Head2Head: Video-based Neural Head Synthesis

Mohammad Rami Koujan*[1,3], Michail Christos Doukas*[2,3], Anastasios Roussos[4,3], Stefanos Zafeiriou[2,3]
[1]College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
[2]Department of Computing, Imperial College London, UK
[4]Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), Greece
[3]FaceSoft.io, London, UK

***Abstract*— In this paper, we propose a novel machine learning architecture for facial reenactment. In particular, contrary to the model-based approaches or recent frame-based methods that use Deep Convolutional Neural Networks (DCNNs) to generate individual frames, we propose a novel method that (a) exploits the special structure of facial motion (paying particular attention to mouth motion) and (b) enforces temporal consistency. We demonstrate that the proposed method can transfer facial expressions, pose and gaze of a source actor to a target video in a photo-realistic fashion more accurately than state-of-the-art methods.**

I. INTRODUCTION

Facial reenactment aims at transferring the expression of a source actor to a target face. It is a challenging problem, in the cutting edge of research and technology, with many applications in video editing, movie dubbing, telepresence and virtual reality. Recent works have produced impressive results and have attracted the attention of the research community, the industry, as well as the general public.

The majority of facial reenactment methods transfer the expressions of the source actor by modifying the deformations within solely the inner facial region of the target actor, without altering the head movements of the target video [1], [2], [3], [4], [5], [6]. Thus, even in cases where this expression transfer is performed well, the overall reenactment result might seem uncanny and non-plausible, as the head motion of the target may not match with the transferred expressions. Various recent methods attempt to transfer the entire head motion [7], [8], [9], [10]. Despite their promising results, these methods have limitations in terms of how realistic the reenactment result looks. *X2face* [7] is a warping-based approach, which is sometimes unsuccessful on preserving the identity of target person and the results often look unrealistic. *Deep Video Portraits* [8] performs image-based head reenactment, by conditioning an image-to-image translation neural network on 3D facial information. As frames are synthesised independently from each other, there are cases where videos exhibit temporal incoherence between the generated frames, especially in the mouth region.

Our proposed approach overcomes all the aforementioned limitations. We fully transfer the pose, facial expression and eye gaze movement from a source to target video, while preserving the identity of the target and maintaining a consistent motion of the upper body part. Given that people easily detect mistakes in the appearance of a human face (uncanny valley effect), we give special attention in the details of the mouth and teeth. Our system generates photo-realistic and temporally consistent videos of faces.

* Equal contribution.

II. RELATED WORK

A. 3D Face Reconstruction

Human faces have attracted substantial attention in the computer vision field due to their centrality in many applications. A growing line of research revolves around reconstructing the 3D geometry of the face and employing this information in further ad-hoc steps, as we do in this work. Some attempts on 3D face reconstruction [11], [12], [13], known as Shape From Shading, work on approximating the image formation process and postulate simplified assumptions about the lighting and illumination models leading to the formation of the image, while others, known as Structure from Motion (SfM), capitalise on the geometric constraints in multiple images of the same object to approach the reconstruction task [14], [15], [16]. 3D Morphable Models (3DMMs) have been researched and used substantially in the literature since the pioneering work of Blanz and Vetter [17], with many extensions [18], [19], [20]. 3DMMs are generative parametric models for the 3D representation of human faces. They are built from a set of 3D facial scans, coupled to each other with anatomical correspondences, and can represent any unseen faces as a linear combination of the training set.

B. Facial Reenactment and Full Head Reenactment Methods

A plethora of works is devoted to the problem of facial reenactment [2], [4], [5], [9]. Expressions are transferred either by 2D warping techniques, based on dense motion fields [1], [21], [9], or by fitting and manipulating parametric 3D face models [6], [4], [5]. A well-known facial reenactment system is *Face2Face* [5], which relies on monocular 3D face reconstruction on both the source and target videos and operates in real time. *Neural Textures* [22] is another approach to facial reenactment, achieving high quality results. In this method, texture features are learned from the target video and translated to RGB images with a neural renderer.

Although many of these facial reenactment systems produce highly realistic results, they do not provide complete control over the generated video. There is a limited number of works in the direction of full head motion transfer [7], [8], [9], [10]. Averbuch-Elor et al. [9] use a target portrait

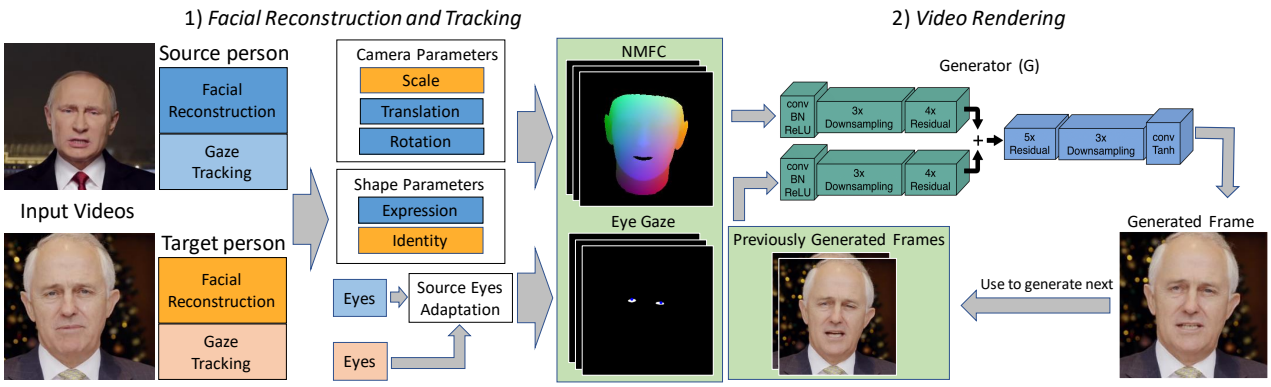

Fig. 1: The pipeline of our Head2Head approach marked by two subsequent stages: 1) facial reconstruction and tracking, 2) video rendering. The **NMFC** and eye gaze sequences computed in the first stage are used to drive the video synthesis. During training, both the source and target frames come from the same (target) video, since we perform self-reenactment.

photo instead of a video and cannot deal well with large variations in the head pose. The target gaze is not controlled and the mouth appearance is copied from the source actor. Zakharov *et al.* [10] propose a network that synthesises frames conditioned on landmarks, using a few-shot adversarial learning approach. However, their method does not adapt the landmarks extracted from the source actor to the head geometry of the target, leading to identity mismatches of the generated video.

C. Image and Video Synthesis

Generative Adversarial Networks (GANs) [23] have been widely used for photo-realistic image and video synthesis. Instead of conditioning synthesis on noisy distributions [23], [24], other data types are frequently used as input to the generator, such as class labels [25] or images [26], [27], [28]. There is extended research on image-to-image translation [29], [30], [26], [28]. Other methods [31], [32], [33], [34] utilise the GAN framework on video synthesis tasks, such as *vid2vid* [35]. In this work, we exploit a GAN-based approach for rendering photo-realistic video frames, providing temporal stability and paramount image quality.

III. Methodology

Our proposed approach addresses the highly-challenging problem of full-head reenactment. More specifically, we fully transfer head pose, facial expression and eye gaze movement from a source to target actor's video. Our **head2head** pipeline, illustrated in Fig. 1, consists of two main and consecutive stages: 1) 3D facial reconstruction and tracking, and 2) learning-based video rendering, using a neural network.

A. Facial Reconstruction and Tracking

To effectively disentangle the human head characteristics in a transferable and photo-realistic way between different facial videos, we benefit from the prior knowledge about the targeted object. This knowledge has been effectively addressed and captured with 3DMMs [17]. Thus, we harness the power of 3DMMs for 3D reconstructing and tracking the faces appearing in the input sequences. Given a sequence of T frames $\mathcal{F}_{1:T} = \{f_t \mid t = 1, \ldots, T\}$, the 3D reconstruction and tracking stage produces two sets of parameters: 1) shape parameters $\mathcal{S} = \{\boldsymbol{s}_t \mid \boldsymbol{s}_t \in \mathbb{R}^{n_i+n_e}, t = 1, ..., T\}$, and 2) camera parameters $\mathcal{P} = \{\boldsymbol{p}_t \mid \boldsymbol{p}_t \in \mathbb{R}^7, t = 1, ..., T\}$, embodying rotation, translation and orthographic scale.

Shape representation. With 3D Morphable Models, a 3D facial shape $\mathbf{x}_t = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^T \in \mathbb{R}^{3N}$ can be represented mathematically as:

$$\mathbf{x}(\boldsymbol{s}_t^i, \boldsymbol{s}_t^e) = \bar{\mathbf{x}} + \mathbf{U}_{id}\boldsymbol{s}_t^i + \mathbf{U}_{exp}\boldsymbol{s}_t^e \tag{1}$$

where $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is the overall mean shape vector of the morphable model, given by $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{id} + \bar{\mathbf{x}}_{exp}$, with $\bar{\mathbf{x}}_{id}$ and $\bar{\mathbf{x}}_{exp}$ depicting the mean identity and expression of the model, respectively. $\mathbf{U}_{id} \in \mathbb{R}^{3N\times n_i}$ is the identity orthonormal basis with n_i principal components ($n_i \ll 3N$), $\mathbf{U}_{exp} \in \mathbb{R}^{3N\times n_e}$ is the expression orthonormal basis with the n_e principal components ($n_e \ll 3N$) and $\boldsymbol{s}_t^i \in \mathbb{R}^{n_i}$, $\boldsymbol{s}_t^e \in \mathbb{R}^{n_e}$ are the identity and expression parameters of the morphable model. We denote the joint identity and expression parameters of a frame $\boldsymbol{t}$ by $\boldsymbol{s}_t = [\boldsymbol{s}_t^{iT}, \boldsymbol{s}_t^{eT}]^T$. In the adopted model (1), the 3D facial shape $\mathbf{x}$ is a function of both identity and expression coefficients ($\mathbf{x}(\boldsymbol{s}_t^i, \boldsymbol{s}_t^e)$), with expression variations being effectively represented as offsets from a given identity shape.

Video Fitting. To estimate the shape and pose parameters of the source and target sequences in a quick and robust way, we follow a novel batch landmark-based approach that takes into account the information from all video frames simultaneously and exploits the rich dynamic information usually contained in facial videos. Similar to [36], we exploit the fact that the current state-of-the-art in facial landmarking can achieve highly-reliable landmark localisation and therefore fuse the landmarks information with high-quality 3D face models, as the one described in (1), to achieve robust and accurate 3D face reconstruction results. We assume that the camera performs scaled orthographic projection (SOP) and that the identity parameters $\boldsymbol{s}_t^i$ are fixed (but unknown) over all the frames, letting however the expression parameters $\boldsymbol{s}_t^e$ as well as the camera parameters (scale and 3D pose) to vary

from one frame to another. In brief, for a given sequence of frames, we minimise a cost function that consists of three terms, see (2) : **a)** a sum of squared 2D landmark reprojection errors over all frames (E_l), **b)** a shape priors term (E_{pr}) that imposes a quadratic prior over the identity and per-frame expression parameters and **c)** a temporal smoothness term (E_{sm}) that enforces smoothness of the expression parameters in time, by using a quadratic penalty of the second temporal derivatives of the expression vector.

$$E(\mathcal{S},\mathcal{P}) = w_l E_l(\mathcal{S},\mathcal{P}) + w_{pr} E_{pr}(\mathcal{S}) + w_{sm} E_{sm}(\mathcal{S}^e) \quad (2)$$

In addition, to deal with outliers (e.g. frames with strong occlusions that cause gross errors in the landmarks), we also impose box constraints on the identity and per-frame expression parameters. Assuming that the camera parameters ($\mathcal{P}$) in (2) have been estimated in an initialisation stage, the minimisation of the cost function results in a large-scale least squares problem with box constraints, which we solve efficiently by using the reflective Newton method of [37]. More details about the initialisation stage of our video fitting step are available in the supplementary material.

3DMM details. For our set of experiments, the identity part of the 3DMM, $\{\bar{\mathbf{s}}_{id}, \mathbf{U_{id}}\}$, originates from the Large Scale Morphable Model (LSFM) [38], [39] built from approximately 10,000 scans of different people, the largest 3DMM ever constructed, with varied demographic information. In addition, the expression part of the model, $\{\bar{\mathbf{x}}_{exp}, \mathbf{U}_{exp}\}$ originates from the work of Zafeiriou et al. [40], who built it using the blendshapes model of Facewarehouse [41] and adopting Nonrigid ICP [42] to register the blendshapes model with the LSFM model.

Gaze tracking. Since the human gaze direction is not captured generally by 3D face scanners, 3DMMs of facial shapes do not represent this characteristic. We, therefore, employ a state-of-the-art gaze tracking technique [43] for tracking the eyes in the source and target sequences.

Many state-of-the-art approaches [8], [5], etc. rely on the analysis-by-synthesis framework for fitting 3DMMs to images, which requires estimating more parameters (e.g. illumination and reflectance) and solving a highly ill-posed problem. On the other hand, our facial reconstruction and tracking stage is a sparse-landmarks-based fast approach, which requires only 68 facial landmarks extracted by [44], as well as the frame sequence. Thanks to our novel video rendering framework, the facial representation extracted by our face tracker encapsulates adequate information for synthesising photo-realistic and temporally smooth videos, removing the need for more elaborate and slower 3D facial reconstruction and tracking techniques.

B. Conditioning Images Generation

Given the estimated shape and camera parameters from both the source and target frames at time t, we replace the identity coefficients and the scale parameter of the source actor with the ones from the target, creating the "hybrid" shape and camera parameters $\boldsymbol{s}_t$, $\boldsymbol{p}_t$, as shown in Fig. 1. Then, instead of feeding these per-frame face parameters directly to the video rendering network, we create a more meaningful representation in the image space. We rasterize the 3D facial shape $\mathbf{x}(\boldsymbol{s}_t)$, producing a visibility mask ($\mathbf{M} \in \mathbb{R}^{W\times H}$) in the image space. Each pixel of $\mathbf{M}$ stores the ID of the corresponding visible triangle on the 3D face from this pixel. Then, we encode the normalised x-y-z coordinates of the centre of this triangle in another image, termed as Normalised Mean Face Coordinates ($\mathbf{NMFC} \in \mathbb{R}^{W\times H\times 3}$) image, and utilise it as conditional input of the video rendering stage, see equation (3) below.

$$\mathbf{NMFC}_t = \mathcal{E}(\mathcal{R}(\mathbf{x}(\boldsymbol{s}_t), \boldsymbol{p}_t), \bar{\boldsymbol{x}}), \quad (3)$$

where $\mathcal{R}$ is the rasterizer, $\mathcal{E}$ is the encoding function and $\bar{\boldsymbol{x}}$ is the normalised version of the utilised 3DMM mean face (see (1)), so that the x-y-z coordinates of this face $\in [0,1]$. This representation is very convenient, as the rendering neural network learns to associate it with the corresponding RGB values, pixel by pixel, and, therefore, results in a realistic and novel video synthesis.

In addition to the $\mathbf{NMFC}$ images, we condition the neural video renderer on the gaze images ($\mathbf{G}_t$), as can be seen in Fig. 1. Gaze images are generated by connecting the extracted eye landmarks at the gaze tracking stage with edges, and filling the interior with color. The produced eye gaze frames are in exact correspondence with the NMFC frames, meaning that in both representations eyes should appear at the same pixel locations.

C. Video Rendering Neural Network

Given a sequence of $\mathbf{NMFC}$ frames $\mathbf{NMFC}_{1:T}$ and the corresponding sequence of eye gaze frames $\mathbf{G}_{1:T}$, the neural network learns to translate the conditional input video $\boldsymbol{x}_{1:T} \equiv \{x_t = (\mathbf{NMFC}_t, \mathbf{G}_t) | x_t \in^{H\times W\times 6}\}_{t=1,\dots,T}$ to a highly realistic and temporally coherent output video $\tilde{\boldsymbol{y}}_{\mathbf{1:T}}$, which shows the target actor performing exactly the same head motions and eye blinks as the actor in the source video. We train this network in a self-reenactment setting, where the source actor coincides with the target actor. Therefore, the generated video $\tilde{\boldsymbol{y}}^i_{1:T}$ should be a reconstruction of the ground truth $\boldsymbol{y}^i_{1:T}$, which is considered both as the source and target video. We adopt a GAN framework for video translation, where the generator G is trained in an adversarial manner, alongside an image discriminator D_I and a multi-scale dynamics discriminator D_D, which ensures that the generated video is realistic, temporally coherent and conveys the same dynamics of the target video. We further improve the visual quality of the mouth area, by designing a dedicated mouth discriminator D_M.

Generator G**.** In order to model the dependence of the produced frames on previous video time steps, we condition synthesis of the t-th frame $\tilde{y}_t$ not only on the conditional input x_t, but also on the previous inputs x_{t-1} and x_{t-2}, as well as the previously generated frames $\tilde{y}_{t-1}$ and $\tilde{y}_{t-2}$, thus:

$$\tilde{y}_t = G(\boldsymbol{x}_{t-2:t}, \tilde{\boldsymbol{y}}_{t-2:t-1}; \theta_G). \quad (4)$$

The generator is applied sequentially and the frames are produced one after the other, until the entire output

sequence has been produced. The architecture of this network is inspired by *vid2vid* [35]. It consists of two identical encoding pipelines (see Fig. 1), where the first one receives the concatenated **NMFC** and eye gaze images $\boldsymbol{x}_{t-2:t}$, while the second one takes in the two previously generated images $\tilde{\boldsymbol{y}}_{t-2:t-1}$. Their resulting features are first added and then passed through the decoding pipeline, which brings the output $\tilde{y}_t$ in a normalised [-1,+1] range, using a $\tanh$ activation.

Image discriminator D_I and mouth discriminator D_M. Both of these networks learn to distinguish real frames from synthesised ones. During training, a random time step t' in the range $[1, T]$ is uniformly sampled. Then, the real pair $(x_{t'}, y_{t'})$ and the fake one $(x_{t'}, \tilde{y}_{t'})$ are fed in D_I. Moreover, the corresponding mouth regions $(x^m_{t'}, y^m_{t'})$ and $(x^m_{t'}, \tilde{y}^m_{t'})$ are cropped and passed to D_M. In order to force these generators to create high-frequency details in local patches of the frames, we use a Markovian discriminator architecture (PatchGAN), as in [26] and [35].

Dynamics discriminator D_D. The dynamics discriminator is trained to detect videos with unrealistic temporal dynamics. This network receives a set of $K = 3$ consecutive real frames $\boldsymbol{y}_{t':t'+K-1}$ or fake frames $\tilde{\boldsymbol{y}}_{t':t'+K-1}$ in its input, which were randomly drawn from the video. To be more precise, D_D is not conditioned only on these short video clips of length K. Given the optical flow $\boldsymbol{w}_{1:T-1}$ of the ground truth video $\boldsymbol{y}_{1:T}$, the purpose of D_D is to ensure that the flow $\boldsymbol{w}_{t':t'+K-2}$ corresponds to the given video clip. Therefore, the dynamics discriminator should learn to identify the pair $(\boldsymbol{w}_{t':t'+K-2}, \boldsymbol{y}_{t':t'+K-1})$ as real and the pair $(\boldsymbol{w}_{t':t'+K-2}, \tilde{\boldsymbol{y}}_{t':t'+K-1})$ as fake. In practice, we employ a multiple scale dynamics discriminator, which performs the task described above in three different temporal scales. The first scale receives sequences in the original frame rate. Then, the two extra scales are formed by choosing not subsequent frames in the sequence, but sub-sampling the frames by a factor of K for each scale.

Objective function: The objective of our GAN-based framework can be expressed as an adversarial loss. We use the LSGAN [45] loss, thus the adversarial objective of the generator is:

$$\mathcal{L}^G_{adv} = \frac{1}{2}\mathbb{E}_{t'\sim\mathcal{U}\{1,T-2\}}[(D_D(\boldsymbol{w}_{t':t'+1}, \tilde{\boldsymbol{y}}_{t':t'+2}) - 1)^2] + \frac{1}{2}\mathbb{E}_{t'\sim\mathcal{U}\{1,T\}}[(D_I(x_{t'}, \tilde{y}_{t'}) - 1)^2 + D_M(x^m_{t'}, \tilde{y}^m_{t'}) - 1)^2]. \quad (5)$$

We add two more losses in the learning objective function of the generator: 1) a VGG loss $\mathcal{L}^G_{vgg}$ and 2) a feature matching loss $\mathcal{L}^G_{feat}$ [28], [35], which is based on the discriminators. Given a ground truth frame y_t and the synthesised frame $\tilde{y}_t$, we use the VGG network [46] to extract visual features in different layers for both frames and compute the VGG loss as in [28] and [35]. In a similar way, we compute the discriminator feature matching loss, by extracting features with the two discriminators D_I and D_D and computing the ℓ_1 distance of these features for a fake frame $\tilde{y}_t$ and the corresponding ground truth y_t. The total objective for G is given by:

$$\mathcal{L}^G = \mathcal{L}^G_{adv} + \lambda_{vgg}\mathcal{L}^G_{vgg} + \lambda_{feat}\mathcal{L}^G_{feat} \quad (6)$$

In order to balance out the loss terms above, we set $\lambda_{vgg} = \lambda_{feat} = 10$. The image and mouth discriminators as well as the dynamics discriminator are optimised under their corresponding adversarial objective functions. All discriminators share the same architecture, which is adopted from pip2pixHD [28]. More details on our video rendering stage are in the supplementary material.

Optical facial flow. To generate as realistic dynamics as possible by our **head2head** video rendering network, it is essential to condition the dynamics discriminator D_D on a very accurate facial flow of the target video. Most state-of-the-art methods for optical flow estimation solve this problem without any prior assumptions about the objects appearing in consecutive images. Since human facial performances exhibit non-rigid and composite deformations as a result of very complex facial muscular interactions, capturing their flow using off-the-shelf state-of-the-art optical flow methods might not always produce visually convincing synthesis. To overcome this issue, we capitalise on the prior knowledge, as we target facial videos, and train a specific network for this task. To start with, we utilise a state-of-the-art network, called FlowNet2, for the optical flow estimation [47]. This network was trained on publicly available images after rendering them with synthesised chairs modified by various affine transformations. We use the pretrained models of [47] and fine-tune their network on the 4DFAB dataset [48], which comes with dynamic high-resolution 4D videos of subjects eliciting spontaneous and posed facial behaviours. To create the ground truth 2D flow, we use the provided camera parameters of the acquisition device and rasterize the 3D scans of around 750K frames so that the difference between each pair of consecutive frames represents the 2D flow. For the background, we generate the 2D flow estimates of the same 750K frames using the original FlowNet2 and use a masked End-Point-Error (EPE) loss so that the background flow stays the same and the foreground follows the ground truth 2D flow coming from the 4DFAB dataset.

IV. Experiments

In this section, we demonstrate the capability of our **head2head** framework in transferring the full head pose, gaze, eye blinking and expression from a source to a target video. Our approach was compared with state-of-the-art methods and achieved very competitive, visually appealing and realistic results. We conduct comprehensive experiments and ablation studies, probing the performance of our proposed approach both quantitatively and qualitatively. We collected a database from publicly-available videos, all having a spatial resolution of 256×256 pixels. This database consists of multiple subjects, mainly well-known politicians (see supplementary material for more results and visualisations).

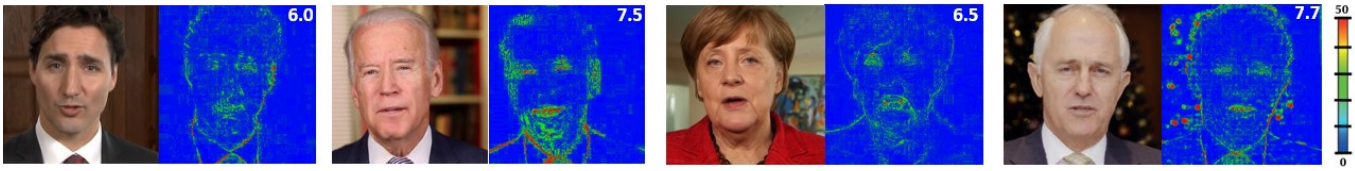

Fig. 2: Quantitative assessment of the self-reenactment experiment on four different videos. In each pair, (left) synthesised frame, (right) corresponding error (heat) map. White numbers in the corner of the heat maps represent the average error for the entire corresponding test sequence. All images are in the range [0, 255].

The **head2head** pipeline requires only a footage of a few minutes for training a model on the given target subject and takes only around 5 hours on an NVIDIA Titan V GPU to finish the training task.

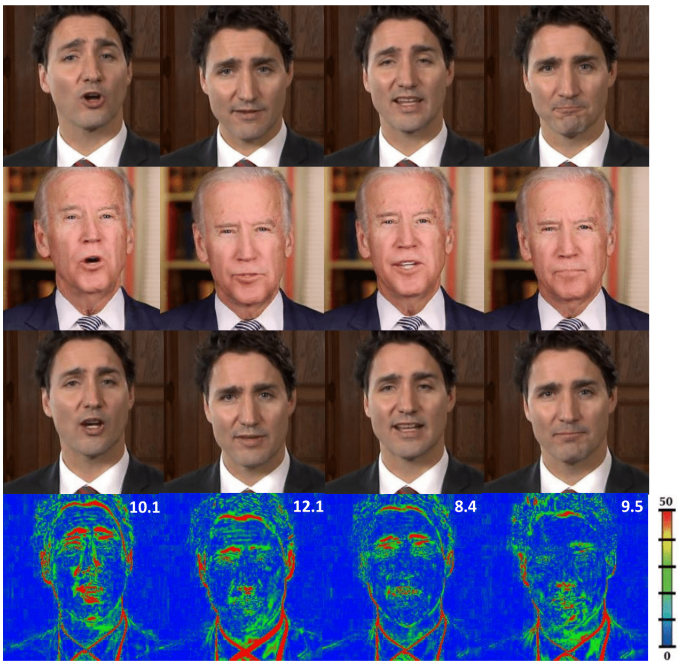

Fig. 3: Cycle-head2head-reenactment results. Top to bottom: driving (source) sequence, manipulated target sequence, manipulated source sequence driven by the manipulated target sequence in the row before, per-pixel l_2 distance between first and third row. Numbers in the top-right corners of error maps represent the average per-pixel error for each image.

A. Quantitative Results

We quantitatively evaluate the performance of our method by conducting two sets of experiments: 1) self-reenactment, and 2) cycle-reenactment. For self-reenactment, each target video was split into a train (first two thirds) and a test (last third) set, and the average per-pixel error was computed over the test split. Note that the per-pixel error is defined as the ℓ_2 distance between RGB images, assuming that the range of values in each color channel is from 0 to 255. Fig. 2 demonstrates the obtained quantitative results on four different subjects (Justin, Joe, Merkel and Turnbull), with an average of $4K$ training frames per video.

As a second experiment, we conducted a cycle-reenactment test to assess the performance of our approach in transparently transferring the human head attributes from a source to a target and then back to the same source. More specifically, given a source sequence $\boldsymbol{X}$ and a target actor $\boldsymbol{Y}$, we first train a network $\mathcal{N}_{\boldsymbol{Y}}$ on the target $\boldsymbol{Y}$. Then, during test time, we transfer expression, pose and gaze direction from the source to the target, generating a video $\tilde{\boldsymbol{Y}} = \mathcal{N}_{\boldsymbol{Y}}(\boldsymbol{X})$. To complete the cycle, another network $\mathcal{N}_{\boldsymbol{X}'}$ was trained on the same source subject of $\boldsymbol{X}$, using the training video $\boldsymbol{X}'$ with different frames from $\boldsymbol{X}$, such that $\boldsymbol{X} \cap \boldsymbol{X}' = \emptyset$. Then, another fake video was generated, as $\tilde{\boldsymbol{X}} = \mathcal{N}_{\boldsymbol{X}'}(\tilde{\boldsymbol{Y}})$ and the average per-pixel distance between $\tilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ was computed. Fig. 3 shows some randomly selected source-target-source ($\boldsymbol{X} \rightarrow \tilde{\boldsymbol{Y}} \rightarrow \tilde{\boldsymbol{X}}$) frames and the heat-maps (per-pixel ℓ_2 distance) between the frames in the first ($\boldsymbol{X}$) and third row ($\tilde{\boldsymbol{X}}$). The average per-pixel error over the entire test sequence, composed of 1K frames, is 9.3.

B. Qualitative Results

We show that our **head2head** framework performs a highly accurate transfer of the expression, head pose and eye gaze from the source to the target subject. We further compare our method with the state-of-the-art *Deep Video Portraits* [8] and two strong baselines: 1) *pix2pix* [26]

Fig. 4: Comparison of our method with the *pix2pix* [26] baseline conditioned on NMFC images, on the self-reenactment task. We achieve better image quality and ground truth reconstruction. Please zoom in for details.

Fig. 5: Qualitative comparison with *vid2vid* conditioned on landmarks [35]. The use of 3DMM and conditioning on **NMFC** images enables identity and expression disentanglement. Our neural video renderer preserves the identity of the target, in contrast to *vid2vid*, which distorts it.

conditioned on our novel **NMFC** and the eye gaze images and 2) *vid2vid* conditioned on facial landmarks, as in [35].

We trained both baseline methods and our model on two different target sequences (Obama and Turnbull). Then, we performed a self-reenactment experiment for the two target subjects and a full head transfer experiment from May to Obama, Putin to Obama and Putin to Turnbull. We found that the lack of sequential modeling when using an image-to-image transfer method, such as *pix2pix*, results in low quality and temporally incoherent frames. Examples from our self-reenactment experiment on the test set of Obama and Turnbull, are shown in Fig 4. It can be seen that our method produces more convincing results than the pix2pix baseline conditioned on **NMFC** and eyes images, for the task of video reconstruction. Next, we demonstrate the importance of the 3DMM and our novel **NMFC** conditioning input to the neural video renderer. Given a sequence of driving frames, our method successfully disentangles identity and expression and generates a set of target frames with the head movements of the source, while preserving facial attributes of the target identity. On the other hand, *vid2vid* model conditioning on landmarks distorts the appearance of the target, since landmarks contain identity information from the source sequence, which is then passed to the generated target video, as seen in Fig. 5. To make the comparison fairer, we rigidly registered the extracted landmarks of Putin to the target's. This helps to reduce the distortion coming from the difference in scale, otherwise the *vid2vid* network generates totally distorted and nonsensical images.

When comparing with [8], our method performs equally well in terms of image quality and photo-realism, while we achieve superior results in the interior of the mouth. As can be seen in Fig. 6, our method outperforms the pose and expression transfer of [8] in many cases. The head movements and facial expressions of our generated sequence match precisely with the ones in the source video.

C. Ablation Study

To evaluate the design choices made to build our **head2head** framework, we carry out an ablation study demonstrating the effect of each. We first show the effect of combining our framework with FlowNet2 [47] versus our facial flow. Table I (a) reports the average errors achieved when doing the same self-reenactment test in figure 2. We report smaller errors in all cases when using facial flow, which both justifies the significance of utilising our facial flow and makes it more descriptive of the facial dynamics. This is expected, as our flow capitalises on the prior knowledge and

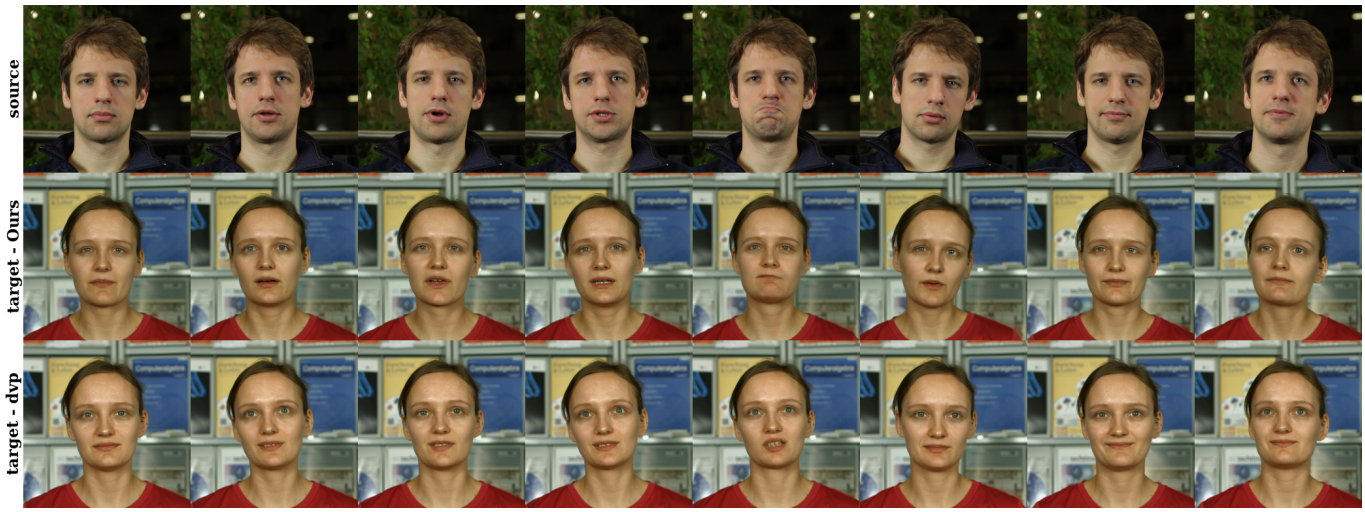

Fig. 6: Comparison with *Deep video Portraits* (dvp) [8]. Top to bottom: the driving (source) sequence, the target sequence generated with our **head2head** method, the corresponding frames with *Deep video Portraits* method. Our poses and expressions match better to the source. Please zoom in for details.

was exclusively trained on facial videos. Furthermore, we explore the significance of the mouth discriminator network D_M in the self-reenactment scenario, on four individual target subjects. We measure the average pixel distance in a constrained area around the mouth, between the frames generated with our method and the ground truth. As can be seen in Fig. 2 and Fig. 3, mouth cavity is a challenging region, with high errors, mainly due to the absence of conditional input information about the teeth and mouth interiors. The average pixel distances reported in Table I (b) indicate that the use of a mouth discriminator significantly improves our results. The improvements are demonstrated visually in Fig. 7.

TABLE I: Ablation study results.

(a) Average pixel distance obtained under a self-reenactment setup on the videos in figure 2 when combining our method with either FlowNet2 [47] or our facial flow.

Video	**Ours** (↓) **FlowNet2**	**Ours** (↓) **Face-Flow**
Justin	6.9	6
Joe	9.3	7.7
Merkel	9	7.5
Turnbull	9.4	6.5

(b) Average pixel distance around the mouth area, obtained under a self-reenactment scenario, with and without our dedicated mouth discriminator D_M.

Video	**Ours** (↓) **w/o** D_m	**Ours** (↓) **w/** D_m
May	11.2	10.1
Justin	11	10.4
Turnbull	11.3	10.7
Obama	22.3	19.4

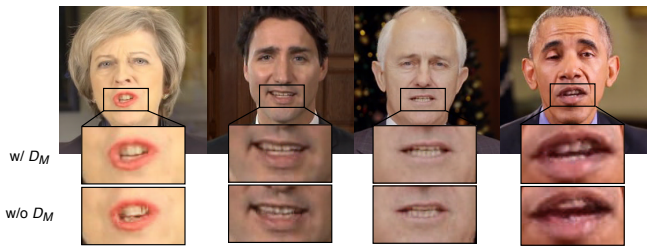

Fig. 7: Generated frames by our **head2head** method, with and without the mouth discriminator D_M.

D. Automated Study

A recent eye-catching attempt in the field to automatically detect fake videos manipulated by state-of-the-art facial manipulation methods was made by Rossler et al. [49]. With the help of a well-trained Convolutional Neural Network (CNN), the authors of [49] manage to outrank the performance of human observers in detecting manipulated videos. Their dataset (the largest publicly available dataset with facial manipulations) was created by manipulating $1,000$ YouTube videos, depicting real-world scenarios, with graphics-based [5], [50] and learning-based [51], [22], facial reenactment methods. In total, their forgery detection network was trained on 1.8 million facially manipulated frames and reported very high detection accuracy on the test split of their dataset (around 99%).

To assess the realism of our synthesised videos automatically, we utilise the trained forgery-detection network provided by the authors of [49]. We randomly choose a subset of 50 videos from their training dataset and manipulate them based on the selected source-target combinations in their work. The accuracy obtained by their network on our generated fake videos is only 1.88%. This demonstrates high photo-realism and consistency of our fake videos, since distinguishing them from real videos is challenging even for such a well-trained system.

E. User Study

In addition, we performed a user study with two parts. In the first one, we presented to the participants both real and synthesised videos from our method and asked them to assess how realistic they appear. The set of fake videos was generated under three different scenarios, namely: self-reenactment, face-reenactment, full-reenactment. All videos were 5-seconds long. We asked the following question *"On a scale of 1-5, how real does this video look?"*, with 1 meaning "absolutely fake" and 5 "absolutely real". Videos which received a 4 or 5 rating can be considered as "real". The ratings given by 95 anonymous participants are presented in Table II and Table III. The percentage of "real" videos according to the ratings of the participants is comparable for both synthetic and actually real videos. This strongly suggests that our method generates results almost indistinguishable from real videos and indicated that participants were excessively cautious about detecting fake videos, as they had the task explained in advance.

TABLE II: Ratings on the self-reenactment task. Columns 1-5 show the number of participants that gave this rating, while column "real" shows the percentage of people that rated the video with a 4 or 5.

Video	**Our synthesised videos**						**Real videos**					
	1	2	3	4	5	"real" (↑)	1	2	3	4	5	"real"
May	5	8	12	34	36	**74%**	3	6	6	26	54	84%
Turnbull	11	15	20	29	20	**52%**	2	6	14	40	33	77%
Putin*	4	10	18	36	27	**66%**	2	3	13	37	40	81%
Justin	5	8	18	36	28	**67%**	5	7	11	28	44	76%
dvp*	3	11	14	26	41	**71%**	9	12	21	29	24	56%
average	6	10	16	32	30	**66%**	4	7	13	32	39	**75%**

TABLE III: Ratings on the face-reenactment and full-reenactment tasks.

Video	**Our synthesised videos**						**Real videos**					
	1	2	3	4	5	"real" (↑)	1	2	3	4	5	"real"
Merkel (face-reenact.)	16	21	20	22	16	**40%**	1	6	17	31	40	**75%**
Justin (full-reenact.)	8	12	25	39	11	**53%**	5	7	11	28	44	76%
dvp* (full-reenact.)	37	19	10	22	7	**31%**	9	12	21	29	24	56%
average (full-reenact.)	22	16	18	30	9	**42%**	7	10	16	28	34	**66%**

As a second experiment, we presented a source (driving) video along with two target videos, one generated by our **head2head** method and the other one synthesised by [8]. Both fake videos were created under a full-reenactment setting, with the source and target subjects being those shown in Fig. 6. Then, we asked the question: *"Which video follows the movements of the person in the source better?"*. Out of

the 70 anonymous participants that answered this question, **60% have chosen the video produced by our method**, showing an indisputable preference towards our result.

V. Conclusion

We presented **head2head**, a full-reenactment approach for transferring the expression, head pose and the eye gaze from a source to a target actor. Different to the state-of-the-art, the proposed approach produces results that are temporally consistent and capitalises on priors on dense face flow. Furthermore, we pay particular attention to the mouth region in order to further improve quality and realism of the result. We demonstrate that our method can transfer facial motion and head movement more accurately than the state-of-the-art.

VI. Acknowledgements

S. Zafeiriou acknowledges funding from a Google Faculty Award, as well as from the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1).

References

[1] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *SIGGRAPH*, 2001.
[2] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li, "Realistic dynamic facial textures from a single image using gans," in *ICCV*, 2017.
[3] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*
[4] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM TOG*, 2015.
[5] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016.
[6] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," in *SIGGRAPH*, 2005.
[7] O. Wiles, A. Koepke, and A. Zisserman, "X2face: A network for controlling face generation by using images, audio, and pose codes," in *ECCV*, 2018.
[8] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Péerez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM TOG 2018*, 2018.
[9] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)*.
[10] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," *CoRR*, vol. abs/1905.08233, 2019.
[11] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *TPAMI*, 2014.
[12] P. Snape, Y. Panagakis, and S. Zafeiriou, "Automatic construction of robust spherical harmonic subspaces," in *CVPR*, 2015.
[13] P. Snape and S. Zafeiriou, "Kernel-pca analysis of surface normals for shape-from-shading," in *CVPR*, 2014.
[14] R. Garg, A. T. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *CVPR*, 2013.
[15] S. Grahof, H. Ackermann, F. Kuhnke, J. Ostermann, and S. S. Brandt, "Projective structure from facial motion," in *IAPR MVA*, 2017.
[16] M. Hernandez, T. Hassner, J. Choi, and G. Medioni, "Accurate 3d face reconstruction via prior constrained structure from motion," *Computers & Graphics*, 2017.
[17] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999.
[18] M. R. Koujan and A. Roussos, "Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction," in *ACM SIGGRAPH CVMP*, 2018.
[19] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou, "3d reconstruction of in-the-wild faces in images and videos," *TPAMI*, 2018.
[20] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM TOG*, 2017.
[21] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormaehlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *CVPR*, 2014.
[22] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM TOG 2019*, 2019.
[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.
[25] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv: 1411.1784*, 2014.
[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016.
[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
[28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018.
[29] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *arXiv:1703.00848*, 2017.
[30] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv:1611.02200*, 2016.
[31] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *ICCV*, 2017.
[32] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *CVPR*, 2018.
[33] E. L. Denton and v. Birodkar, "Unsupervised learning of disentangled representations from video," in *NIPS* (I. G. et al., ed.), 2017.
[34] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua, "Coherent online video style transfer," in *ICCV*, 2017.
[35] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *NeurIPS*, 2018.
[36] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *IJCV*, 2018.
[37] T. F. Coleman and Y. Li, "A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM Journal on Optimization*, 1996.
[38] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *CVPR*, 2016.
[39] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *IJCV*, 2018.
[40] S. Zafeiriou, G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis, "The 3d menpo facial landmark tracking challenge," in *ICCV*, 2017.
[41] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *TVCG*, 2014.
[42] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic, "Statistical non-rigid icp algorithm and its application to 3d face alignment," *Image and Vision Computing*, 2017.
[43] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *ECCV*, 2018.
[44] J. Guo, J. Deng, N. Xue, and S. Zafeiriou, "Stacked dense u-nets with dual transformers for robust face alignment," *arXiv preprint arXiv:1812.01936*, 2018.
[45] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[47] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.
[48] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications," in *CVPR*, 2018.
[49] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.
[50] Marek, "Faceswap." github.com/MarekKowalski/FaceSwap/, 09/2019.
[51] torzdf, "Deepfakes." github.com/deepfakes/faceswap, 09/2019.