Head2Head++: Deep Facial Attributes Re-Targeting

Michail Christos Doukas[*1,4], Mohammad Rami Koujan[*2,4], Viktoriia Sharmanska[1], Anastasios Roussos[2,3,4], and Stefanos Zafeiriou[1,4]

Abstract—Facial video re-targeting is a challenging problem aiming to modify the facial attributes of a target subject in a seamless manner by a driving monocular sequence. We leverage the 3D geometry of faces and Generative Adversarial Networks (GANs) to design a novel deep learning architecture for the task of facial and head reenactment. Our method is different to purely 3D model-based approaches, or recent image-based methods that use Deep Convolutional Neural Networks (DCNNs) to generate individual frames. We manage to capture the complex non-rigid facial motion from the driving monocular performances and synthesise temporally consistent videos, with the aid of a sequential Generator and an ad-hoc Dynamics Discriminator network. We conduct a comprehensive set of quantitative and qualitative tests and demonstrate experimentally that our proposed method can successfully transfer facial expressions, head pose and eye gaze from a source video to a target subject, in a photo-realistic and faithful fashion, better than other state-of-the-art methods. Most importantly, our system performs end-to-end reenactment in nearly real-time speed (18 fps).

Index Terms—face reenactment, full head reenactment, neural rendering, video renderer, 3DMM, 3D reconstruction, gaze tracking, temporal discriminator, facial flow.

1 Introduction

Image and video synthesis receives a rapidly increasing amount of attention in Computer Vision and Deep Learning research, as "synthetic" data appear more and more realistic, as well as especially promising for real-world applications. Video editing, film dubbing, social media content creation, teleconference and virtual assistance are some indicative examples. However, generating artificial human faces indistinguishable from real ones is a very challenging task, particularly when it comes to video data. Adding the extra dimension of time, might give rise to the so-called problem of temporal incoherence. As the uncanny valley effect suggests, people are extremely perceptible in unnatural facial and head movements, thus even small discontinuities can expose a synthetic video.

Over the past years, various methods that target human faces have emerged, taking advantage of different modalities, such as audio [1], [2] and video [3], [4], [5] to drive synthesis and dictate the movements of the generated subject. Facial reenactment is a widely-studied approach [3], [6], [7], which aims to transfer the facial expressions from a source to a target subject, by conditioning the generative process on the driving video of the source. In most cases [3], [7], this is done by modifying the deformations solely within the internal facial region of the target identity and placing the manipulated face back to the original target frames, using an image interpolation method. Given that face reenactment systems offer no control over the target's head pose and eye gaze, there might be cases where the person's expressions do not match with the overall head movement and therefore seem unnatural. Even so, face reenactment can be very useful for specific applications, such as video dubbing [8], which aims to alter the mouth motion of the target actor to match the audio track spoken by the dubber.

A more holistic approach on reenactment involves generating all pixels within frames, including the upper body, hair and background of the target identity. In contrast to video manipulation techniques, such as face reenactment, full head reenactment methods [4], [9] aim to transfer the entire head motion from a source identity to a target one, along with the eye blinking and gaze, providing complete control over the target subject. Most head reenactment approaches fall into two categories: a) Warping-based methods [2], [10], [11], which are mainly learning-based and do not involve computing priors such as facial landmarks or 3D face models, b) object-specific methods [3], [4], [9], [12], which assume knowledge (e.g. 3D reconstruction, 68 landmarks) of the source and target faces. On the one hand, warping-based methods usually suffer from distortions in the face and background [2]. On the other hand, object-specific methods relying on facial keypoints are affected from the so-called identity preservation problem. As keypoints encapsulate identity attributes of the source (e.g, head geometry), the more the identity of the source diverges from that of the target, the more distorted the head shape of the generated subject might appear. On the contrary, 3D morphable models (3DMMs) [13], [14] have proven to be a reliable means of decoupling expression and identity from each other. *Deep Video Portraits (DVP)* [4] is a full head reenactment system that capitalises on 3DMMs and utilises a neural network to translate 3D face reconstructions to realistic frames. Nonetheless, *DVP*, as a purely image-based model, does not take into account temporal dependencies

- $_*$ *denotes equal contribution*
- [1] *Department of Computing, Imperial College London, UK*
- [2] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK*
- [3] *Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), GR*
- [4] *FaceSoft.io, London, UK*

Manuscript received June 15, 2020; revised October 3, 2020

between frames.

Our recently proposed *Head2Head* [15] model overcomes the aforementioned limitations, as it combines the benefits of conditioning synthesis on 3D facial shapes with the advantages of a sequential, video-based, neural renderer. In this paper, we extend the work of *Head2Head* in the following directions:

- We prevail over the limitations of the 3D reconstruction stage of *Head2Head* by designing a novel **DenseFaceReg** network for the robust and fast estimation of semantic facial images based on 3D facial geometry. Our semantic facial images, referred to as **Normalised Mean Face Coordinates (NMFC)** images, capture pose, expression and identity information of the subject from the 3D facial mesh and are used to condition video synthesis.
- We propose a simple yet fast method for detecting the eye gaze, based on 68 facial landmarks. When combined with 3D Facial Recovery and Video Rendering Network, our system can operate in nearly real-time speeds.
- We conduct an extensive set of experiments, including user, automated and ablation studies. Our results reveal the significance of video-based modeling.
- We make our code and dataset publicly available[1].

2 Related Work

2.1 3D Face Reconstruction

Recovering the 3D geometry of human faces from monocular images is a challenging problem that has attracted much attention due to its major role in many applications, ranging from facial reenactment, performance capture and tracking [16], facial expression recognition [17], [18], etc. Owing to their pose and illumination invariance, 3D facial data constitute an indispensable geometrical description of faces for various facial image processing systems. The Computer Vision field is rich in approaches targeting this problem under different assumptions and constraints. Some of these attempts [19], [20], named *Shape from Shading*, approximate the image formation process and make simplified assumptions about the lighting and illumination models leading to the formation of the image, while others, known as *Structure from Motion (SfM)*, benefit from the geometric constraints in multiple images of the same object to solve the problem [21], [22]. One common approach for addressing this task are the *3D Morphable Models (3DMMs)* of the human face. 3DMMs are linear statistical models that have been used substantially since the pioneering work of Blanz and Vetter [23], with many extensions [13], [14], [24]. With the rise of deep neural networks, nonlinear face models have been proposed to recover the 3D geometry and appearance of human faces from images or videos using *deep Convolutional Neural Networks (CNNs)* [25], [26], [27], [28]. For a very recent and comprehensive review of the state-of-the-art methods on monocular 3D face reconstruction and the open challenges of 3DMMs, we refer the readers to [29], [30].

In this work, we employ 3DMMs and a 3D shape regression network for: 1) estimating the 3D geometry of faces appearing in videos, and 2) utilising the estimated 3D geometrical information for driving our Deep Video Rendering Neural Network.

1. https://github.com/michaildoukas/head2head

2.2 Facial Synthesis and Re-targeting

Various deep architectures have been proposed for the image and video synthesis tasks with the aid of Recurrent Neural Networks (RNNs), Variation Auto-encoders (VAE) [31], Gaussian mixture VAE [32], Hierarchical VAE [33], Generative Adversarial Networks (GANs) [34] and VAE-GAN [35].

Traditional methods of reenactment, transfer the facial expressions either with 2D warping techniques [10], [11], or by utilizing 3D face models [3], [6], [7]. These methods do not provide complete control over the generated video, as they manipulate only the interior of the face. Recently, there has been a substantial effort in the direction of both expression and pose transfer [2], [4], [9], [36], [37]. One of the first approaches, *X2Face* [2], designs an embedding and a set of auto-encoders. *X2face* follows a warping-based approach that causes deformations in the generated heads and inconsistent upper-body motions. Wang et al. [12] propose *vid2vid*, a GAN-based spatio-temporal approach for the video-to-video synthesis, relying on a Temporal Discriminator for improving the temporal quality of the synthesised videos. In a follow-up work, Wang et al. [38] extend their approach with an attention mechanism, making it trainable in a few-shot manner and leading to better generalisation performance. Both [12] and [38] can perform reenactment with face sketches drawn from landmarks, leading to a target identity preservation problem. Moreover,. the synthesised mouth regions do not look very realistic.

Zakharov et al. [9] propose a few-shot, image-based adversarial learning approach, as their network learns to generate unseen target identities even from a single image. Nonetheless, their neural network relies on landmarks, causing an identity distortion of the target. Siarohin et al. [39] animate objects via a deep motion transfer framework. Given a single image and a driving sequence, their method estimates a dense motion field appearing in the sequence and transfers it to the target image while preserving its appearance. When used for facial reenactment, the faces synthesised by this approach suffer from head distortions and non-naturalistic mouth and teeth areas. To tackle this issue, the authors extended their work [40] to account for complex motions with a first-order motion model and an occlusion-aware Generator. Although this extension considerably improved the results on various tasks, synthesised faces still exhibited visual artifacts appearing as expression-dependent continuous scale changes. This can be attributed to the estimated 2D dense motion field that does not fully describe the actual intricate 3D facial motion.

As far as we are aware, *DVP* [4] is the only learning-based head reenactment system, prior to our work, that uses 3D facial information to condition video synthesis. Their image-based model requires a long video footage of the target person, while training a new model for each target takes many hours. Moreover, generated mouths look unnatural, since the 3D reconstruction method they adopt does not encode the inner-mouth region or the teeth.

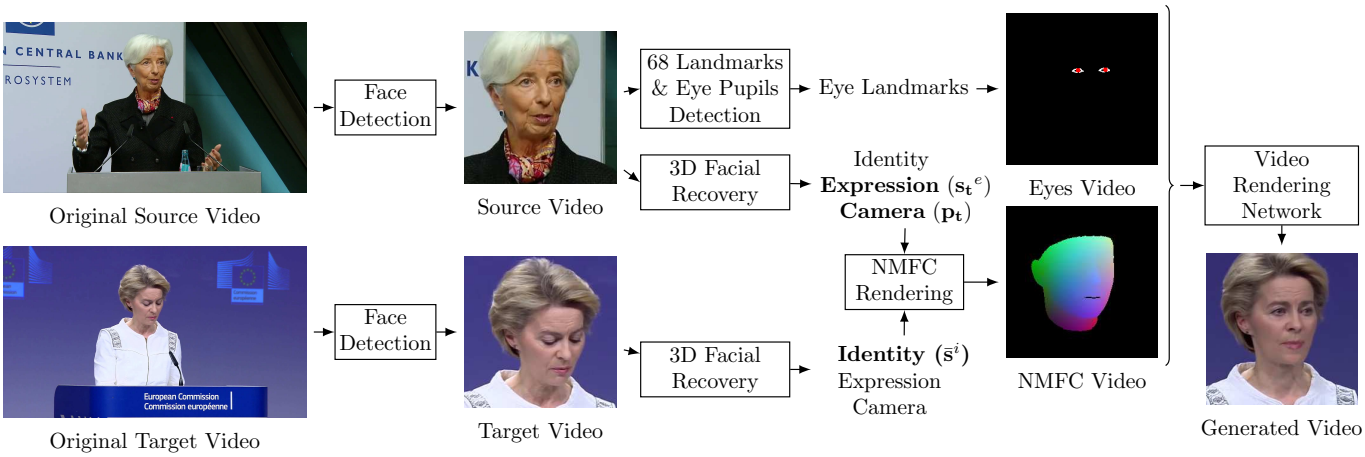


Fig. 1. Our Head2Head++ pipeline for head reenactment. First, the facial region of interest is extracted from both the source and target original videos. Next, 3D Facial Recovery is performed. The average identity parameters ($\bar{\mathbf{s}}^i$) of the target, along with the expression ($\mathbf{s}_t^e$) and camera ($\mathbf{p}_t$) parameters of the source for each time-step t, are passed to the NMFC renderer, which produces the NMFC video. At the same time, the eye landmarks of the source are used to detect the pupils and sketch the eyes video, which represents eyes movements. Finally, the combined NMFC and eyes video is given as conditional input to the Video Rendering Network, which computes the generated video (reenactment result). The Video Rendering Network is a person specific model, as it is trained using footage of the target person. During training, the source video coincides with the target video, enabling us to perform self-reenactment and get access to ground truth data.

Unlike other studies, our approach employs efficiently the 3D geometry and conditions frame generation on a compact and meaningful representation in the image space derived from 3D facial reconstruction. When combined with our novel video-based neural rendering stage, the result is a faithful and photo-realistic full head video reenactment, indistinguishable from real videos. Additionally, we focus specifically on the mouth area and improve its visual quality by designing a dedicated discriminator.

3 Methodology

Our *Head2Head++* framework proposes a solution for the largely ill-posed full head reenactment problem. Our method is capable of transferring the time-varying head attributes (pose, expression, eye gaze) and mainly consists of two subsequent modules: a) **3D Facial Recovery** (Sec. 3.1), and b) a **Video Rendering Network** (Sec. 3.4). Our video rendering network capitalises on a carefully designed GAN-based framework. Please refer to Fig. 1 for an overview of our *Head2Head++* pipeline.

3.1 3D Facial Recovery

We aim to generate a reliable estimation of the facial 3D geometry, capturing the temporal dynamics, while separating the identity and expression contributions of the pictured subject in each frame. We utilise this separability to effectively disentangle the human head characteristics in a transferable and photo-realistic way between different videos. Towards that aim, we benefit from the prior knowledge in our problem space and harness the power of 3DMMs [23] for reconstructing the faces that appear in the input monocular sequences. Given a sequence of T frames $\mathcal{F}_{1:T} = \{f_t \mid t = 1, \ldots, T\}$, the 3D reconstruction stage produces two sets of parameters: 1) shape parameters $\mathcal{S} = \{\mathbf{s}_t \mid \mathbf{s}_t \in \mathbb{R}^{n_i+n_e}, t = 1, ..., T\}$, and 2) camera parameters $\mathcal{P} = \{\mathbf{p}_t \mid \mathbf{p}_t \in \mathbb{R}^6, t = 1, ..., T\}$, depicting rotation, translation and orthographic scale.

Shape representation. Using 3DMMs, a 3D facial shape $\mathbf{x}_t = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^\top \in \mathbb{R}^{3N}$ can be written mathematically as:

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

where $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is the mean shape of the morphable model, given by $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{id} + \bar{\mathbf{x}}_{exp}$, with $\bar{\mathbf{x}}_{id}$ and $\bar{\mathbf{x}}_{exp}$ standing for 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 $\mathbf{s}_t^i \in \mathbb{R}^{n_i}$, $\mathbf{s}_t^e \in \mathbb{R}^{n_e}$ are the identity and expression parameters of the morphable model. We designate the joint identity and expression parameters at time-step t by $\mathbf{s}_t = [\mathbf{s}_t^{i\top}, \mathbf{s}_t^{e\top}]^\top$. In the adopted model (1), the 3D facial shape $\mathbf{x}$ is a function of both identity and expression coefficients ($\mathbf{x}(\mathbf{s}_t^i, \mathbf{s}_t^e)$), where expression variations are effectively represented as offsets from a given identity shape.

Video-based 3D reconstruction. The video fitting approach followed in *Head2Head* [15] to estimate the 3D facial geometry is based on a set of sparse landmarks extracted from the entire input sequence. This method has three main drawbacks: 1) the fidelity of the 3D reconstruction relies heavily on the accuracy of extracted landmarks which are also sparse (68 in total), 2) it might require a large number of frames with enough reconstruction cues (various rotations) to produce good accuracy, 3) it makes a quite strong assumption in the initialisation stage about the rigidity of the face to estimate the camera parameters. To overcome these limitations, we propose to perform the 3D facial reconstruction in this work by training a deep CNN, we call **DenseFaceReg**, the purpose thereof is to produce a dense 3D facial mesh from a single RGB frame. We use a thousand

annotated videos from the **Face3DVid** dataset of Koujan et al. [16] to train this network in a supervised manner. The adopted loss function during training is:

$$\mathcal{L}(\Phi) = \sum_{i=1}^{N} ||\mathbf{v}_i^{GT} - \mathbf{v}_i||^2. \quad (2)$$

Equation (2) penalises the deviation of each vertex from the corresponding ground-truth vertex ($\mathbf{v}_i = [x, y, z]^\top$). We use the camera parameters provided with the **Face3DVid** dataset to project the 3D ground-truth mesh, so that our **DenseFaceReg** produces 3D vertices (dense landmarks) directly in the image space. We use the dense 3D vertices ($\sim$5K) estimated by our trained **DenseFaceReg** on each video frame to: 1) estimate the camera parameters, 2) generate the 3DMM identity and expression coefficients by projecting the dense shape onto the 3DMM bases. For all our experiments in this work, we use the same 3DMMs utilised in [15].

The analysis-by-synthesis approach, which is used by many state-of-the-art approaches [3], [4], estimates a lot of parameters (e.g. illumination, reflectance, shape, etc) and solves a highly ill-posed problem for fitting 3DMMs to images. On the contrary, our facial reconstruction stage is a fast CNN-based approach (6ms test runtime) trained on a large number of in-the-wild videos. The facial representation extracted with our method is informative enough to synthesise photo-realistic and temporally smooth videos, eliminating the need for more elaborate and slower 3D facial reconstruction techniques.

3.2 Facial Semantic Representation - NMFC Rendering

Our video rendering network receives as input a facial semantic representation of the target subject, with the head pose and facial expression guided by the source frames. This representation disentangles identity from expression, allowing us to train our video rendering network on a specific target person. During test, we are able to transfer the expression and pose of any source, with different head characteristics, to the target. Given the recovered identity and expression parameters (facial shape) $\mathbf{s}_t = [\mathbf{s}_t^{i\top}, \mathbf{s}_t^{e\top}]^\top$ and camera parameters $\mathbf{p}_t$, at frame t, we rasterize the 3D shape, producing a visibility mask ($\mathbf{M} \in \mathbb{R}^{W \times H}$) in the image space. Each pixel of $\mathbf{M}$, stores the index of the corresponding visible triangle on the 3D face seen from this pixel. Thereafter, we store the normalised x-y-z coordinates of the centre of this triangle in the $\mathbf{NMFC} \in \mathbb{R}^{W \times H \times 3}$ image, which is the facial semantic representation that is utilised as conditional input to the video rendering network. Equation (3) details this process.

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

where $\mathcal{R}$ is the rasterizer, $\mathcal{E}$ is the encoding function and $\bar{\mathbf{x}}$ is the normalised version of the 3DMM mean face (see (1)), so that the x-y-z coordinates of this face belong in $[0, 1]$. The NMFC image generation relies not only on the 3D reconstructed face and camera parameters of the current frame but also on an always-fixed normalised mean face ($\bar{\mathbf{x}}$) of the employed 3DMM. The (x, y, z) coordinates of the normalised mean face are used as constant colors to texture the 3D face of the current input frame f_t. The NMFC image of a frame f_t is generated then by rendering the corresponding textured 3D mesh. Since we texture any reconstructed 3D mesh with always a fixed set of distinctive colors, the same semantic point, say the tip of the nose, in any NMFC image (regardless of the subject and the target video) will always have the same color. This is why we refer to NMFCs as semantic images. The main advantage of using NMFCs over a UV 2D parameterisation is that it proved experimentally to be easier for the video renderer to learn the mapping to the output RGB images, since both the NMFCs and the output are in the same space. Additionally, this representation is more compact and easy-to-interpret by our video rendering network, since it associates well with the corresponding RGB frame to be generated, pixel by pixel, and, subsequently, leads to a realistic and novel video synthesis. Note that during test time, the expression coefficients $\mathbf{s}_t^e$ and camera parameters $\mathbf{p}_t$ are estimated from the source video at frame t, while the identity coefficients $\mathbf{s}^i$ are the average identity parameters estimated using all frames of the target video (see Fig. 1).

3.3 Eye Pupils Detection

We choose a real-time operating method for the extraction of the eye movements from the source frames, which does not require fitting an eye model [41], but is based on 68 facial landmarks [42]. Given the subset of landmark points $\mathbf{L}^{eye} \in \mathbb{R}^{6\times 2}$ that correspond to the left or right eye, we estimate one more landmark $\mathbf{l}^{pupil} \in \mathbb{R}^2$, which corresponds to the eye pupil. The eye pupil is obtained as the centre of mass within the set of pixels $\mathbf{\Omega}$ that are bounded by the polygon formed with the eye landmarks $\mathbf{L}^{eye}$ by computing a weighted sum, using the inverse intensity of pixels in $\mathbf{\Omega}$ as weights. Following [43], we are based on the assumption that eye pupils are the "darker" areas within the eye region in terms of pixel intensity. Mathematically, the eye centre of mass is given as

$$\mathbf{l}^{pupil} = \frac{\sum_{\mathbf{p}\in\mathbf{\Omega}} I(\mathbf{p})\mathbf{p}}{\sum_{\mathbf{p}\in\mathbf{\Omega}} I(\mathbf{p})}, \quad (4)$$

where $I(\mathbf{p})$ is the inverse intensity of pixel $\mathbf{p}$ in the source frame. In this way, $\mathbf{l}^{pupil}$ corresponds to the center of "darker" pixels within the eye region.

Once the eyes and pupils landmarks are estimated, we create the eyes video $\mathbf{E}_{1:T}$, which is a sequence of T RGB frames of size $256 \times 256 \times 3$. An example of these frames is shown in Fig. 1. We connect the eye landmarks with white edges to create an outline of each eye. Then, two red circles are drawn on the same plane, using as centres the eye pupil coordinates of each eye, in order to indicate eye gaze.

3.4 Deep Video Rendering Neural Network

Our carefully-designed video rendering network receives as conditional input two sequences of images, namely: the NMFC video $\mathbf{NMFC}_{1:T}$ and the corresponding eye video $\mathbf{E}_{1:T}$, collectively as $\mathbf{X}_{1:T} \equiv \{\mathbf{X}_t = (\mathbf{NMFC}_t, \mathbf{E}_t)\}_{t=1,\dots,T}$, with $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 6}$. With this input, the video rendering

network yields a highly realistic and temporally coherent output video $\tilde{\mathbf{Y}}_{1:T}$ picturing the target actor mimicking exactly the same expressions, head pose and eyes reflected in $\mathbf{X}_{1:T}$. While training, we follow a self reenactment setting where the source and target videos are the same footage. In this way, the reenacted video in the output should be a replication of the RGB training target video $\mathbf{Y}_{1:T}$, which serves as ground truth. Our video rendering network is trained within a GAN-based structure with: 1) a Generator G, which is the video rendering network itself, 2) an Image Discriminator D_I, 3) a multi-scale Dynamics Discriminator D_D enforcing the temporal coherence and realism of the produced facial performance, and 4) a dedicated Mouth Discriminator D_M, which further improves the visual quality of the mouth area.

Generator G. To successfully synthesise smooth and convincing temporal facial performances, we condition the synthesis of the t-th frame $\tilde{\mathbf{Y}}_t$ not only on the conditional input $\mathbf{X}_t$, but also on the previous inputs $\mathbf{X}_{t-1}$ and $\mathbf{X}_{t-2}$, as well as the previously generated frames $\tilde{\mathbf{Y}}_{t-1}$ and $\tilde{\mathbf{Y}}_{t-2}$, thus:

$$\tilde{\mathbf{Y}}_t = G(\mathbf{X}_{t-2:t}, \tilde{\mathbf{Y}}_{t-2:t-1}). \tag{5}$$

The Generator is applied sequentially, producing the output frames one after the other, until the entire output sequence has been created. Similar to *Head2Head* [15], the Generator consists of two identical encoders, operating in parallel, as well as a decoder. The first encoder receives the concatenated NMFC and eye images $\mathbf{X}_{t-2:t}$, while the second is given the two previously generated frames $\tilde{\mathbf{Y}}_{t-2:t-1}$. The two extracted feature maps are first added and then passed through the decoder, which brings the output $\tilde{\mathbf{Y}}_t$ in a normalised [-1,+1] range, using a $\tanh$ activation function.

Image Discriminator D_I and Mouth Discriminator D_M. Both of these networks aim at telling real and synthesised frames apart, and are used only during training. At time-step t, the Image Discriminator D_I processes the real pair $(\mathbf{X}_t, \mathbf{Y}_t)$ and the fake one $(\mathbf{X}_t, \tilde{\mathbf{Y}}_t)$. Concurrently, the corresponding mouth regions $(\mathbf{X}_t^m, \mathbf{Y}_t^m)$ and $(\mathbf{X}_t^m, \tilde{\mathbf{Y}}_t^m)$ are cropped and sent to the Mouth Discriminator D_M.

Dynamics Discriminator D_D. With the aim of learning complex facial dynamics in mind, we further equip our GAN framework with a Dynamics Discriminator D_D. During training, D_D receives a set of three consecutive real frames $\mathbf{Y}_{t:t+2}$ or fake frames $\tilde{\mathbf{Y}}_{t:t+2}$. They are passed to D_D after being associated with the optical flow $\mathbf{W}_{1:T-1}$, computed from the real frames (training video $\mathbf{Y}_{1:T}$ of target subject). Following this, the Generator is encouraged to create fake frames showing the same flow (dynamics) as the corresponding real ones. The Dynamics Discriminator learns to differentiate between the real $(\mathbf{W}_{t:t+1}, \mathbf{Y}_{t:t+2})$ and fake $(\mathbf{W}_{t:t+1}, \tilde{\mathbf{Y}}_{t:t+2})$ pairs. In practice, we employ a multiple-scale Dynamics Discriminator, which operates in three different temporal scales. The first scale receives frame sequences in the original frame rate. Then, the two extra scales are formed by sub-sampling the frames by a factor of two for each scale.

Objective function: In order to train our Generator to synthesise samples as real as possible, we formulate an adversarial loss. More specifically, we use LSGAN [44] with labels $a = c = 1$ for fake samples and label $b = 0$ for real ones, resulting in the following adversarial objective for the Generator:

$$\begin{aligned}\mathcal{L}_{adv}^G = &\frac{1}{2}\mathbb{E}_t[(D_D(\mathbf{W}_{t:t+1}, \tilde{\mathbf{Y}}_{t:t+2}) - 1)^2] \\ &+ \frac{1}{2}\mathbb{E}_t[(D_I(\mathbf{X}_t, \tilde{\mathbf{Y}}_t) - 1)^2 + (D_M(\mathbf{X}_t^m, \tilde{\mathbf{Y}}_{t'}^m) - 1)^2].\end{aligned} \tag{6}$$

We add two more losses in the learning objective function of the Generator: 1) a VGG loss $\mathcal{L}_{vgg}^G$ and 2) a feature matching loss $\mathcal{L}_{feat}^G$, first proposed in the work of Xu et al. [45]. Our feature matching loss term is based on the activations of both the Image and Dynamics Discriminators. Given a ground-truth frame $\mathbf{Y}_t$ and the synthesised frame $\tilde{\mathbf{Y}}_t$, we use the VGG network [46] to extract visual features in different layers for both frames and compute the VGG loss as in [47] and [12]. Likewise, the feature matching loss is computed 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 $\tilde{\mathbf{Y}}_t$ and the corresponding ground truth $\mathbf{Y}_t$. For the computation of the overall feature matching loss, we compute a loss term $\mathcal{L}_{feat}^{G-D_I}$ using features extracted by D_I, as well as another loss term $\mathcal{L}_{feat}^{G-D_D}$, using the dynamics discriminator D_I and then we add them: $\mathcal{L}_{feat}^G = \mathcal{L}_{feat}^{G-D_I} + \mathcal{L}_{feat}^{G-D_D}$. The total objective for G is given by:

$$\mathcal{L}^G = \mathcal{L}_{adv}^G + \lambda_{vgg}\mathcal{L}_{vgg}^G + \lambda_{feat}\mathcal{L}_{feat}^G \tag{7}$$

To achieve a balance between the loss terms above, we set $\lambda_{vgg} = \lambda_{feat} = 10$. The Image, Mouth and Dynamics Discriminators are optimised under their corresponding adversarial objective functions (see Supplementary Material).

Facial dynamics. One key factor to consider while designing the Dynamics Discriminator is how it could learn the complex realistic facial muscular interactions observed in monocular videos of facial performances. As discussed and validated experimentally in [16], off-the-shelf state-of-the-art optical flow methods solve this problem without any prior knowledge, since they target any moving objects in the scene. When applied to faces captured in the wild, the estimated flow does not reflect faithfully the non-rigid and composite facial deformations. To tackle this, we employ a state-of-the-art network, termed as FlowNet2, for the optical flow estimation [48]. The authors of [48] trained this network on publicly available images after rendering them with synthesised chairs modified by various affine transformations. Starting from the pretrained models of [48], we fine-tune their network on the 4DFAB dataset [49]. This dataset has dynamic high-resolution 4D videos of subjects eliciting spontaneous and posed facial behaviours. Our fine-tuned FlowNet2 network is therefore utilised to estimate the optical flow matrix $\mathbf{W}$ referred to in equation 6.

4 Experiments

4.1 Dataset and Implementation Details

We train and test *Head2Head++* on a newly collected video dataset, which we make publicly available, consisting of eight different individuals. Each individual is depicted in one video, which is at least 10 minutes long and presents the target in diverse head poses and facial expressions. First,

Fig. 2. Head reenactment aims at transferring the expressions, pose and eye movements from the driving sequence to the target identity.

we apply face detection on each video, as a pre-processing step. We utilise [50] to obtain a bounding box per frame, and then we compute the average bounding box throughout frames. We extract the facial ROI of 256×256 pixels for each frame, according to this fixed bounding box. This way, the background remains as stable and unchanged as possible, from the beginning of the video until the end. Finally, the frames of each subject are split into a training and a test set. Please refer to the Supplementary Material for more details on the dataset.

We base our Generator's architecture on our previous work [15]. All discriminators have the same architecture, adopted by *pip2pixHD* [47]. We train a separate person- and video-specific Video Rendering Network for each one of the eight target identities. Each trained model at the end is dedicated for the target video used during training and does not generalise to synthesise the same target person under different cloths and backgrounds (only video specific). Given a target video footage of 10 minutes, extracting the eyes and NMFC conditional input sequences and training, the GAN framework requires around 20 hours. Networks are optimised with Adam [51], for 40 epochs with an initial learning rate $\eta = 2 \cdot 10^4$, $\beta_1 = 0.5, \beta_2 = 0.999$, on a single NVIDIA GeForce RTX2080 Ti. During test time, our *Head2Head++* pipeline performs head reenactment from web-camera captures in nearly real-time speeds (18 fps).

4.2 Supported Reenactment Methods

As already mentioned, we have chosen to use 3DMMs for the retrieval of 3D facial shapes from the source and target videos in order to address the target identity preservation problem, allowing us to disentangle expression from identity seamlessly. Therefore, *Head2Head++* can be used both for full head reenactment and simple facial reenactment, depending on which set of camera parameters (target or source) are used during the creation of the conditional NMFC sequence.

Head Reenactmenent. This is the main functionality of our method and arguably the most challenging, since it involves transferring the complete head motion from any driving video to the desired target subject. A head reenactment example is demonstrated in Fig. 2. Given the driving sequence (Johnson), our method synthesises a highly photo-realistic and temporally coherent video of the targeted identity (Trump).

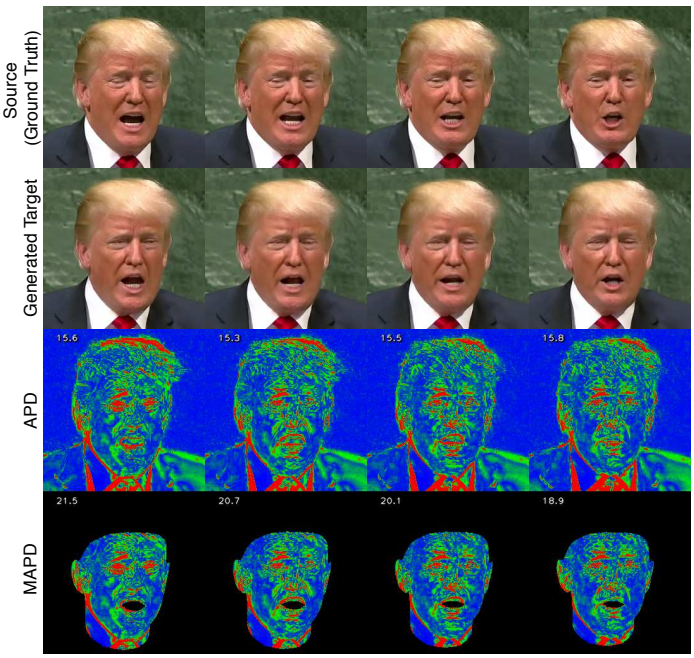


Fig. 3. Self reenactment is used for evaluating the performance of models. Here, the source identity coincides with the target one. We display Average Pixel Distance (APD) and Masked Average Pixel Distance (MAPD) between the generated and ground truth frames, in the form of RGB heatmaps. Please refer to Sec. 4.2 for more details on those metrics.

Face Reenactmenent. Instead of transferring all head movements from the source video, we can simply pass the facial expressions to the target identity. This is done by utilizing the original camera parameters, estimated from the target training sequence. In this case, we use the original eye landmarks extracted from the training sequence to drive synthesis. The advantage of face reenactment is that our Generator has encountered the same head poses and positions during training, which is beneficial when it comes to synthesising the hair, background and upper body areas, where no conditional information is available in the input. Although prior works have mainly used it as a manipulation method [3], [7], we perform face reenactment by generating all pixels within frames, not only a masked area of the face.

Self Reenactmenent. During self reenactment, the source identity coincides with the target one. This is extremely useful for training and evaluating our model, since we are given access to the ground truth frames. Note that the driving sequences used during test have not been seen during training, since they belong to the test data split. Fig. 3 shows a self reenactment experiment. Ideally, the generated frames should be identical with the source ones. For the evaluation of the reconstructive ability of models in the experiments that follow, we use metrics such as the average pixel distance (APD) or masked average pixel distance (MAPD) between the synthesised and ground truth frames. In Sec. 4.3, we provide more details on APD and MAPD.

Fig. 4. Significance of Mouth Discriminator. Please zoom in for details.

4.3 Quantitative Evaluation Metrics

We evaluate the performance of our system both qualitatively and quantitatively. Most experiments were performed under self reenactment. That is, given a set of test frames $\mathbf{Y}^{(i)}$ for each target identity i in the dataset, we generated the corresponding synthetic frames $\tilde{\mathbf{Y}}^{(i)}$ and used $\mathbf{Y}^{(i)}$ as ground truth. We assessed our method's reconstructive ability, target identity preservation, pose and expression transferability as well as the photorealism of generated frames, using the metrics below:

Average Pixel Distance (APD): is computed as the average L2-distance of RGB values across all spatial locations and frames, between the ground truth and generated data.

Masked Average Pixel Distance (MAPD): similar to APD, it tests the reconstructive performance. A mask computed from NMFC frames is used to constrain the metric on the facial area, where conditional information is available.

Distance between Average Identities (DAI): after performing 3D face reconstruction on $\mathbf{Y}^{(i)}$ and $\tilde{\mathbf{Y}}^{(i)}$ sequences, we compute the average identity coefficients for each sequence, and then the L1-distance between those average identities.

Average Expression Distance (AED): after reconstructing the fake and ground truth sequences, we measure the average L1-distance between the expression coefficients across frames.

Average Rotation Distance (ARD): this metric is used to measure pose transferability. Using the estimated camera parameters from the generated and ground videos, we compute the Euler angles that correspond to head poses. Then, the average Euler angles discrepancy across sequences is determined in terms of degrees.

Fréchet Inception Distance (FID): this is a widely used metric for evaluating the visual quality of individual frames, originally proposed in [52]. We use [53] as a feature extractor and we report the mean FID throughout the experiments.

Maximum Mean Discrepancy (MMD^2): is a very useful metric for measuring the discrepancy between real and fake frames [54].

4.4 Ablation Study

We conducted an ablation study, in order to assess the significance of several "key" components of our *Head2Head++* system. First, we evaluate numerically the contribution of the two reconstructive loss terms, namely the feature matching and VGG loss. As can be seen in the first two rows of Table 1, all metrics demonstrate the contribution of these loss terms on the quality of the generated samples. The next row of the table validates the importance of a sequential Generator, as opposed to its non-sequential variation ("w/o seq. G"). The results indicate that considering previously generated frames in the Generator has a very positive effect on the visual quality of samples. Finally, the majority of metrics show that removing the Dynamics Discriminator ("w/o D_D") might slightly improve the quality of frames, when seen as individual images. We believe this is due to the fact that the proposed metrics analyse solely the spatial content of frames and do not evaluate temporal information. Removing D_D network would actually reduce the temporal stability and coherence of videos. Such behaviour is more apparent in areas where no conditional information is available (e.g. hair, background, upper body). In order to understand better the significance of the Dynamics Discriminator in the visual plausibility of the output video, please refer to our Supplementary Video.

Arguably, the eyes and the mouth, are the facial areas that might be the first to expose the "fakeness" of a synthetic video. In Fig. 4, we illustrate the importance of the dedicated Mouth Discriminator, initially proposed in [15]. As can be

Fig. 5. Significance of conditioning on the eyes. The blinking of the source is successfully transferred to the target when using eye landmarks to condition synthesis. On the contrary, conditioning solely on NMFC frames is not sufficient for transferring eye gaze and blinking.

TABLE 1
Ablation study results under the self reenactment setting, averaged across all eight target identities in our dataset. For all metrics, lower values indicate better quality or performance. Bold and underlined values correspond to the best and the second-best value of each metric, respectively.

Variations	APD	MAPD	AED	ARD	$\times 10^2$ FID	$\times 10^5$ MMD^2
w/o $\mathcal{L}^G_{feat}$	17.62	16.70	0.627	0.539°	6.98	13.23
w/o $\mathcal{L}^G_{vgg}$	15.74	13.95	0.514	0.452°	5.40	9.42
w/o seq. G	15.71	13.63	0.486	0.436°	4.98	9.06
w/o D_D	<u>15.09</u>	**13.03**	**0.456**	**0.408°**	**3.95**	**5.57**
Full model	**14.82**	<u>13.06</u>	<u>0.467</u>	<u>0.412°</u>	<u>4.15</u>	<u>6.50</u>

TABLE 2
Ablation study of the effect of Mouth Discriminator D_M and conditioning on the eyes in *Head2Head++*.

Variations	APD	AELD
w/o D_M	13.43	-
w/o Eyes input	-	1.52
Full model	**12.32**	**1.06**

TABLE 3
Ablation study of the effect of Sequential Generator and Dynamics Discriminator in *Head2Head++*.

Variations	FVD
w/o seq. G	74.42
w/o D_D	66.48
Full model	**57.46**

seen in the second row of Fig. 4, when the Mouth Discriminator is not included in GAN training, teeth details make the generated video appear less realistic, as opposed to the results in the first row, where the Generator was supervised by D_M network. Fig. 5 demonstrates the significance of conditioning the Generator on the eyes input sequence. Given that NMFCs do not adequately reflect eye gaze and blinking, conditioning solely on the NMFC input sequence does not provide substantial information to the Generator, for transferring the eye movements of the source to the target. Additionally, we quantitatively evaluate the significance of the mouth discriminator and the eye gaze conditioning. Table 2 presents the numeric measures obtained when assessing their effect. In the second column of Table 2, we report the **average pixel distance (APD)** within the **mouth area** under self-reenactment, between the generated and ground truth regions of interest, for all eight identities in our dataset. As can be seen, the numeric results agree with the qualitative ones, in Fig. 4. In the third column of Table 2, we assess the importance of **conditioning the Generator on the Eyes input** (Eyes Video sketch shown in Fig. 1). For that, we use the landmarker to extract eye landmarks, both from the ground truth and synthetic frames, created under self-reenactment. Then, we compute the **average eye landmarks $L2$-distance (AELD)**, between real and generated frames, across all eight identities and report the results in the Table 2. We observe that conditioning on the eyes helps to synthesise faces that better follow the eye movement and gaze appearing in the source video.

Lastly, in order to evaluate the importance of the Dynamics Discriminator D_D on the temporal consistency of generated frames, we use the Fréchet Video Distance (FVD) [55]. We also provide the FVD score for the non-sequential variation of Head2Head++ (w/o seq. G). The results of Table 3 indicate that our full model with a sequential Generator trained alongside a Dynamics Discriminator outperforms both variations by a significant margin.

4.5 Effect of Training Video Length

Fig. 7 displays the influence of training footage duration on the generative performance of the the Video Rendering Network. For that experiment, we trained five separate models per target identity. Each model was trained on frame sequences of different length: one, two, four, eight, up to sixteen thousand frames. We found that FID and MMD^2 scores, which indicate visual quality, are the metrics affected the most by the number of training frames. Furthermore, it is visible that there is still room for improvement in the performance of our Video Rendering Network, provided that even longer training videos are available. However, this would require additional footage with enough head pose variability, which in practice does not scale linearly with the number of training frames available.

Fig. 6. Significance of training video length. We show the same image generated with a GAN trained on 1K, 2K, 4K, 8K and 16K frames, as well as the ground truth (last image).

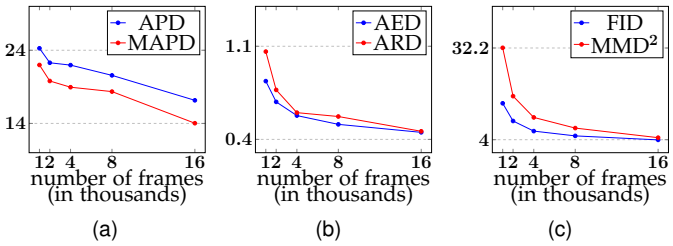


Fig. 7. The effect of the number of training frames on the quality of generated frames. All metrics were computed for the subset of four target identities (Macron, Sanchez, Trump, Von Der Leyen).

In Fig. 6, we demonstrate the impact of training video length on the visual quality of synthetic frames. We observe visually that as the number of available training frames increases, artifacts become less severe and eventually disappear. This behaviour comes in agreement with the decrease of evaluation metrics, shown in Fig. 7.

4.6 Pre-training on Face Forensics++ Dataset

We further explore how the generative performance of the Video Rendering Network can be improved in the case of limited video footage, for instance when training sequences are less than four thousand frames (about 2.5 minutes). To that end, we pre-trained *Head2Head++* on the large-scale *Face Forensics++* dataset [56], containing 1000 different identities. At this stage, the Generator learned to create a "average identity", since there is no mechanism to dictate which identity to be synthesised over this multi-person dataset. As a next step, we fine-tuned eight person-specific models, one for each target identity in our dataset, using a subset of their available footage (1K, 2K or 4K frames). In Table 4, we report a quantitative evaluation with and without pre-training on [56] data. As suggested by the results,

Fig. 8. The effect of pre-training the Video Rendering Network on Face Forensics++ dataset. The importance of pre-training is very prominent when a limited number of training frames (1K, 2K) is available.

TABLE 4
Effect of pre-training GAN on Face Forensics++ dataset, when training video footage is limited, namely less than 4K frames. All metrics are averaged across all eight target identities in our dataset. For all metrics, lower values indicate better quality or performance.

Number of train frames	APD	MAPD	AED	ARD	$\times 10^2$ FID	$\times 10^5$ MMD^2
1K	22.62	20.32	0.844	0.955°	16.01	39.59
w/ pre-train	**21.95**	**19.02**	**0.575**	**0.562°**	**15.93**	39.76
2K	20.41	18.27	0.719	0.737°	11.04	24.83
w/ pre-train	**20.17**	**17.84**	**0.526**	**0.516°**	11.22	24.93
4K	18.87	17.46	0.610	0.577°	7.41	15.08
w/ pre-train	18.90	**16.71**	**0.483**	**0.453°**	7.80	15.63

pre-training on *Face Forensics++* data appears advantageous, especially for shorter sequences (1K or 2K frames). This could be explained if we consider pre-training as a head start for the GAN, where the Generator is already capable of creating a realistic "average identity" and then it remains to learn a specific one. The same trend can be seen in Fig. 8. Especially for a training footage of 1K or 2K frames, the quality of generated frames is clearly inferior in comparison with samples created with pre-training.

4.7 Comparison with the State of the Art

We compare our *Head2Head++* system with the image-based method of *DVP* [4] and video-based *Head2Head* [15]. This is done by performing a full head reenactment experiment on the same source and target video footage. First, we trained our method and *Head2Head* on the target sequence, which was kindly provided by the authors of [4], and then we used the same source to drive synthesis. In Fig. 9 we display some

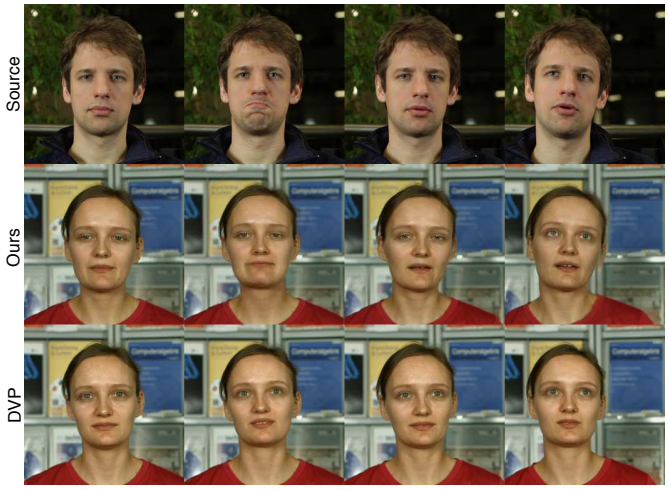


Fig. 9. Comparison of our method with *DVP*. In many cases our method outperforms *DVP*, especially in terms of head pose transferability.

TABLE 5
Quantitative comparison with *DVP* and *Head2Head* under full head reenactment. For all metrics, lower is better.

Method	AED	ARD	DAI	$\times 10^2$ FID	$\times 10^5$ MMD^2	AELD
DVP	2.089	2.54°	23.01	15.94	61.32	0.85
Head2Head	**1.426**	**0.70°**	9.56	39.35	186.11	0.72
Ours	1.509	0.83°	**8.86**	**12.42**	**39.15**	**0.71**

examples, in which our method outperforms *DVP* in terms of head pose and expression transferability.

For a quantitative comparison of the three methods, we performed 3D face reconstruction on the two synthetic videos and computed the average expression distance (AED) and average rotation distance (ARD) across frames, between the source video and the three generated videos. Then, we applied 3D face reconstruction on the target video and computed the distance between the average identities extracted from this real and each one of the three fake sequences. Finally, we used the FID and MMD^2 scores as a photo-realism metric, both computed between the fake videos and a set of frames from the target's original footage. For the eye region, where significant visual improvements have been made over *Head2Head*, we computed the average eye landmarks $L2$-distance (AELD), between source and generated videos.

The results presented in Table 5 suggest that our method outperforms DVP on every single metric and *Head2Head* in several. More specifically, data samples produced by *Head2Head++* exhibit significantly lower FID and MMD^2 scores, when compared to these created by *Head2Head++*. This can be explained by the fact that the network (ArcFace [53]) we employ to extract facial features for FID and MMD^2 is very sensitive to the shape of the eyes and thus imperfections and inconsistencies in the eyes region generated by *Head2Head* are reflected in both photo-realism metrics. Considering the AELD metric, the difference between the two methods is very small, which can be justified by the fact that the eye tracker used in *Head2Head* is quite reliable

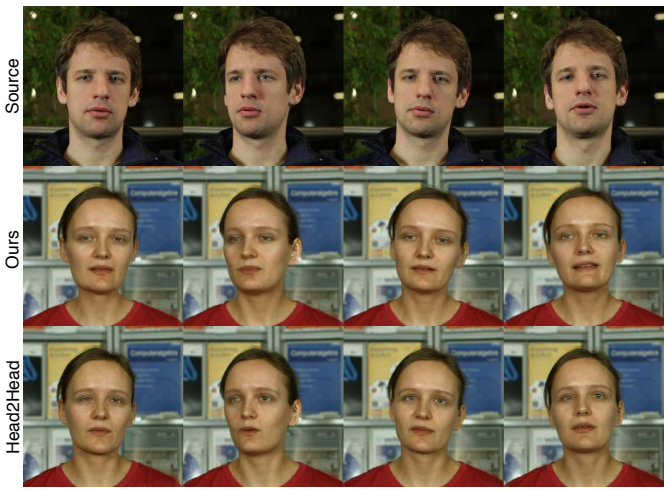


Fig. 10. Comparison of our *Head2Head++* method with *Head2Head* [15] on eye region synthesis. We synthesise more photo-realistic eyes in comparison to [15]. Please refer to the supplementary video for more examples.

TABLE 6
Quantitative comparison with *X2Face* and *FOMM*, using all eight target identities in our dataset, under self reenactment.

Method	APD	FID ($\times 10^2$)	MMD^2 ($\times 10^5$)
X2Face [2]	30.54	63.3	253.2
FOMM [40]	22.43	24.6	90.3
Ours	14.82	4.2	6.5

and the video mostly frontal, therefore differences with *Head2Head++* can be better understood in terms of photo-realism. In Fig. 10, we show that our new eye motion extraction method results in more reliable eye synthesis, compared to *Head2Head* [15], in terms of photo-realism and eye shape. We observe that the eye regions generated by *Head2Head* seem unnatural, while *Head2Head++* has successfully rendered photo-realistic eyes.

The AED, ARD and DAI metric values obtained for *Head2Head* seem to be slightly better than *Head2Head++*. We attribute these results to the fact that *Head2Head* is equipped with a video-based 3D reconstruction approach that relies on all video frames to generate the 3D faces, taking a few minutes compared to few milliseconds in the case of *Head2Head++*. Nonetheless, we strongly believe that performing 3D face recovery in nearly real-time outweighs the small disadvantage we noticed in terms of facial expression accuracy.

We further compare our method with the warping-based *X2Face* method [2] as well as the one-shot, image-based *First Order Motion Model for Image Animation (FOMM)* [40] system. As suggested visually in Fig. 11 and the quantitative results in Table 6, *Head2Head++* outperforms both few-shot methods by a large margin. Such a discrepancy can be explained by the fact that we train an individual person-specific model for each target identity, instead of relying on a few image samples of the subject. On the other hand, methods such as first order motion model capitalise on a single image to generate the target person, which results in

Fig. 11. Comparison with *X2Face* and *FOMM*. Our method exhibits better photo-realism. In many cases, the results of *X2Face* appear distorted, while *FOMM*'s performance drops significantly for poses distant from the one provided in the target image.

Fig. 12. Comparison with *vid2vid* conditioned on facial landmarks. Here, the identity preservation problem of landmarks is evident, as the head geometry of the source is reflected on the generated target.

the inherent ambiguity when the desired head pose is very different from the one displayed on that single reference image. Unavoidably, this makes the identity preservation problem more prominent in one-shot models.

In order to demonstrate the significance of the NMFC representation, we conducted a head reenactment experiment conditioning on landmarks. To that end, we trained *vid2vid* [12] on the task of landmark-to-RGB video translation. We demonstrate the results in Fig. 12. As can be seen, the facial shape of the source has been transferred along with the pose and expression to the generated target.

4.8 Automated Study

The progress recently made by generative deep learning methods is so impressive to the extent that manipulated videos (deepfakes) by such approaches are indistinguishable from the real ones. Consequently, a new research topic has emerged to tackle the detection of what is commonly known as 'deepfakes'. A recent attempt in that direction is the *Face Forensics++* [56]. The authors of [56] collected a dataset of 1000 YouTube videos and manipulated them with graphics-based [3], [57] and learning-based [7], [58],

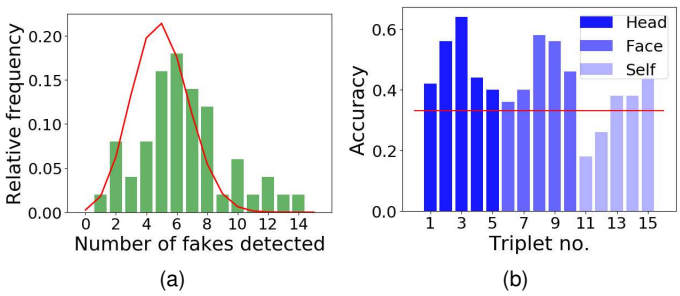


Fig. 13. User study results. (a) The predictive performance of participants against random selection (Binomial distribution on $n = 15$ independent Bernoulli trials with success probability $p = 1/3$). (b) The fake detection accuracy of participants on each triplet of video samples.

facial reenactment methods. With this dataset, a CNN was trained to outrank the performance of human observers in detecting manipulated videos. Such a network was trained on 1.8 million facially manipulated frames and reached a very high detection accuracy on the test split of their dataset (around 99%). We use their pre-trained best-performing network and fine-tune it on around 14K frames synthesised from eight different subjects by our *Head2Head++* approach and another 14K frames synthesised by *Head2Head*. While fine-tunning, we set the learning-rate to 0.0001 and batch-size to 32. Adam optimiser [51] was also used with the default parameters. We then test the fine-tuned network on 3.2K frames generated by *Head2Head++* and *Head2Head* methods (1.6K frames each) and not seen during training. The trained network reports an accuracy of 76% vs 80.2% of fake frames detection for *Head2Head++* and *Head2Head*, respectively. This indicates that the fine-tuned forgery detection network finds it easier to spot fake frames generated by Head2Head compared to Head2Head++. Moreover, the same network performs way better (around 99%) in detecting fake frames synthesised by any of the four head-reenactment methods tested in [56]. Thanks to our carefully designed *Head2Head++* framework, the trained forgery detection network finds it harder to identify our fake frames, which can be attributed to the increased realism of our results.

4.9 User Study

We further evaluate the photo-realism of our generated videos, by conducting a user study. For that, we synthesised five videos of 75 frames (3 seconds) each, using different reenactment methods (head, face and self reenactment). Then, we coupled each fake video with two real ones, with the same duration and same target identity, forming triplets of videos. Next, we asked 50 participants on MTurk to detect the fake sample of each triplet. We presented the video triplet to the participants, allowing them to watch them only once.

Our recorded results indicate a human fake detection accuracy of 43.1%, which demonstrates the strong photo-realism of samples created by our *Head2Head++* system. In Fig. 13a, we show the relative frequency of participants on the number of fake videos that were successfully detected. As can be seen, the majority managed to spot between 4 and 8 out of the 15 synthetic samples we displayed. The comparison of our statistics with a Binomial distribution (red curve) indicates that user study participants performed slightly better than random selection. Nonetheless, a significant percentage of them, around 18%, managed to detect 9 or more fake samples. Finally, Fig. 13b displays the predictive accuracy of participants on each video triplet. It is evident that self reenactment samples were indistinguishable from real ones, with just 32.8% predictive accuracy. Videos generated under face and head reenactment were identified a little more successfully, with an accuracy of 47.2% and 49.2% respectively.

5 Conclusion and Future Work

We proposed *Head2Head++*, a pipeline consisting of a novel 3D facial reconstruction system and a Video Rendering Network, able to perform full head reenactment from a source to a target identity. Our new 3D face recovery and eye tracking methods allow our system to operate in nearly real-time speeds. The extensive set of experiments we performed demonstrated the generative abilities of our system, as well as the significance of its components. Furthermore, qualitative and quantitative comparisons suggest that *Head2Head++* outperforms other state-of-the-art methods, in terms of photo-realism, expression and pose transferability, as well as target identity preservation. For future work, we aim to make the network trainable in a few-shot manner, reducing the training time and eliminating the need for learning a different model per person. Additionally, we plan to transfer the facial expressions from the source in a way that preserves the target's speaking style.

References

[1] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*

[2] O. Wiles, A. Koepke, and A. Zisserman, "X2face: A network for controlling face generation by using images, audio, and pose codes," in *ECCV*, 2018.

[3] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016.

[4] 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.

[5] M. Koujan*, M. Doukas*, A. Roussos, and S. Zafeiriou, "Reenact-net: Real-time full head reenactment," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2020, pp. 327–327.

[6] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM TOG*, 2015.

[7] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM TOG 2019*, 2019.

[8] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt, "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track," 2015.

[9] 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. [Online]. Available: http://arxiv.org/abs/1905.08233

[10] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *SIGGRAPH*, 2001.

[11] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormaehlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *CVPR*, 2014.

[12] 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.

[13] 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.

[14] 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.

[15] M. Koujan*, M. Doukas*, A. Roussos, and S. Zafeiriou, "Head2head: Video-based neural head synthesis," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2020, pp. 319–326.

[16] M. R. Koujan, A. Roussos, and S. Zafeiriou, "Deepfaceflow: In-the-wild dense 3d facial motion estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] M. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos, "Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2020, pp. 539–546.

[18] G. Giannakakis, M. Koujan, A. Roussos, and K. Marias, "Automatic stress detection evaluating models of facial action units," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, 2020, pp. 817–822.

[19] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *TPAMI*, 2014.

[20] W. A. Smith and E. R. Hancock, "Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics," *IJCV*, 2008.

[21] R. Garg, A. T. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *CVPR*, 2013.

[22] S. Graßhof, H. Ackermann, F. Kuhnke, J. Ostermann, and S. S. Brandt, "Projective structure from facial motion," in *IAPR MVA*, 2017.

[23] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999.

[24] N. Faggian, A. Paplinski, and J. Sherrah, "3d morphable model fitting from multiple views," in *FG 2018*, 2008.

[25] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1126–1135.

[26] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2549–2559.

[27] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355.

[28] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: Face model learning from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 812–10 822.

[29] Z. et al., "State of the art on monocular 3d face reconstruction, tracking, and applications," in *Computer Graphics Forum*, 2018.

[30] E. et al, "3d morphable face models–past, present and future," *arXiv preprint arXiv:1909.01815*, 2019.

[31] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in neural information processing systems*, 2015, pp. 2539–2547.

[32] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7083–7092.

[33] P. Goyal, Z. Hu, X. Liang, C. Wang, and E. P. Xing, "Nonparametric variational auto-encoders for hierarchical representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5094–5102.

[34] 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.

[35] A. Larsen, S. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning, PMLR.*, 2016.

[36] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Headon: Real-time reenactment of human portrait videos," *ACM Transactions on Graphics 2018 (TOG)*, 2018.

[37] 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)*.

[38] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[39] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2377–2386.

[40] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[41] E. Wood, T. Baltrusaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Gazedirector: Fully articulated eye gaze redirection in video," *Computer Graphics Forum*, vol. 37, 04 2017.

[42] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[43] J. Saragih, S. Lucey, and J. Cohn, "Real-time avatar animation from a single image," 04 2011, pp. 117 – 124.

[44] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[45] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 251–260.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[47] 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.

[48] 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.

[49] 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.

[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[51] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[52] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 6626–6637.

[53] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[54] D. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton, "Generate models and model criticism via optimized maximum mean discrepancy," in *ICLR*, 2017.

[55] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[56] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.

[57] Marek, "Faceswap," github.com/MarekKowalski/FaceSwap/, 09/2019.

[58] torzdf, "Deepfakes," github.com/deepfakes/faceswap, 09/2019.

Michail Christos Doukas is a PhD Student at Imperial College London and is currently doing his internship at Huawei UK Research Centre. He received the MSc degree in Computing from Imperial College London, in 2017. Prior to that, he has studied Electrical and Computer Engineering (MEng 2016) at the National Technical University of Athens (NTUA), Greece. His research interests are focused on Deep Learning and Computer Vision, including Generative Adversarial Neural Networks, Video-to-Video Translation, Visual Speech Synthesis, Face and Head Reenactment and Few-shot Learning.

Mohammad Rami Koujan is currently a PhD student at the University of Exeter (UK). Before his PhD, he did a 2-year Erasmus Mundus Joint MSc. course in Computer Vision and Robotics at Heriot-Watt University, University of Bourgogne and University of Girona and graduated with distinction. As an undergraduate student, he completed a Communications and Information engineering course with first class honors at Yarmouk Private University, Syria. His research interests lie mainly in 3D Computer Vision and Deep Learning techniques. He has been focusing on 3D facial shape modelling and analysis, non-rigid 3D facial motion capture and photo-realistic synthesis.

Viktoriia Sharmanska is a research fellow at Imperial College London, leading the project in 'Deep Understanding of Human Behaviour from Video Data: Action plus Emotion Approach', since October 2017. Prior to her current position, she was a visiting research fellow at the University of Sussex, UK, working on cross-modal and cross-dataset learning with privileged information. She got her MSc in Applied Mathematics from the Taras Shevchenko National University of Kyiv, Ukraine, and her PhD in Computer Vision and Machine Learning from the Institute of Science and Technology Austria. Her research interests include deep learning methods for understanding human behaviour from facial and bodily cues, algorithmic fairness, designing machine learning models that can overcome human and dataset collection biases.

Anastasios Roussos is a Principal Researcher (Associate Professor level) at the Foundation for Research and Technology - Hellas (FORTH), Greece. He is also an Honorary Senior Lecturer in the University of Exeter, UK. Prior to his current positions, he was a Lecturer in Computer Science at the University of Exeter and a Fellow of the Alan Turing Institute, UK. Before these positions, he was a postdoctoral researcher at Imperial College London, University College London and Queen Mary, University of London. He has studied Electrical and Computer Engineering (PhD 2010, Dipl-Ing 2005) at the National Technical University of Athens (NTUA), Greece. His research specialises in the fields of Computer Vision and Machine Learning.

Stefanos Zafeiriou is currently a Professor in Machine Learning and Computer Vision with the Department of Computing, Imperial College London, London, U.K, and an EPSRC Early Career Research Fellow. Between 2016-2020 he was also a Distinguishing Research Fellow with the University of Oulu under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011. He was the recipient of the President's Medal for Excellence in Research Supervision for 2016. His research specialises in machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behaviour analysis.

Head2Head++: Deep Facial Attributes Re-Targeting (Supplementary Material)

Michail Christos Doukas[1,4]*, Mohammad Rami Koujan[2,4]*, Viktoriia Sharmanska[1], Anastasios Roussos[3,4], and Stefanos Zafeiriou[1,4]

1 DenseFaceReg Network Details

The aim of this network is to accept an input image and estimates the per-vertex (x, y, z) coordinates reflecting the subject's facial geometry. ResNet50 [1] network architecture was selected and trained for this purpose after replacing the output fully-connected (**fc**) layer with a convolutional one $(3\times3, 512)$ and then a linear **fc** layer with $\sim 1.5k\times3$ neurons. This network was trained initially and separately from the rest of the framework on a 1000 videos from the **Face3DVid** dataset [2]. Adam optimiser was used [3] during the training with learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size 32.

2 Video Rendering Network Details

2.1 The architecture of the generator G

Let ck denote a 7×7 Convolution2D,BatchNorm,ReLU layer with k filters and unitary stride. dk represents a 3×3 Convolution,BatchNorm,ReLU layer with k filters, and stride 2. uk denotes a 3×3 transpose Convolution2D,BatchNorm,ReLU layer with k filters, and stride of $\frac{1}{2}$. Then, the encoding architecture is c64, d128, d256, d512, R512, R512, R512, R512 and the decoding architecture is R512, R512, R512, R512, R512, u256, u128, u64, conv, tanh.

2.2 The objective of the discriminators

The Image Discriminator D_I is optimised under the objective:

$$\begin{aligned}\mathcal{L}_{adv}^{D_I} =& \frac{1}{2}E_t[(D_I(\mathbf{X}_t, \mathbf{Y}_t) - 1)^2] \\ &+ \frac{1}{2}E_t[(D_I(\mathbf{X}_t, \tilde{\mathbf{Y}}_t))^2].\end{aligned} \tag{1}$$

The mouth discriminator D_M is optimised under a similar adversarial loss with $\mathcal{L}_{adv}^{D_I}$, which results by replacing the whole images in (1) with the cropped mouth areas.

$$\begin{aligned}\mathcal{L}_{adv}^{D_M} =& \frac{1}{2}E_t[(D_I(\mathbf{X}_t^m, \mathbf{Y}_t^m) - 1)^2] \\ &+ \frac{1}{2}E_t[(D_I(\mathbf{X}_t^m, \tilde{\mathbf{Y}}_t^m))^2].\end{aligned} \tag{2}$$

We note that we perform image cropping around the center of the mouth on the real $\mathbf{Y}_t$, fake $\tilde{\mathbf{Y}}_t$ and conditional input images (eyes and NMFC), ending up with images $\mathbf{X}_t^m, \mathbf{Y}_t^m, \tilde{\mathbf{Y}}_t^m$ of size 72×72. Finally, the dynamics discriminator is trained to minimise the adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv}^{D_D} =& \frac{1}{2}E_t[(D_D(\mathbf{W}_{t:t+1}, \mathbf{Y}_{t:t+2}) - 1)^2] \\ &+ \frac{1}{2}E_t[(D_D(\mathbf{W}_{t:t+1}, \tilde{\mathbf{Y}}_{t:t+2}))^2].\end{aligned} \tag{3}$$

3 Dataset Details

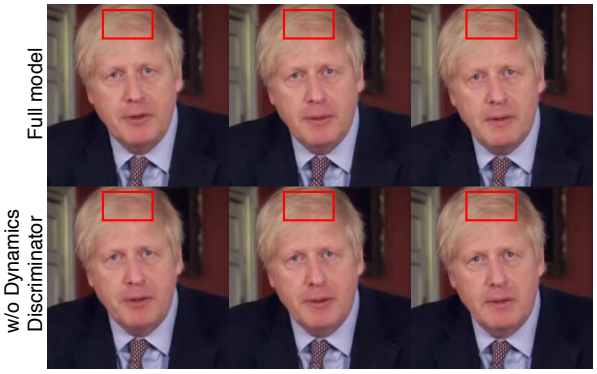
TABLE 1
Number of frames in the training and test split, for each one of the eight target videos in our dataset.

Target Identity	Number of test frames	Number of training frames
Lagarde	1948	10800
Macron	1498	16400
Von Der Leyen	1648	16600
Trump	3568	28800
Mitsotakis	1732	9200
Johnson	1904	14100
Sanchez	1998	16350
Tusk	998	8550

- * *denotes equal contribution*
- [1] *Department of Computing, Imperial College London, UK*
- [2] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK*
- [3] *Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), GR*
- [4] *FaceSoft.io, London, UK*

Manuscript received June 15, 2020; revised October 3, 2020

Fig. 1. Our dataset consists of eight different target identities, exhibiting various facial expressions and head poses.



(a) Hair consistency

(b) Background consistency

(c) Upper body consistency

Fig. 2. Significance of Dynamics Discriminator. Supervising the Generator by such a network increases the temporal coherence and smoothness across frames. Without the Dynamics Discriminator, various discontinuities might appear in the hair, background and upper body, which are even more prominent when watching the generated videos.

(a) Face reenactment

(b) Cycle reenactment

Fig. 3. More reenactment methods. During face reenactment (a), only the expressions (inner facial movements) are passed on to the target. Cycle reenactment (b), involves two consecutive full head reenactments, aiming at reconstructing the initial source.

4 Significance of Dynamics Discriminator

If Fig. 2, we illustrate the significance of the Dynamics Discriminators in the quality of generated frames, in terms of temporal coherence and continuity. In areas were no conditional information is available, such as hair, background and upper body, the Dynamics Discriminators helps to ensure smooth transitions. Please refer to the supplementary video for more.

5 Face and Cycle Reenactment Examples

Our *Head2Head++* method is also able to perform face reenactment, as depicted in Fig 3a. Under this reenactment setting, only the facial expression are passed on from the driving video to the target identity, while the head pose and eyes are obtained from the original target video.

Cycle reenactment (see **??**) is another interesting method for evaluating the performance of our system. It consists of two head reenactments, where the synthesised sequence of the first one is used as a driving video to for the second one. Ideally this should produce a reconstruction of the initial source sequence. This method can be used instead of simple self reenactment, for evaluating the performance of a reenactment system quantitatively, since it allows comparison with the ground truth.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[2] M. R. Koujan, A. Roussos, and S. Zafeiriou, "Deepfaceflow: In-the-wild dense 3d facial motion estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[3] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.