Free-HeadGAN: Neural Talking Head Synthesis With Explicit Gaze Control

Michail Christos Doukas, Evangoulos Ververas, Viktoriia Sharmanska, and Stefanos Zafeiriou, Member, IEEE

Abstract—We present Free-HeadGAN, a person-generic neural talking head synthesis system. We show that modeling faces with sparse 3D facial landmarks is sufficient for achieving state-of-the-art generative performance, without relying on strong statistical priors of the face, such as 3D Morphable Models. Apart from 3D pose and facial expressions, our method is capable of fully transferring the eye gaze, from a driving actor to a source identity. Our complete pipeline consists of three components: a canonical 3D key-point estimator that regresses 3D pose and expression-related deformations, a gaze estimation network and a generator that is built upon the architecture of HeadGAN. We further experiment with an extension of our generator to accommodate few-shot learning using an attention mechanism, in case multiple source images are available. Compared to recent methods for reenactment and motion transfer, our system achieves higher photo-realism combined with superior identity preservation, while offering explicit gaze control.

Index Terms—Canonical 3D key-points, gaze estimation, gaze redirection, neural talking head synthesis, pose editing, reenactment.

I. INTRODUCTION

GENERATING photo-realistic images or videos of human faces has lately become a very popular topic in Computer Vision, with a remarkable amount of research and discussion revolving around it. Aside from considerable advancements accomplished by the computer graphics community [1], [2], [3], deep neural networks and more specifically Generative Adversarial Neural Networks (GANs) [4] have made a significant contribution towards this direction. For instance, StyleGAN2 [5] has shown incredible results in unconditional face synthesis.

Neural talking head synthesis, here also referred to as reenactment, is the task where a reference (source) image or video is manipulated or re-animated to match with the facial expressions and head poses performed by a driving (target) actor. Reenactment systems have numerous applications in social media, teleconference, image and video editing, as well as virtual reality and games. They can be further used for facial video compression and reconstruction [6].

Learning-based reenactment methods are classified either as person-specific or person-generic. On the one hand, most person-specific approaches train neural networks to generate samples of a single identity [7], [8], [9]. Despite their impressive visual results, such networks are optimised using a long video of the particular source actor we wish to reenact and cannot generalise to other identities. On the other hand, person-generic systems [6], [10], [11], [12], [13], [14], [15], [16], [17] have the ability to adapt to the source identity during inference, given only a few reference images of the subject, even a single one. Although very promising, these models in general produce samples of lesser quality, when compared to their person-specific counterparts.

Another important aspect of talking head synthesis relates to face modeling. A common facial representation that is adopted in [10], [11], [12] are 2D sparse landmarks (e.g., 68 points). More recent methods, such as [13] and [18] use 3D landmarks instead. Among other benefits, the selection of a three-dimensional representation enables free-view control, as one can intervene in the head pose to be generated by rotating the landmarks. Other widely used facial representations are dense 3D facial meshes combined with statistical priors of the face, such as 3D Morphable Models (3DMMs) [19], [20], [21], [22], which have been adopted for full head reenactment [7], [8], [17]. This modeling enables to disentangle shape (identity) from expression-related deformations, and therefore helps to tackle the identity preservation problem in cross-identity motion transfer, which is neither straightforward nor trivial with sparse landmarks. In a quite different line of work, Siarohin et al. [15], [16] attempt to solve the more general problem of motion transfer for arbitrary objects, proposing a model-free method that is based on unsupervised 2D key-point detection. Following up, Wang et al. [6] present a model that focuses on faces and learns to detect canonical 3D key-points without supervision, as a way to disengage identity from expression and pose.

Given that they deal with motion transfer, most recent systems are flow-based [6], [13], [14], [15], [16], [17], [18], [23], meaning that they make use of optical flow to warp the source image(s) or visual features into the target facial pose and expression, as a first step prior to image synthesis. In fact, the aforementioned model-free approaches [6], [15], [16] are able
to learn important key-points for motion by approximating the optical flow around the key-points. For instance, [16] and [6] adopt a first order approximation of motion using local affine transformations. On the contrary, methods such as [13], [14], [17] learn a dense optical flow for every pixel location, without resorting to approximations.

In this work, we do not use a first order approximation of motion, but rather use a module that learns to predict dense optical flow, similarly to HeadGAN [17]. We propose Free-HeadGAN, a system based on the architecture of [17], but “free” from statistical priors of faces like 3DMMs. Instead, we adopt a neural network that regresses 3D key-points along with head pose and expression deformations from the driving frames, and also deals with the identity preservation problem of reenactment. Our choice to replace the 3DMM of HeadGAN with this network stems from the fact that: a) its feed-forward operation is considerably faster than 3DMM fitting, b) our approach leads to slightly better image quality and identity preservation, c) HeadGAN relies on a complicated 3D face rendering operation for creating conditional inputs [24], d) the LSFM [21] 3DMM adopted in HeadGAN is not publicly available, making its reproducibility difficult. Our new approach is closely related to [6], however, our proposed network does not predict canonical key-points directly. Instead, we compute them based on estimated 3D points after removing pose and expression-related information. Moreover, we supervise our key-point estimator with pseudo-ground truth 3D facial landmarks extracted with RetinaFace [25]. In this way, we ensure that key-points are placed on meaningful parts of the face, avoiding cases where no key-points are assigned to important regions such as the eyes and mouth, as it happens with unsupervised models [6], [15], [16]. Finally, we propose a gaze estimation network that makes explicit gaze transfer possible. Our contributions can be summarised in the following:

- We release HeadGAN [17] from 3DMM priors of the face, by proposing a network that estimates 3D key-points, pose and expression. As suggested by our experiments, we achieve comparable and in many cases better results.
- This network performs disentanglement of identity from expression and pose, based on canonical key-points.
- We propose a second network that regresses 3D meshes of the eyes. We exploit these meshes to obtain the direction of gaze, which is then used to condition image synthesis.

To the best of our knowledge, we deliver the first person-agnostic reenactment system with explicit gaze control.

- We show a N-shot extension of our generative framework.

II. BACKGROUND

A. GANs and Image Synthesis

Since their introduction, GANs [4] have been extensively used both for unconditional and conditional [26] data synthesis. They have been successfully applied to various computer vision tasks, such as image-to-image [27], [28], [29], [30], video-to-video [11], [31] and audio-to-image translation [32], [33]. In this work, we focus on the problem of conditional image synthesis and propose a GAN-based system for talking head synthesis, which achieves highly photo-realistic results and improved identity preservation compared to recent state-of-the-art approaches.

B. Talking Head Synthesis

1) Face Reenactment: Most of the early attempts to reenact human faces focus on the task of face reenactment, which aims to transfer the facial expressions performed by a driving actor to the face of a source video. The majority of face reenactment methods, i.e., Face2Face [2], manipulate the source footage, by re-writing only the facial region of the source video stream, while keeping the remaining parts (e.g., hair, body, background) unchanged [34], [35]. Bringing Portraits to Life [36] has been a first exception to these systems, which is able to slightly animate the entire head with the application of 2D warping.

2) Full Head Reenactment: Deep Video Portraits (DVP) [7] is allegedly one of the earliest learning-based head reenactment methods, as it utilises an image translation network and manages to fully transfer the head motions of the driver to the source, including eye gaze. In a more recent work, Head2Head [8], [9] adopts a sequential video-based translation network that performs full head reenactment while taking into consideration the temporal dynamics of talking faces. Both aforementioned methods rely on strong statistical priors of faces with 3D information, such as 3DMMs [19], [20], [21], [22]. Despite their
remarkable results, these person-specific models are optimised on long videos of the source identity and cannot generate new identities without re-training.

On the other hand, person-generic models are able to adapt their generative process on the appearance of new source identities, even if they are not introduced to them during training. Numerous identity-agnostic talking head synthesis systems are 2D-based [10], [11], [12], [13], as they condition image synthesis on 2D facial landmarks. Zakharov et al. [10] introduce a few-shot framework that is composed of an identity embedding network and a generator with AdaIN [37] layers for receiving the embedding vectors. They propose a training strategy consisting of a meta-learning stage which involves optimisation on a multi-person dataset, followed by fine-tuning on a few images of a new and unseen face. The video-based system [11], namely few-shot vid2vid, employs a sequential generator equipped with dynamic SPADE [30] layers, enabling to adapt synthesis on the appearance of new reference images. Zakharov et al. [12] propose a SPADE-based system that achieves real-time one-shot talking head synthesis on mobile phones, focusing on the foreground as it disregards background information. In contrast to 2D face modeling, 3D-based approaches such as Warp-Guided GANs [18] and MarioNETte [13] depend on 3D landmarks to represent faces. As opposed to the preceding works, MarioNETte [13] applies a PCA-based disentanglement on identity and expression with 3D facial landmarks, in order to alleviate the identity mismatch problem during reenactment. HeadGAN [17] is among the first one-shot talking head synthesis systems that relies on identity and expression 3DMMs for the disentanglement of expression and identity parameters in dense 3D facial shapes. HeadGAN yields image samples of unprecedented quality in reenactment, while maintaining the identity of the source and exhibiting free-view control. A very recent work, namely MegaPortraits [23], explores the idea of representing the appearance of the faces as a 3D volume, and combines it with latent motion representations.

There is a considerable amount of research involving person-generic model-free head synthesis [6], [14], [15], [16]. X2Face [14] is one of the pioneering approaches on the problem, which does not rely on any 2D or 3D priors. It is able to animate faces driven by multiple modalities, such as images, audio, and pose codes. In order to address motion transfer for arbitrary objects, Siarohin et al. [15] propose Monkey-Net, a framework composed of a 2D key-point detector which is jointly trained with a motion prediction and motion transfer network. Key-points are learned directly from data in an unsupervised way, while being assigned to meaningful regions of the objects, in order to enable reliable optical flow approximation for motion transfer. In the follow-up work, First Order Motion Model for Image Animation (FOMM) [16] learns a first order approximation of optical flow, based again on learnable 2D key-points, yielding very promising results. However, during cross-identity motion transfer it uses relative key-points to preserve the identity of the source, which makes the assumption that the face in the first driving frame is in the same pose with the source face. In their recent work, Wang et al. [6] apply first order motion approximation on 3D facial key-points. This enables their so-called face-vid2vid model to perform free-view reenactment, as they can manipulate pose by rotating the 3D points. They further devise two networks, one for predicting canonical key-points and one that estimates pose and expression deformations, which enable to tackle the identity preservation problem. As opposed to Wang et al. [6], we do not regress canonical key-points directly, but rather obtain them by removing the estimated head pose and expression deformations from regular 3D key-points, all of which are predicted using a single network.

In Table I we present a summary of the key features and design choices of recent state-of-the-art systems for reenactment and motion transfer, as discussed in this section.

### C. Gaze Estimation and Gaze Redirection

Gaze estimation is the problem of predicting the direction that someone is looking at, based on input images or videos. Ever since [38] employed CNNs to tackle the task, appearance-based methods have been the default approach. Multiple methods attempt to recover geometric features of eyes and use them to infer gaze [39], [40], [41]. Others focus on adapting generic gaze estimation networks to specific test domains, aiming to produce person-specific models [42], [43], [44] or adapting to
unseen image domains [45], [46]. Recently, significant progress has been made to reduce the labeled data required to build effective gaze estimation systems, by proposing to learn gaze in unsupervised or weakly-supervised settings [47], [48], [49]. In this work, we aim to implicitly recover gaze from the dense geometry of the eyes, and employ it for driving our image generator.

Related to our approach are also methods that perform gaze redirection. DeepWarp [50] is one of the earliest works on this field, which uses a deep network to perform coarse-to-fine image warping. GazeDirector [51] employs a 3D model of the eyes and renders the compositing 3D eyeballs onto the output image in a photo-realistic way. He et al. [43] present a GAN-based architecture and enforce cycle consistency to supervise the gaze redirection process. Unlike the aforementioned methods that manipulate only the eye region, the method of Zheng et al. [52] operates on the entire face, although it struggles with appearance preservation. Unlike [50] and [53] that rely on gaze datasets captured under controlled conditions, our system learns to manipulate source images from in-the-wild data [54], and offers control over the gaze, head pose and facial expression.

III. METHODOLOGY

Our proposed talking head synthesis system consists of three discrete networks: a network for inferring canonical key-points (Section III-A), a gaze estimation network (Section III-B) and an image generator (Section III-C). Each component of Free-HeadGAN is trained separately on its individual task.

A. Computation of Canonical Key-Points

Given a facial image, our goal here is to extract a set of key-points in a canonical space, which are independent both from the head pose and facial expressions of the subject. This representation merely depends on the geometry of the input face and therefore encodes only identity-related information. To that end, we propose a neural network $E_{can}$ that learns to a) regress sparse 3D facial key-points $p = \{p^{(k)}\}_{k=1,...,K}$, with $p^{(k)} \in [-1, 1]^3$, b) estimate head pose, as an affine transformation $T = \{s, R, t\}$ and c) compute a 3D vector $d^{(k)}$ for each key-point $k = 1, \ldots, K$, which models the non-linearity of deformations caused by facial expressions. Please note that $s$ corresponds to head scale, $R$ is a rotation matrix and $t$ is the translation. The architecture of $E_{can}$ follows that of the head pose and expression deformation estimator in [6], with the addition of one more affine output layer that predicts $K$ points $p$.

The canonical key-points are obtained by first estimating the 3D points and then subtracting the expression deformations and removing translation, rotation, and scale

$$
\hat{p}^{(k)} = \frac{1}{s}R^{-1}(p^{(k)} - d^{(k)} - t), \quad k = 1, \ldots, K. \quad (1)
$$

We further define the inverse operation, which brings canonical points back to the original 3D space, by adding scale, rotation, translation as well as the expression deformation

$$
p^{(k)} = sRp^{(k)} + t + d^{(k)}, \quad k = 1, \ldots, K. \quad (2)
$$

In order to train network $E_{can}$, we are given a source and a target image pair $(y_{src}, y_{tgt})$, which depict the same person performing a different and random pose and expression. We use $E_{can}$ to predict 3D key-points as well as pose transformations and expression perturbations from each image, which gives $\{p_{src}, T_{src}, d_{src}\}$ and $\{p_{tgt}, T_{tgt}, d_{tgt}\}$. After that, we apply (1) to get the corresponding canonical points $\hat{p}_{src}$ and $\hat{p}_{tgt}$, as shown in Fig. 2. Then, based on (2) we aim to recover the target key-points, by transforming the canonical representation of the source using the target pose and expression, and vice versa, which results in the reconstructed points $\tilde{p}_{tgt}$ and $\tilde{p}_{src}$.

We optimise $E_{can}$ in a supervised manner using $K = 68$ 3D facial landmarks estimated by a pre-trained RetinaFace [25] network. That is, given the pseudo-ground truth source and target landmarks $(p^{*}_{src}, p^{*}_{tgt})$ extracted with RetinaFace, we minimise the distance

$$
L_p = ||p_{src} - p^{*}_{src}||_2^2 + ||p_{tgt} - p^{*}_{tgt}||_2^2, \quad (3)
$$

which forces $E_{can}$ to predict accurate facial key-points. We noticed that learning to estimate the 3D key-points helps $E_{can}$ to predict the perturbations caused due to expressions more efficiently. We further minimise the key-point reconstruction distance

$$
L_{rec} = ||\tilde{p}_{src} - p^{*}_{src}||_2^2 + ||\tilde{p}_{tgt} - p^{*}_{tgt}||_2^2, \quad (4)
$$

which forces $E_{can}$ to learn the affine pose transformations as well as the deformation vectors, in an effort to recover the source and target key-points. In order to assist $E_{can}$, distinguish between rigid and non-rigid transformations, we penalise the error between the predicted target head rotation $R_{tgt}$ and the ground truth $R_{tgt}^*$ (here expressed as Euler angles). Furthermore, a regularisation term on the expression deformation vectors ensures that key-point perturbations due to expressions are kept small, as we want to avoid encoding identity-specific details in these vectors

$$
L_R = ||R_{tgt} - R_{tgt}^*||_2^2, \quad L_d = ||d_{src}||_2^2 + ||d_{tgt}||_2^2. \quad (5)
$$
Combining the loss terms defined above, the overall objective function for network $E_{can}$ is given as

$$L_{E_{can}} = \lambda_p L_p + \lambda_{rec} L_{rec} + \lambda_R L_R + \lambda_d L_d,$$

with the hyper-parameters set to the following values: $\lambda_p = \lambda_{rec} = 200, \lambda_R = 2$ and $\lambda_d = 5$.

### B. Gaze Estimation

The gaze estimation network $E_{gaze}$ learns to predict the 3D orientation of the eyes from an input image. Particularly, our aim is to recover a 3-dimensional vector that represents gaze direction and use it to condition the process of image synthesis during reenactment. To that end, we adopt mesh regression as our approach to gaze estimation, meaning that we train $E_{gaze}$ to predict 3D meshes of the eyes instead of 3D gaze directions. This approach is based on the fact that estimating the dense geometry of the object instead of a few parameters, has been shown to benefit recent face and body tracking systems [55], [56], [57]. In our experimental section we present an ablation study that further justifies our design choice to estimate a 3D mesh of the eyes, rather than a 3D gaze vector.

Given that most datasets for gaze estimation provide gaze directions as 3D vectors, angles or points on screen [38], [58], [59], we create training and validation data compatible with our mesh regression approach by first defining a template $T_{eye}$ of the 3D eye mesh. This mesh template consists of $N^{eye}_{v} = 481$ vertices and $N^{eye}_{e} = 928$ triangles, as shown in Fig. 3. We fit $T_{eye}$ on the available eye images, based on 2D sparse landmarks around the iris contour and the available gaze labels. To obtain the iris landmarks we employ the network from [39]. We perform fitting by first rotating the 3D eye template according to the ground truth gaze direction and then translate and scale it according to the 2D iris landmarks. In this way, we obtain a pseudo-ground truth 3D mesh for each eye image.

For our gaze estimation network $E_{gaze}$ we adopt a simple architecture, consisting of a ResNet-34 backbone and a fully connected layer that takes an eye image and outputs $v$, which is a vector of $3N^{eye}_{v}$ real values representing the coordinates of the vertices that constitute the 3D mesh of the eye. These values lie in a normalized space $[-1, 1]$. We optimize $E_{gaze}$ in a supervised fashion, based on the 3D eye meshes we fitted on the available gaze estimation data. To that end, we minimise the distance

$$L_v = ||v - v^*||_1,$$

between the predicted coordinates $v$ and the corresponding pseudo-ground truth values $v^*$. Additionally, to maintain a feasible eye shape, we minimize the distance

$$L_e = ||e - e^*||_1,$$

between the lengths of the edges $e$ computed from $v$ and the edges $e^*$ computed from $v^*$. We define the length of an edge to be the euclidean distance between two vertices that belong to the same triangle, according to $T_{eye}$. Lastly, we improve the accuracy of gaze estimation by employing the gaze loss

$$L_g = (180/\pi) \arccos(g^T g^*),$$

where $g$ and $g^*$ are normalized 3D gaze vectors, calculated from the predicted and pseudo-ground truth eye meshes respectively. Combining the losses above, the overall objective function for $E_{gaze}$ is given as

$$L_{E_{gaze}} = \lambda_v L_v + \lambda_e L_e + \lambda_g L_g,$$

with the hyper-parameters set to the following values: $\lambda_v = \lambda_e = 0.1$ and $\lambda_g = 1$.

Considering that the predicted 3D eye meshes encapsulate person-specific information, such as eyeball size, for the purposes of reenactment we utilise the 3D mesh merely as an intermediate representation. That is, we use it as a means to recover the 3D gaze vector $g = (g_x, g_y, g_z)$, as we do not "feed" the 3D mesh directly into our generative network. We give more details about how we condition image synthesis on gaze information in the next section.

### C. Image Synthesis

We perform talking head synthesis with the assistance of an image-to-image translation network $G$, which is based on the generator of HeadGAN [17], without AdaIN [37] layers as we do not process audio features. Given a source image $y_{src}$ and a target frame $y_{tgt}$, first we use the networks $E_{can}$ and $E_{gaze}$ to compute the facial keypoints $p_{src}$, $p_{tgt}$ and gaze vectors $g_{src}$ and $g_{tgt}$. Then, we use the key-points to draw 2D sketches of the source and target faces, denoted as $x_{src}$ and $x_{tgt}$. Gaze vectors are coded in the sketches as RGB colours, within the areas defined by the key-points that belong to the eyes, as seen in Fig. 1. The two sketches, together with the source image, serve as inputs to network $G$, which learns to generate a photo-realistic image $\hat{y}$, according to the target head pose and facial expression.

In more detail, as in [17], our network $G$ is comprised of two modules: a flow network and a rendering network. First, the flow network extracts visual feature maps from the source image and the corresponding source sketch in multiple spatial resolutions: $\{h^{(i)}\}_{i=1,\ldots,L}$, $L = 3$. Then, the target sketch is injected into...
the network through SPADE [30] layers, as modulation input, and guides the prediction of an optical flow $w$, which warps the source image to the target expression and pose. In addition, we perform warping the visual feature maps, in order to align them spatially with the desired expression and pose. In practice, we utilize the backward optical flow warping operator from FlowNet 2.0 [60]. Subsequently, the rendering network passes the target sketch through an encoder and combines the resulting feature map with the warped visual features $\{h^{(i)}\}_{i=1,\ldots,L}$ and warped image $\hat{y} = w(y_{src})$, which enter the network through SPADE [30] layers, as modulation inputs. Up-sampling is performed with PixelShuffle [61] layers. The final output is a photo-realistic image $\tilde{y}$ of the source identity, imitating the facial expressions and head pose shown in the target image. Please refer to Fig. 4 for an illustration of our generative model.

**N-Shot Extension.** We further extend our method to enable few-shot learning, in cases where more than one source images are available. For that, we propose an optional attention mechanism, with the addition of one more output layer to the flow network, which now learns to compute a set of 2D weights $m \in \mathbb{R}^{H \times W}$, alongside the optical flow field. Given $M$ source frames $\{y_j\}_{j=1,\ldots,M}$, we pass each one through the flow network, gaining flows $\{w_j\}_{j=1,\ldots,M}$ and weights $\{m_j\}_{j=1,\ldots,M}$. Next, the warped image is computed with the assistance of a Softmax function

$$\tilde{y} = \frac{\sum_j^M \exp(m_j)w_j(y_j)}{\sum_j^M \exp(m_j)},$$

and the warped features are given as

$$\tilde{h}^{(i)} = \frac{\sum_j^M \exp(m_j)w_j(h^{(i)}_j)}{\sum_j^M \exp(m_j)}, \quad i = 1,\ldots,L.$$

**Training.** The flow estimation and rendering networks that make $G$ are trained jointly, on the task of self-reenactment (reconstruction). To that end, we sample the source image from the frames of the target video to obtain image pairs that belong to the same person and scene. In this case, the generated image $\hat{y}$ should match the target frame $y_{tgt}$ that serves as ground truth. We optimise the generator $G$ by minimising the distance between

Fig. 5 shows a visual example of optical flows and weights.
D. Free-HeadGAN Inference

During inference in cross-identity reenactment, the source and target images belong to different identities. In order to adapt the target key-points to the facial shape of the source identity, we use $E_{can}$ to regress 3D key-points, head pose and expression deformations from $y_{src}$, and then evaluate Equation (1) to obtain the canonical key-points $\tilde{\mathbf{p}}_{src}$. At the same time, we estimate the target pose and expression $T_{tgt}, d_{tgt}$ from $y_{tgt}$. The adapted target key-points $\mathbf{p}_{adv}$ are obtained with the application of Equation (2), which transforms the source canonical key-points using the estimated target pose and expression. In this way, we remove the target identity-related information from key-points and inject the source one, which helps to overcome to problem of identity mismatching. In parallel, we estimate eye gaze from both images with $E_{gaze}$ network and finally draw the sketches that serve as input to generator $G$, which hallucinates the output image $\hat{y}$. For a visual inspection of our proposed pipeline during reenactment please refer to Fig. 6.

IV. Experiments

A. Datasets and Training

We train each component of Free-HeadGAN independently. Both the network $E_{can}$ that infers canonical key-points and the generative network $G$ (with discriminators $D_I, D_M$) are optimised on VoxCeleb [54] video dataset, which contains over one hundred thousand clips and more than one thousand identities, at a $256 \times 256$ resolution. We keep the original train and test split. As a pre-processing step, we prepare three-dimensional 68 landmarks for each frame of the training split with RetinaFace [55], which serves as pseudo-ground truth annotations. We note that RetinaFace has been trained on WIDERFACE dataset [65]. For the extraction of ground truth head pose, we employ a least-squares solver to determine the transformation between dense 3D points regressed with RetinaFace and a fixed template that represents the frontal pose. The aforementioned annotations are used only for training. During inference, our method requires only the source and target images.

For the optimisation of network $E_{can}$, we use ADAM solver [66] with $\beta_1 = 0.5, \beta_2 = 0.999$ and learning rate $\eta = 0.0002$. We use the same optimiser and hyper-parameters for training $G$ and its adversaries $D_I, D_M$. All models are optimised for 5 epochs on the entire VoxCeleb database. We extend Free-HeadGAN to accommodate N-shot learning, by first training in one-shot without predicting attention weights. Then, we add the weight layer and fine-tune our GAN for one more epoch using two source images (2-shot), while freezing the parameters of the flow network.

We optimize network $E_{gaze}$ on a combination of recent gaze estimation datasets, aiming to include variation from multiple image domains. Particularly we employ ETH-XGaze [67] which includes large variation in gaze and face pose and consists of 756 thousand frames from 80 subjects, Gaze360 [58] which is captured both indoors and outdoors and includes 127 thousand

...
training sequences from 365 subjects and MPIIGaze [38] which provides over 213 thousand frames of 15 subjects, captured with laptop cameras. We employ the default training and validation sets from Gaze360 and MPIIGaze, while we perform a manual split for ETH-XGaze, since gaze labels for its test data are not available. Lastly, we use ADAM [66] solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\eta = 0.0001$.

### B. Evaluation Metrics

We evaluate the reconstruction capabilities of our method quantitatively, using L1 distance (L1), Peak signal-to-noise ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS) [68]. All three metrics measure the distance, between the synthesised and ground truth target frames. L1 distance is computed across RGB channels that are in the range [0, 255]. PSNR is the ratio between the maximum possible power of a signal and the power of noise that affects its correctness. It is defined as $20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10} MSE$, where $MAX_I = 255$ and $MSE$ denotes the mean squared distance. LPIPS [68] is another widely used metric to measure the fidelity of reconstruction, which uses a pre-trained AlexNet model for the extraction of a feature maps. The similarity score between two images is calculated as the distance of their visual features.

Furthermore, we assess the photo-realism of generated images with Fréchet Inception Distance (FID) [69] and Fréchet Video Distance (FVD) [70], as we handle video data and it is crucial to measure the performance of models considering the temporal coherence among frames.

The identity preservation in the synthesised data is calculated with Cosine Similarity (CSIM) between embedding vectors from the target and corresponding generated images. All embeddings are computed with the assistance of ArcFace [71]. For reenactment, where we have no access to ground truth data, we extract the embedding from the source image(s) and compare it with the embeddings coming from generated data. This leads to lower CSIM values, as the source and target pose do not usually match and ArcFace’s output is slightly affected by pose variations.

We use Action Units Hamming distance (AU-H) to measure the expression transferability of models. We run OpenFace [72] with [73] on target and synthetic images for the detection of Action Units (AU). OpenFace recognises whether or not a set of AUs is present in a facial image, with the prediction of an AU boolean vector. We measure the Hamming distance between boolean vectors that are extracted from the corresponding target and synthetic data.

We measure the correctness of pose transfer with Average Rotation Distance (ARD), which is the $l_1$-distance between Euler angles that come from the head pose of target and generated frames, in degrees. We compute pose with the assistance of dense 3D facial points (around 1 K) extracted with RetinaFace, different from the ones we supervised $E_{can}$.

Finally, we evaluate gaze transfer with Average Gaze Distance (AGD). For that we employ a third-party gaze estimator and more specifically Gaze360 [58]. We compute the gaze vector $g_{tgt}$ from the target frame, and the vector $\bar{g}$ from the corresponding synthetic image. Then, we compute the angle between vectors as $\frac{180}{\pi} \arccos \left( \frac{\bar{g}^\top g_{tgt}}{||\bar{g}|| ||g_{tgt}||} \right)$ and average across frames to obtain AGD in degrees.

### C. Comparison With State-of-the-Art

We compare our one-shot Free-HeadGAN system both numerically and visually with state-of-the-art methods under two setups: a) same-identity reconstruction or self-reenactment, where the source and target identities coincide, and b) cross-identity motion transfer or reenactment, in which source and target identities are different. Please refer to Fig. 7 for a visualisation of the two tasks. More specifically, we compare our approach with X2Face [14], few-shot vid2vid (fs-vid2vid) [11], Bi-layer Neural Avatars (Bi-layer) [12], First Order Motion Model (FOMM) [16] and HeadGAN [17]. Please note that we tested two variations of FOMM, one with absolute (FOMM-abs) and one with relative key-point coordinates (FOMM-rel), which is the default setting for cross-identity motion transfer. For X2Face [14], FOMM [16] and HeadGAN [17], we use the models provided by their authors, all trained on VoxCeleb [54] dataset. For Bi-layer method, we used the network parameters provided by its authors, trained on VoxCeleb2 dataset* [74]. We trained fs-vid2vid [11] using the official open source implementation, as pre-trained models are available. Given that the source code for MarioNETte [13], Warp-guided GANs [18] and face-vid2vid [6] is not publicly available, we were not able to measure Free-HeadGAN’s performance against these systems.

First, we present a quantitative comparison of our proposed method with the aforementioned baselines. For the task of self-reenactment, we reconstructed the entire test set of VoxCeleb. For cross-identity reenactment, we employed the same 15 pairs of driving videos and reference images used in HeadGAN [17]. These pairs were formed by first sampling 15 random videos from the test split of VoxCeleb dataset. These constitute the driving videos of the experiment. After that, 15 more videos with different identities were sampled and a random frame from
TABLE II
NUMERICAL COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TASKS OF SELF-REENACTMENT (SAME-IDENTITY RECONSTRUCTION) AND REENACTMENT (CROSS-IDENTITY MOTION TRANSFER) FOR VOXCELEB [54] TEST SET

| Method         | Self-reenactment | Reenactment |
|----------------|------------------|-------------|
|                | L1  | PSNR | LPIPS | FID  | FVD  | CSIM | FID  | CSIM | ARD  | AU-H | AGD  |
|----------------|-----|------|-------|------|------|------|------|------|------|------|------|
| X2Face [14]    | 13.49 | 20.69 | 0.250 | 130.2 | 697 | 0.600 | 122.1 | 0.520 | 4.39 | 0.346 | 21.9 |
| fs-vid2vid [11]| 17.15 | 18.32 | 0.197 | 62.8  | 471 | 0.542 | -    | -    | -    | -    | -    |
| Bi-layer* [12] | 12.18 | 20.19 | 0.152 | 92.2  | 394 | 0.590 | 172.8 | 0.563 | 1.01 | 0.296 | 16.3 |
| FOMM-abs [16]  | 12.34 | 20.93 | 0.153 | 64.9  | 338 | 0.754 | 100.7 | 0.587 | 1.46 | 0.298 | 13.6 |
| FOMM-rel [16]  | -   | -    | -     | -    | -   | -    | 63.7  | 0.765 | 12.53 | 0.400 | 21.4 |
| HeadGAN [17]   | 13.32 | 21.14 | 0.112 | 36.1  | 254 | 0.807 | 58.0  | 0.688 | 1.35 | 0.326 | 16.7 |
| Free-HeadGAN   | 9.96  | 22.16 | 0.100 | 35.4  | 248 | 0.810 | 53.9  | 0.789 | 1.26 | 0.351 | 13.1 |

Fig. 8. Visual comparison of our method with baselines on the task of self-reenactment.

In Table II, our method creates superior samples with respect to image quality and reconstruction fidelity, both in same-identity reconstruction and cross-identity motion transfer. In addition, our approach preserves the identity of the source better in all experiments. For expression transferability, we observe that Free-HeadGAN is left slightly behind from [12] and [17]. This is attributed to the following two reasons: First, the pseudo-ground truth 3D landmark annotations have been extracted with RetinaFace [25], which is not a model focusing exclusively on landmark prediction and occasionally misses fine details. Second, during reenactment the adaptation of target key-points to the source identity with network $E_{can}$ relies on the regression of expression deformations, which also comes with small inaccuracies. However, this is a compromise we make, as we supervise network $E_{can}$ with the 3D facial key-points of RetinaFace and we pay particular attention to the identity preservation of the source, through the canonical key-point space.

In Figs. 8 and 9 we show examples from a visual comparison of Free-HeadGAN with the baselines, on the task of reconstruction and reenactment respectively. As can be seen, the qualitative results confirm our numerical analysis. More specifically, our
proposed system outperforms all state-of-the-art methods in terms of reconstruction fidelity, photo-realism, identity conservation and gaze transfer. In addition, Free-HeadGAN outperforms HeadGAN [17] slightly in terms of image quality, which is also reflected in the metrics. We attribute this incremental improvement partially on the rectification of the eyes region. Our method appears to be significantly more reliable on gaze transfer than HeadGAN, thanks to our gaze prediction network and the explicit gaze control during synthesis. Nonetheless, we believe that gaze correction does not account for the total improvement of the results. Since PNCC images [24] encode considerably more information about the shape of the face compared to sketches of landmarks, our presumption is that HeadGAN’s generator takes over the burden to disregard the occasional inaccuracies of the PNCC representation with regards to facial shape. On the contrary, the simplicity of sketches allows our generator to learn the structural details solely from the reference image. We urge the reader to refer to our supplementary video for a better inspection of our results, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3253243.

Finally, the feed-forward operation of our network $E_{can}$ is considerably faster than the 3DMM fitting algorithm and the subsequent 3D face rendering step of HeadGAN [17]. More specifically, 3DMM fitting requires on average 38.8 msecs per frame, 3D face rendering takes 26.9 msecs, while $E_{can}$ runs at 8.3 msecs per frame. The reported runtimes were measured on a NVIDIA Tesla V100 32 GB GPU. This suggests that estimating 3D facial key-points with $E_{can}$ is nearly 8x faster than HeadGAN’s 3D face reconstruction pipeline, i.e., 3DMM fitting followed by 3D face rendering. However, given that these two operations can be executed concurrently for subsequent driving frames, the implementation of HeadGAN uses concepts from asynchronous processing and reduces the execution speed of the overall pipeline to the speed of the slower process, which is the 3DMM fitting step. Therefore, in practice, the estimation of 3D key-points with network $E_{can}$ is around 4.5x faster than the 3D face reconstruction pipeline of HeadGAN [17].

### D. Ablation Study

We follow a few-shot learning approach to adjust our system in a setting where multiple source images are available. Although our flow network extended to predict weights has been trained in a 2-shot scenario, during inference it can operate for any number of source frames $N$. We examine the setting where $N = 2, 4$ or 8 source frames are given. In Fig. 10 we show the effect of the number of source images $N$ in the quality of synthesised data. The results clearly demonstrate the beneficial effect of N-shot learning.

In Table III, we qualitatively measure the impact of N-shot learning on the task of reconstruction, using the test set of
VoxCeleb. As suggested by all evaluation metrics, the generative performance of our system improves when the number of available source images increases.

Next, we examine the significance of the two face tracking components of our talking head synthesis system and more specifically networks $E_{can}$ and $E_{gaze}$. We evaluate the importance of network $E_{can}$ for cross-identity motion transfer, as the canonical key-point estimator $E_{can}$ is particularly developed to tackle the identity mismatches in reenactment, since it allows to adapt the target key-points to the facial shape of the source. To that end, we develop a variation of Free-HeadGAN by replacing $E_{can}$ network with RetinaFace. That is, we utilise directly the 3D key-points regressed by RetinaFace in order to draw the sketches that serve as conditional input to the generator, without adapting them to the identity of the source. The quantitative results displayed in Table IV indicate that removing $E_{can}$ has a critical effect on CSIM metric that measures identity preservation. Moreover, the overall quality of samples degrades as FID score increases. On the contrary, expression transfer slightly improves, indicating that the canonical space key-points actually retain a small amount of expression information.

In order to validate the importance of explicit eye gaze conditioning in image synthesis, we test a second variation of our system, where we do not encode any gaze information into sketches. To that end, we train a Free-HeadGAN generator that learns to transfer gaze direction relying solely on the facial key-points that belong to the eyes, as we do not colour-code the gaze angles inside the eye cavities. We evaluate the performance of this variation with AGD metric, which measures the average angular distance between the driving and generated gaze on the test data for reenactment. We display AGD results in Table IV, which confirms the significance of gaze estimator $E_{gaze}$ quantitatively. Additionally, in Fig. 11 we illustrate some cases where the model variation without explicit gaze input fails to capture the direction of the target eyes.

Finally, we verify the benefits of inferring the gaze direction implicitly from the predicted 3D eye mesh, instead of learning to predict the 3D gaze vector directly. To that end, we train two versions of $E_{gaze}$. One that predicts dense eye meshes as described in Section III-B, namely $E_{meshes}^{gaze}$, and one that estimates 3D gaze vectors, denoted as $E_{vectors}^{gaze}$. We perform within-dataset, cross-subject experiments on the recent gaze estimation datasets MPIIGaze [38], Columbia [75], UTMV [76], and Gaze360 [58]. We compute the AGD metric and present our results in Table V, which confirm our initial suggestion that estimating a 3D mesh of the eyes is beneficial for gaze estimation.
E. Pose and Gaze Manipulation

Apart from cross-identity motion transfer (reenactment) and facial video compression and reconstruction (self-reenactment), we can use Free-HeadGAN to edit facial images. That is, we can set the driving head pose manually, simply by rotating the 3D source key-points. Moreover, we can change the gazing direction of the reference subject, by editing the estimated gaze angles before feeding them to the generator through the target sketch. In Fig. 12 we demonstrate the ability of our proposed system to edit the head pose and eye gaze of random reference images. Here, the left column shows the original reference image, while the next three columns depict synthesised images by our method, after editing pose and gaze. Our image editing experiments show that Free-HeadGAN is a powerful tool for data augmentation. Our method could possibly replace the naive and widely-used affine transformations on facial images with complex non-linear transformations of human heads. Moreover, considering that our system provides strong identity preservation, such image augmentation could benefit various computer vision tasks related to facial identity recognition.

VI. LIMITATIONS

Although our method achieves very convincing and photorealistic reenactment results, while offering high image quality and strong identity preservation properties, it does not come without limitations. In particular, the performance of Free-HeadGAN in extreme poses such as large head rotations is bounded by the limited head pose distribution of available databases, such as VoxCeleb [54]. Moreover, as suggested by our qualitative analysis and FID scores [69], there exists a significant performance gap between self-reenactment and reenactment. Apart from the information that is lost during the adaptation of key-points in reenactment, this gap is partly attributed to the random selection of source and target images during training, and could be reduced by a more sophisticated sampling strategy.

Although Free-HeadGAN does not rely on 3DMMs [21], its performance highly depends on the accuracy of the pseudo-ground truth data, and more specifically the 3D facial landmarks of RetinaFace [25] and the iris landmarks of [39]. One possible way to avoid this issue would be to follow a self-supervised approach [6], [15], [16], which also comes with limitations, such as inconsistencies in key-point placement. Some more recent works [77], [78] attempt to overcome this problem of self-supervised key-points, by learning to predict distinct facial or body part segments that cover all moving areas of the input frames. Nevertheless, these methods rely on optical flow approximations, and thus combining them with our dense flow estimation and rendering network is not straightforward.

Last but not least, considering that our generative module is merely image-based, there are some cases where jittering effects take place in our videos. This phenomenon is actually more prominent in areas where conditional information is completely missing, such as the upper torso and hair, as there is more uncertainty about the content to be synthesised. On the contrary, jitters are less likely to happen inside the facial region, thanks to the input provided by facial landmarks. Arguably, one could resort to video-based modeling [11], [31] to alleviate this problem. For instance, a video-to-video translation architecture could be explored for reducing the temporal inconsistencies that manifest during one-shot reenactment. This would be an interesting direction for future research.

VI. CONCLUSION

We presented Free-HeadGAN, a model that extends HeadGAN [17] while releasing it from 3DMM fitting and 3D face rendering requirements in pre-processing. Instead, we condition synthesis on sketches of 3D key-points, which also support free-view synthesis. We designed a network that takes care of the identity mismatches in cross-identity reenactment, thought free-view synthesis. We designed a network that takes care of the identity mismatches in cross-identity reenactment, thought

REFERENCES

[1] J. Thies, M. Zollhöfer, M. Nießner, L. Vaegaerts, M. Stamminger, and C. Theobalt, “Real-time expression transfer for facial reenactment,” ACM Trans. Graph., vol. 34, no. 6, 2015, Art. no. 183.
[2] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2387–2395.
[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, “Headon: Real-time reenactment of human portrait videos,” ACM Trans. Graph., vol. 37, 2018, Art. no. 164.
[4] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8107–8116.
[6] T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-shot free-view neural talking-head synthesis for video conferencing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10034–10044.
[7] H. Kim et al., “Deep video portraits,” ACM Trans. Graph., vol. 37, no. 4, 2018, Art. no. 163.
[8] M. Koujan, M. Doukas, A. Roussos, and S. Zafeiriou, “Head2Head: Video-based neural head synthesis,” in Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit., 2020, pp. 16–23.
[9] M. C. Doukas, M. R. Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou, “Head2Head: Deep facial attributes re-targeting,” IEEE Trans. Biom., Behav., Ident. Sci., vol. 3, no. 1, pp. 31–43, Jan. 2021.
[10] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9458–9467.
[61] W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1874–1883.

[62] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 105–114.

[63] J. H. Lim and J. C. Ye, “Geometric GAN,” 2017, arXiv:1705.02894.

[64] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, “Learning to super-resolve blurry face and text images,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 251–260.

[65] S. Yang, P. Luo, C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5525–5533.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[67] 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 Proc. Adv. Neural Inf. Process. Syst., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 6626–6637.

[68] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 586–595.

[69] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2015, arXiv:1412.6980.

[70] J. Deng, J. Guo, X. Niaman, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4685–4694.

[71] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “OpenFace: A general-purpose face recognition library with mobile applications,” CMU School Comput. Sci., Pittsburgh, PA, Tech. Rep. CMU-CS-16-118, 2016.

[72] T. Baltrusaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic action unit detection,” in Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit., 2015, pp. 1–6.

[73] J. S. Chung, A. Nagrani, and A. Zisserman, “V oxCeleb2: Deep speaker recognition,” in Proc. Annua. Conf. Int. Speech Commun. Assoc., 2018, pp. 1086–1090.

[74] B. Smith, Q. Yin, S. Feiner, and S. Nayar, “Gaze locking: Passive eye contact detection for human-object interaction,” in Proc. ACM Symp. User Interface Softw. Technol., 2013, pp. 271–280.

[75] Y. Sugano, Y. Matsushita, and Y. Sato, “Learning-by-synthesis for appearance-based 3D gaze estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1821–1828.

[76] A. Siarohin, S. Roy, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Motion-supervised co-part segmentation,” in Proc. 25th Int. Conf. Pattern Recognit., 2021, pp. 9650–9657.

[77] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, “Motion representations for articulated animation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13653–13662.

Evangelos Ververas received the graduate degree from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTH), Greece, in 2016. He is currently working toward the PhD degree with the Department of Computing, Imperial College London, supervised by Prof. Stefanos Zafeiriou, and a researcher with Huawei U.K. Research Centre. His research interests lie in the fields of computer vision and deep learning for 3D face analysis and reconstruction, as well as 3D gaze estimation.

Viktoria Sharmanska is passionate about designing intelligent systems that can learn concepts from visual data using machine learning models. She has joined the University of Sussex as a lecturer in artificial intelligence, in 2020. Prior to this, 2017–2020, she was an Imperial College Research fellow with the Department of Computing, Imperial College London, working on deep learning methods for emotion and human behaviour recognition. She is also an honorary lecturer with Imperial College London since 2021. She has co-authored numerous papers on novel statistical machine learning methodologies applied to computer vision problems, such as attribute-based object recognition, learning using privileged information, cross-modal learning, human facial behaviour analysis, and recently on algorithmic fairness methods, published in the most prestigious conferences in her field of research, such as CVPR, ICCV/ECCV, NeurIPS. She has built an international reputation such as being among the youngest area chair for top-tier international conferences in computer vision and deep learning such as International Conference on Learning Representations (ICLR) 2019–2021 and Conference on Computer Vision and Pattern Recognition (CVPR) 2021. She is actively involved in organising and promoting Women in Computer Vision, Women in Machine Learning, and in increasing visibility of the underrepresented groups in computing. In 2018, together with colleagues from Facebook AI, Stanford University, Autonomous University of Barcelona she co-organised the international Women in Computer Vision workshop, held in conjunction with CVPR 2018.

Stefanos Zafeiriou (Member, IEEE) is a reader in machine learning and computer vision with the Department of Computing, Imperial College London, U.K., and a distinguishing research fellow with the University of Oulu. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London, in 2011 to start his own independent research group. He was the recipient of the President’s Medal for Excellence in Research Supervision for 2016. He currently serves as an associate editor of IEEE Transactions on Affective Computing and Computer Vision and Image Understanding journal. He has been a guest editor of more than six journal special issues and co-organised more than 13 workshops/special sessions in top venues, such as CVPR/FG/ICCV/ECCV. He has coauthored more than 55 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, published in the most prestigious journals in his field of research, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, the International Journal of Computer Vision, IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Visualization and Computer Graphics, and IEEE Transactions on Information Forensics and Security, and many papers in top conferences. He has more than 18000 citations to his work, h-index 51.

Michail Christos Doukas received the graduated degree from the Department of Electrical and Computer Engineering (MEng), National Technical University of Athens (NTUA), Greece, in 2016, and the master’s degree in computing (MSc) from Imperial College London, in 2017. He is currently working toward the PhD degree with Imperial College London, supervised by Prof. Stefanos Zafeiriou and Dr. Viktoria Sharmanska. He is currently a research intern with Huawei U.K. Research Centre. His research interests lie in the fields of computer vision, deep learning and more specifically deep generative modeling, image and video synthesis, generative adversarial neural networks, face and head reenactment, visual speech synthesis and few-shot learning.