $\mathbf{p}_i$ and $\mathbf{g}_i$, mapping each number $x \in \mathbb{R}$ to a higher dimension embedding $\gamma(x) \in \mathbb{R}^{2 \cdot N+1}$. Replacing the original inputs with their embeddings allows our model to achieve high frequency details in the generated images.

The MLP network described above learns to estimate a feature vector $\mathbf{f}$ for each spatial location of the image plane independently, while considering head pose, facial expression and eye gaze information. In order to render frame i, we first evaluate the MLP network at each spatial position $\mathbf{x} \in \mathbf{X}$ of the coordinate grid that corresponds to resolution $H_f \times W_f$, while keeping all other inputs fixed. After that, we accumulate the resulting features in a visual feature map $\mathbf{F}_i \in \mathbb{R}^{H_f \times W_f \times n_f}$. Then, we employ a decoding network D, which receives the feature map and performs up-sampling, in order to synthesise the output frame $\tilde{\mathbf{I}}_i = D(\mathbf{F}_i)$ at the target resolution $H \times W$. An overview of our proposed framework is shown in Fig. 2. More details on the architecture of networks are available in the supplementary material.

Objective Function and Optimisation. We train our proposed model on the task of reconstruction. Given the generated image $\tilde{\mathbf{I}}_i$ and the corresponding ground truth frame $\mathbf{I}_i$, we define the reconstruction loss as the L2-distance between the predicted and true image. We experimented with L1, perceptual [47] and adversarial [20] losses, but all of them produced substantially inferior results. However, we found that we can improve our method's performance by adding an extra output layer to the MLP, which predicts the colour $\mathbf{c} \in \mathbb{R}^3$ side-by-side with visual features $\mathbf{f}$, and minimise the distance of the predicted colour from the true one. To that end, we accumulate the colour outputs for all 2D spatial locations in $\mathbf{C}_i \in \mathbb{R}^{H_f \times W_f \times 3}$, and penalise the L2-distance from the ground truth image $\mathbf{I}'_i$ down-scaled to resolution $H_f \times W_f$, in order to match the resolution of $\mathbf{C}_i$. The overall loss term for frame i is given as

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}'_{rec} = ||\tilde{\mathbf{I}}_i - \mathbf{I}_i||_2^2 + ||\mathbf{C}_i - \mathbf{I}'_i||_2^2 \quad (4)$$

We optimise the MLP network and CNN-based decoder jointly, in an end-to-end fashion.

3.3. Audio-driven Portrait Synthesis

Our choice to inject the driving signals (i.e. pose, expression and gaze parameters) into our system through an MLP network enables to easily adapt our method to other driving modalities, such as acoustic signals. Unlike previous works [22, 16, 40, 26], which focus exclusively either on expression blendshapes or audio data as driving signals, we show that the same architecture can be used effectively for both modalities. More specifically, we can replace the expression blendshapes with an audio feature vector $\boldsymbol{\alpha}_i$ as input to the MLP network, in case the audio stream is available. We extract high-level features from acoustic signals with the widely-adopted DeepSpeech model [24, 1]. As a first step, we assign a 29-dimensional vector estimated by DeepSpeech to every video frame. Then we create a window of vectors with size $w = 16$ around each frame, taken from its neighboring (past and future) time steps. In this way each frame i is coupled with a DeepSpeech feature $\mathbf{A}_i \in \mathbb{R}^{16 \times 27}$. We follow a similar approach with AD-NeRF [22] and utilise a 1D-convolutional network N_{aud} that learns to compute per-frame latent codes $\mathbf{a}_i \in \mathbb{R}^{n_a}$ from $\mathbf{A}_i$. Next, we employ a self-attention network N_{att}, as proposed in [40] and [22], which operates as a temporal filter on subsequent audio codes $\mathbf{a}_{i-u+1:i+u}$ and mixes them up with the assistance of predicted attention weights $w_{i-u+1:i+u}$, to form the final audio feature vector $\boldsymbol{\alpha}_i = \sum_{j=i-u+1}^{i+u} w_j \mathbf{a}_j$. We set $u = 4$, which results in a window of $2u = 8$ time steps.

4. Experiments

Dataset. Our networks are optimised on monocular RGB videos, of various resolutions: 256^2, 512^2 and 1024^2. We train a new model for each video portrait (different individual). As a pre-processing step, we crop frames around the target face and compute pose, expression and gaze parameters. We experiment with videos from 5K to 20K frames. In the supplementary material we provide more details of the adopted video database and face tracking systems.

4.1. Comparison with State-of-the-Art

Reconstruction. First, we carry out a comparison with Deep Video Portraits (DVP) [26] and NerFACE [16], the best-performing person-specific reenactment methods, as well as First Order Motion Model (FOMM) [38], a representative of person-generic models. We evaluate the generative performance of methods on the task of reconstruction, also know as self-reenactment. We assess the fidelity of reconstruction quantitatively, with the assistance of L1-distance between generated and ground truth test frames, as well as with Learned Perceptual Image Patch Similarity (LPIPS) [51], which tests the perceptual similarity between images. Moreover, we determine photo-realism with Fréchet Inception Distance (FID) [25] and Fréchet Video Distance (FVD) [45] metrics, which appear to correlate well with human perception. The results displayed in Table 1 indicate that our method outperforms all baselines, for three different video portraits. Please note that the reported numbers refer to the average scores, computed across all test frames for each video portrait. An exception to that are FID [25] and FVD [45] metrics, as they are computed on

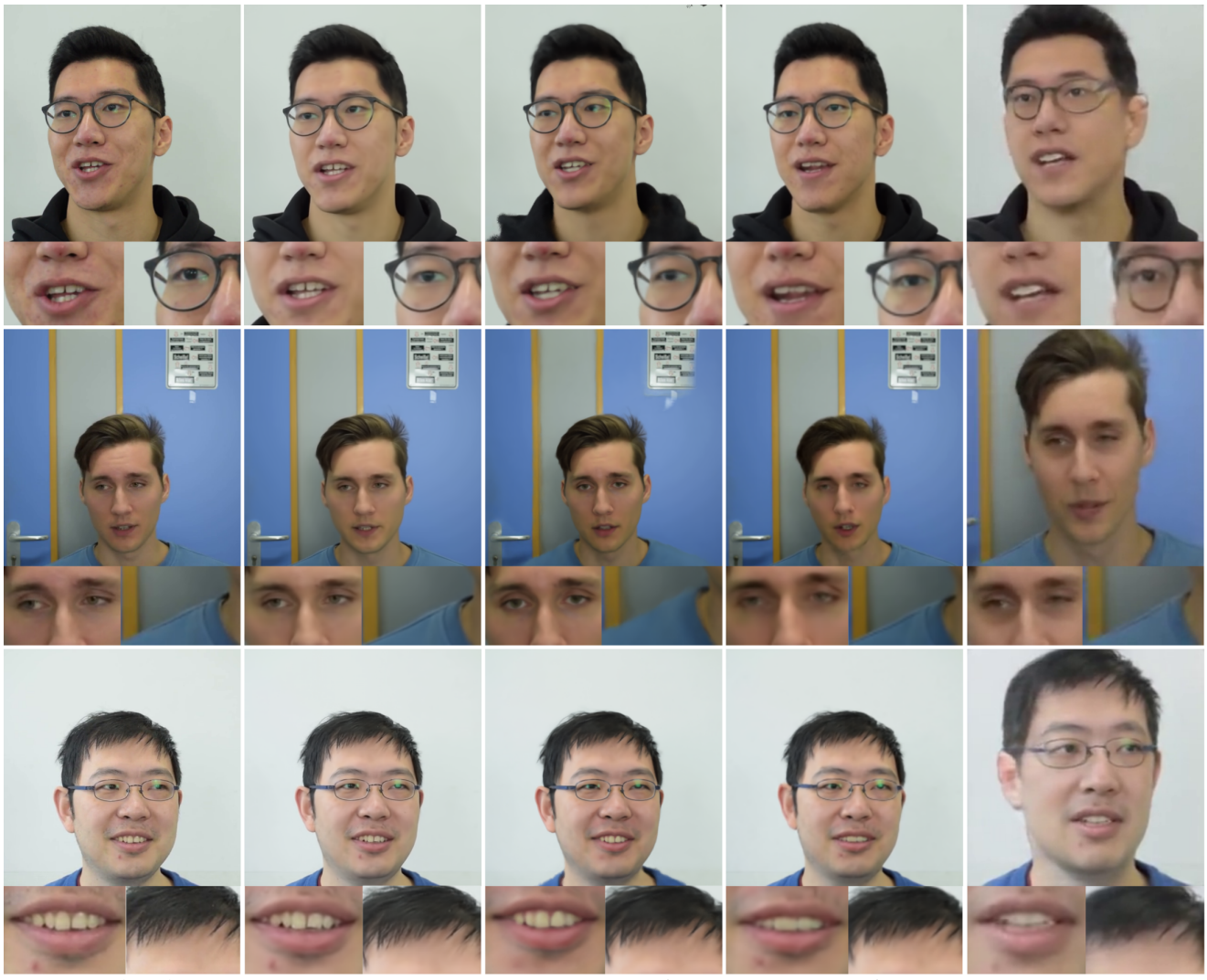


Figure 3. Visual comparison with baselines on the task of reconstruction. Our method consistently generates more realistic samples with finer details than its counterparts. For the evaluation of FOMM [38], images had to be further cropped. Please zoom in for details.

Method	Portrait	L1 (↓)	LPIPS (↓)	FID (↓)	FVD (↓)
FOMM	ID. 1	8.79	0.095	25.48	331.0
	ID. 2	7.60	0.084	37.24	338.3
	ID. 3	9.48	0.130	24.72	254.10
DVP	ID. 1	6.95	0.152	49.35	195.4
	ID. 2	9.08	0.079	37.58	464.3
	ID. 3	8.01	0.123	51.30	196.7
NerFACE	ID. 1	**6.19**	0.136	74.22	278.3
	ID. 2	10.98	0.143	74.02	357.7
	ID. 3	6.28	0.067	34.64	81.17
Ours	ID. 1	6.45	**0.094**	**23.76**	**169.8**
	ID. 2	**5.21**	**0.071**	**24.60**	**222.1**
	ID. 3	**5.15**	**0.051**	**23.94**	**78.06**

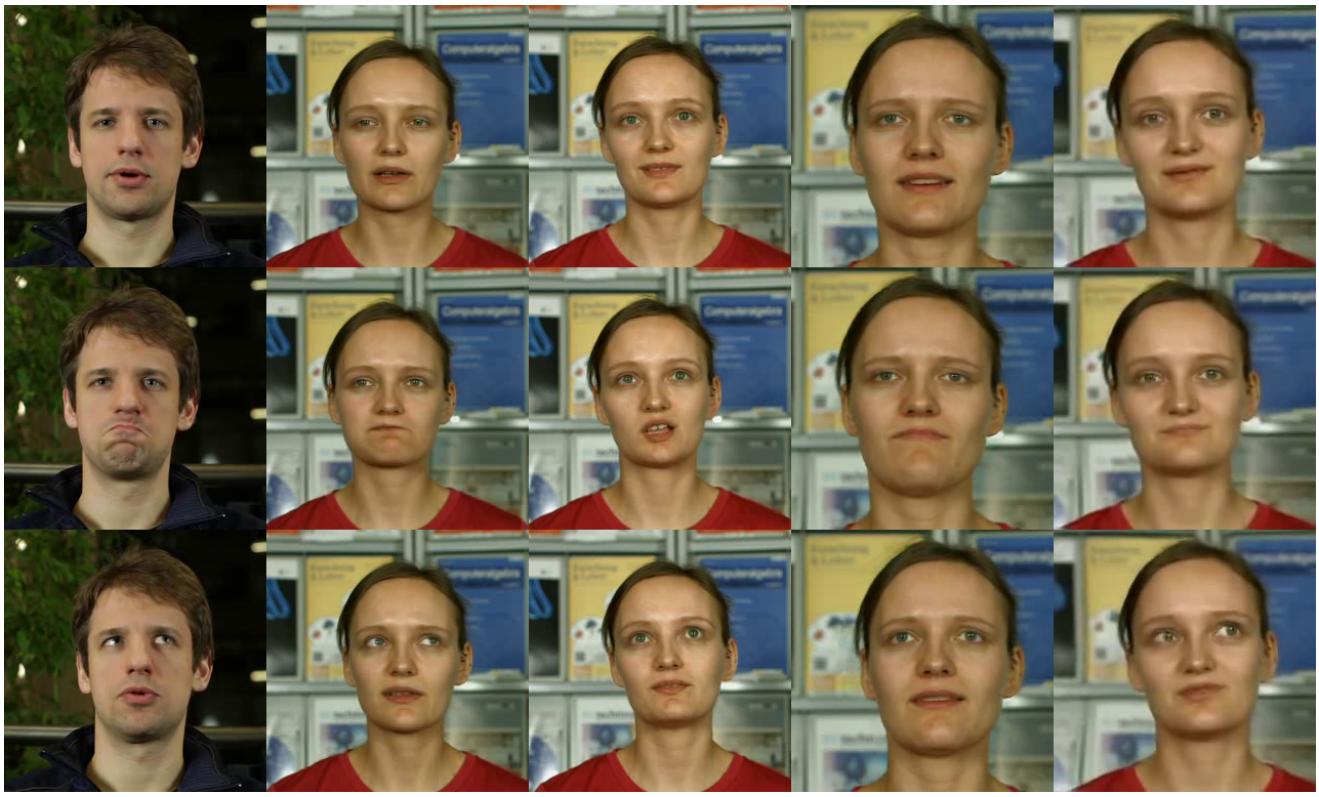
Table 1. Numerical comparison with FOMM [38], DVP [26] and NerFACE [16] for three different portraits on the task of reconstruction (self-reenactment).

pairs of videos. All scores suggest that our approach produces superior samples in terms of visual quality and photorealism. This observation is confirmed visually in Fig. 3. As can be seen, our method generates more crispy images with finer details and more consistent eye gaze. For a better visualisation of our results, we urge the reader to refer to our supplementary video.

Method	CSIM (↑)	Expression Dist. (↓)	Pose Dist. (↓)	Gaze Dist. (↓)
FOMM [38]	0.840	13.28	6.31°	9.42°
HeadGAN [13]	0.755	14.17	3.26°	6.35°
DVP [26]	0.861	11.95	4.93°	8.37°
Ours	**0.885**	**10.69**	**2.57°**	**4.74°**

Table 2. Numerical comparison with FOMM [38], HeadGAN [13] and DVP [26] on the task of cross-identity reenactment.

Reenactment. We further validate our model's performance on the task of cross-identity motion transfer (reenactment) in a quantitative way. This task involves passing on the head pose, facial expression and eye gaze from a driving actor to another target person, while preserving the identity of the later, as the two subjects are now different. We compare our approach with DVP [26], FOMM [38] and HeadGAN [13]. To that end, we use the source and target video portraits provided by the authors of DVP. For the numerical comparison, we measure the target's identity preservation with Cosine Similarity (CSIM), based on identity embeddings extracted with the assistance of ArcFace [11]. Moreover, we use DECA [15] to regress pose and expression parameters, both from the driving and gen-



driver ours DVP [26] HeadGAN [13] FOMM [38]

Figure 4. Visual comparison with baselines on reenactment. Our approach transfers pose, expression and gaze more reliably than both person-specific DVP [26] and person-generic methods HeadGAN [13] and FOMM [38]. Please note that for the correct evaluation of HeadGAN and FOMM, images had to be cropped closer to the face.

Method	Portrait	L1 (↓)	LPIPS (↓)	FID (↓)	FVD (↓)
AD-NeRF	Obama	4.11	0.083	16.67	225.7
	May	**5.13**	0.143	47.31	272.2
Ours	Obama	**4.04**	**0.054**	**9.12**	**110.3**
	May	5.72	**0.087**	**26.72**	**96.1**

Table 3. Numerical comparison with AD-NeRF [22] on audio-driven video reconstruction.

ground truth ours AD-NeRF [22]

Figure 5. Visual comparison with AD-NeRF [22] on audio-driven reconstruction. We generate more accurate lips motions and higher quality details compared to AD-NeRF. Please zoom in for details.

erated frames. Then, we compute the L1-distance between expression parameters, as well as the head rotation distance in degrees. Finally, we use the gaze estimator from [10] to regress gaze vectors, and calculate their angular distance. In Table 2, we show that our method achieves better scores than all three baselines, across all metrics related to successful reenactment. In Fig. 4, we showcase examples of frames where our method transfers pose, expression and gaze more accurately than DVP [26], HeadGAN [13] and FOMM [38].

Audio-driven reconstruction. Except for conditioning our MLP network on expression blendshapes, we demonstrate the generative performance of our system when driven by acoustic signals. For that, we conducted a side-by-side comparison with AD-NeRF [22], the state-of-the-art model on this task. As shown in Table 3, our quantitative analysis reveals that our approach performs better than AD-NeRF, both in terms of reconstruction and image quality. Our findings can be observed also visually in Fig. 5.

Execution Speed. Owing to its lightweight architecture, our method is able to render frames in a resolution up to 1024×1024, in 24 fps. For lower resolutions (e.g. 256×256 and 512×512) our pipeline operates in speeds faster than real time. In comparison to recent NeRF-based state-of-the-art methods [16, 22], our system achieves a significant speed-up, generating images nearly 270 times faster, which makes it much more efficient for real-world applications. In Table 4, we report the execution times of different methods, measured on NVIDIA's Tesla V100 PCIe 32 GB. For DVP* [26] we use the numbers recorded by its authors.

Method	256×256 time (fps)	512×512 time (fps)	1024×1024 time (fps)
AD-NeRF [22]	-	9630 (0.10)	-
NerFACE [16]	-	8465 (0.12)	-
DVP* [26]	65 (15.4)	196 (5.1)	-
HeadGAN [13]	41 (24.5)	-	-
FOMM [38]	21 (47.2)	-	-
Ours	**11** (**90.9**)	**31** (**32.3**)	**42** (**24.2**)

Table 4. Comparison of the execution time between our generative model and related methods. Time is reported in milliseconds (msec). Please note that all reported numbers refer to the forward pass time of models during inference, without considering data pre-processing.

4.2. Ablation study

Next, we evaluate our design choices and validate the significance of different components that make up our model. We conduct quantitative and qualitative experiments for six variations of our system. To that end, we test all variations on the task of reconstruction, for two separate portraits (Biden and Obama). We start with a 2D coordinate-based MLP with controllable dynamics, described in Section 3.1, which is used as the baseline model (A) of our ablation study. In order to demonstrate the advantages of our proposed architecture, we further experiment with a CNN-based decoding network only, without the MLP network (B). Then, we couple the MLP with the decoding network, which leads to variation (C). We form the next variations by first adding the learnable latent variables input (D), then including the reconstruction loss term $\mathcal{L}'_{rec}$ (E) and finally the gaze input (F), ending up with our full model as presented in Section 3.2. Our numerical analysis presented in Table 5 reveals the importance of each component, especially the huge impact of the MLP network when coupled with a CNN-based decoding network. As can be observed, the latent variables increase FVD scores as they help to stabilise movements between frames. Furthermore, according to LPIPS, the $\mathcal{L}'_{rec}$ loss improves the perceptual similarity with ground truth data. Finally, the gaze input corrects eye motions, something that becomes more apparent when inspecting our supplementary video. In Fig. 6, we illustrate some examples of the visual differences between samples generated by variations (A), (B) and our full model (F).

5. Discussion

3D multi-view consistency. Theoretically, our method is not 3D aware. Nonetheless, for the purposes of portrait reenactment this is not essential for achieving consistent results in various head poses. Given that training videos are captured by stationary cameras, they provide access to a single view of the scene. This makes 3D NeRF-based methods an over-parametrisation for such data, which end

ground truth | 2D MLP contr. dyn. (A) | CNN-based decoder (B) | our full model (F)

Figure 6. Qualitative results of our ablation study. Our full model (F) improves the quality of frames synthesised by our baseline (A) (2D coordinate-based MLP with controllable dynamics) or the unaided CNN-based decoder (B) by a noteworthy margin.

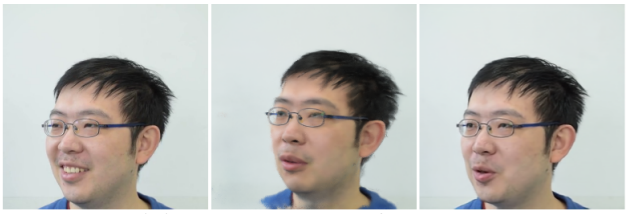
	Variation	Portrait	L1 (↓)	LPIPS (↓)	FID (↓)	FVD (↓)	Gaze Dist. (↓)
(A):	2D coordinate-based MLP with controllable dynamics (baseline, Section 3.1)	Biden	5.33	0.072	10.69	197.2	-
		Obama	3.74	0.077	14.33	132.9	
(B):	CNN-based decoder	Biden	6.67	0.089	11.51	263.8	-
		Obama	4.27	0.068	12.78	127.4	
(C):	MLP + decoder	Biden	5.55	0.061	8.76	114.2	-
		Obama	3.58	0.060	9.37	103.4	
(D):	(C) + latent codes	Biden	5.68	0.062	**8.38**	**92.05**	-
		Obama	3.10	0.047	9.37	83.1	
(E):	(D) + $\mathcal{L}'_{rec}$	Biden	5.14	0.057	9.80	100.9	2.69
		Obama	**2.80**	0.041	9.97	82.1	2.31
(F):	(E) + gaze input (full model, Section 3.2)	Biden	**5.03**	**0.054**	9.42	107.2	**1.86**
		Obama	2.86	**0.041**	**9.33**	**79.3**	**2.18**

Table 5. Quantitative results of our ablation study on two separate video portraits (Biden and Obama), of resolution 512×512.

up violating 3D consistency for upper torso and producing severe shoulder trembling [16]. We argue that 3D modeling would require a video portrait captured by a moving camera from multiple viewpoints, as demonstrated most recently by RingNeRF [3]. For videos captured by stationary cameras, our 2D model is able to generate frames with consistent appearance in diverse head poses and surpass 3D NeRF-based methods [16, 22] in visual quality and inference speed.

Large changes in head pose. The reenactment results we presented are mainly frontal. This is attributed to the training data, which consist mostly of frontal poses. Actually, both 2D and 3D-based methods are limited by the head pose variation that exists in the training video. We observed that all systems struggle to synthesise poses out of the training data span. In Fig. 7, we illustrate the training frame with the most extreme head pose, in terms of yaw angle. Next to it, we show the most extreme pose synthesised by NerFACE [16] and our method, without a substantial drop in quality. Our approach achieves higher photo-realism.

Ethical considerations. Generative models that offer explicit control on movements of faces could be potentially used for unethical purposes. For example, they could be used to synthesise videos of politicians acting in a provocative way. We would like to clarify that the intention of this paper is to make advancements realistic full-head reenactment for benevolent applications. We do not condone using our work to produce fake news or deceive the public.



max yaw (training) | NerFACE [16] | ours

Figure 7. Evaluation of larger changes in head pose.

6. Conclusion

We described *Dynamic Neural Portraits*, a novel method for controllable video portrait synthesis. In contrast to previous attempts, our approach does not require renderings of 3D faces to drive synthesis and does not rely on GANs or NeRFs. Our experiments demonstrate the superiority of our model in terms of visual quality and run-time performance, compared to state-of-the-art video or audio-driven systems.

Acknowledgement. The work of Stefanos Zafeiriou was partially funded by the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1).

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[2] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2021.

[3] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022.

[4] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[6] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57, 2017.

[7] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018.

[8] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[10] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[12] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):31–43, 2021.

[13] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021.

[14] Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021.

[15] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021.

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, June 2021.

[17] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4217–4224, 2014.

[18] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[21] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[22] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[23] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10893–10900, 2020.

[24] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[27] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 16–23. IEEE, 2020.

[28] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5557, 2020.

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[30] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel 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.

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[32] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[33] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[34] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[35] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas 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.

[36] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[38] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

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

[41] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG 2019*, 2019.

[42] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015.

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

[44] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[45] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[46] Thomas Vetter and Volker Blanz. Estimating coloured 3d face models from single images: An example based approach. In *In Proceedings, European Conference on Computer Vision*, pages 499–513. Springer, 1998.

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[48] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.

[49] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.

[50] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

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

[52] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021.

[53] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.