FACEGAN: Facial Attribute Controllable rEenactment GAN

Soumya Tripathy
Tampere University
Finland
soumya.tripathy@tuni.fi

Juho Kannala
Aalto University
Finland
juho.kannala@aalto.fi

Esa Rahtu
Tampere University
Finland
esa.rahtu@tuni.fi

Abstract

The face reenactment is a popular facial animation method where the person's identity is taken from the source image and the facial motion from the driving image. Recent works have demonstrated high quality results by combining the facial landmark based motion representations with the generative adversarial networks. These models perform best if the source and driving images depict the same person or if the facial structures are otherwise very similar. However, if the identity differs, the driving facial structures leak to the output distorting the reenactment result. We propose a novel Facial Attribute Controllable rEenactment GAN (FACEGAN), which transfers the facial motion from the driving face via the Action Unit (AU) representation. Unlike facial landmarks, the AUs are independent of the facial structure preventing the identity leak. Moreover, AUs provide a human interpretable way to control the reenactment. FACEGAN processes background and face regions separately for optimized output quality. The extensive quantitative and qualitative comparisons show a clear improvement over the state-of-the-art in a single source reenactment task. The results are best illustrated in the reenactment video provided in the supplementary material. The source code will be made available upon publication of the paper.

1. Introduction

Face-reenactment is a process of animating a source face according to the motion (pose and expression) of a driving face. In general, the process involves three major steps: 1) creating a representation of the source face identity, 2) extracting and encoding the motion of the driving face, and 3) combining the identity and motion representations to produce a modified source face. Each part has a significant impact on the output quality.

Number of algorithms, including traditional *3D* face models [3, 4], data driven Neural-Networks [5, 6, 7, 1, 8, 9, 10, 11, 12], and their combinations [13] have been presented for creating photorealistic face animations. In *3D*
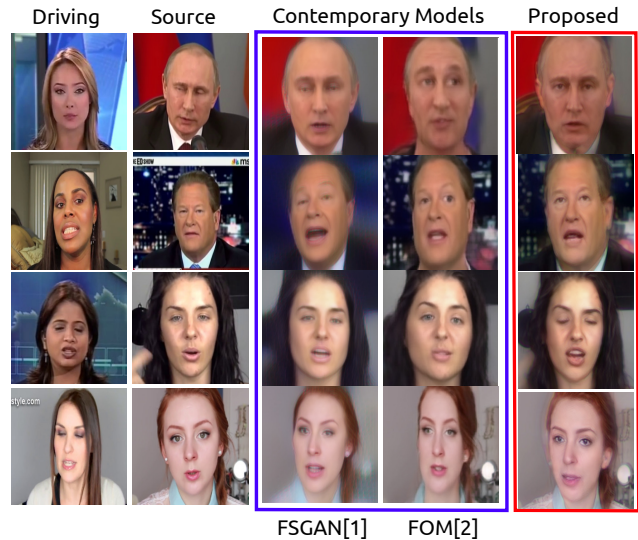

Figure 1. Examples of the face-reenactment results using the proposed FACEGAN and recent baseline works FSGAN [1] and First-Order-motion-Model (FOM) [2]. FACEGAN is able to preserve the source identity faithfully while imposing the driving pose and expression. The baselines suffer particularly in preserving the facial shape of the source.

face-model based approach [3] , the identity and motion features are encoded with *3D* model parameters. The reenacted face is then rendered using the identity parameters of the source and motion parameters of the driving face. Although this approach results in high quality outputs, they require substantial efforts in obtaining the faithful *3D* representations of the faces. Therefore, such approaches are often limited to a few source identities.

In recent years, the data driven deep neural networks have gained popularity in generating and manipulating images. In particular, the models using adversarial loss functions [14] have obtained highly realistic image synthesis results [15, 16]. Similar models are also applied for the face reenactment as demonstrated in [5, 1, 12, 11, 6, 17, 10, 2, 18]. In most of these works, the first step is to represent the source and driving faces with deep networks. These representations can be either latent codes or human interpretable
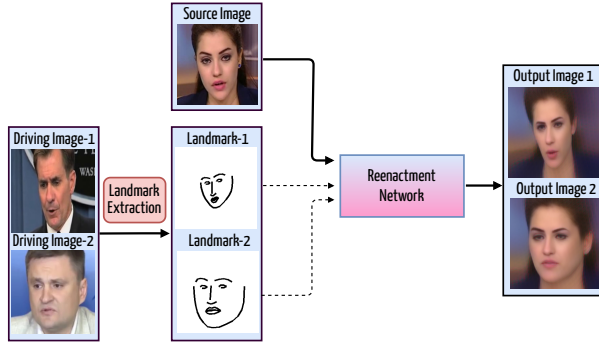
Figure 2. An illustration of the facial structure leaking problem in the landmark based reenactment approach. The facial structure of the generated output clearly follows the shape of the driving face instead of source face as intended.

representations like facial landmarks. Subsequently, the facial feature representations are combined by a generator network to produce the final output. Although these models generalise well to multiple identities, they have numerous challenges discussed in the following.

Landmark based motion representation Most recent works [5, 1] use facial landmark points as the representation of the facial motion (pose and expression). Although landmarks can provide strong supervision of the driving motion, they also contain the overall structure of the driving face in the form of facial contours. If the driving landmarks are directly used to animate the source face, the difference between the facial contours may lead to dramatic distortions in the output result. Therefore, the gap in the facial shapes usually leads to either losing the source identity or low photorealism. Figure 2 illustrates a practical example of such distortions. For this reason, many works search for driving and source pairs with highly similar face structures. However, this greatly limits the applicability of such methods.

Learning based representation Instead of landmarks, some works [11, 5, 18] use unsupervisedly learned latent features to represent the facial identity and motion. These representations are often not disentangled with respect to identity, structure and motion. Hence, the driving motion features may contain the driving identity information, which ends up to the final reenacted output. This identity leakage problem is illustrated in figures 5,1 in the case of X2face [11] and FOM [2].

Generalization and inference time training In some works, the entire reenactment model is trained for a particular source identity [12] or the model is fine tuned using a few shot approach at the inference time [5]. Although such a strategy may improve the output quality, it requires resources and time to adapt to a new identity.

Selective editing It is often desirable to selectively edit the reenactment result, for example, add an extra smile or open the eyes. This type of editing would be easier by directly manipulating the motion representation instead of

searching or producing an alternative driving image. To this end, the representation of the driving motion should be interpretable for the human editor. Although landmark points can be easily moved for editing, one needs sophisticated tools to preserve the structure of the source face while adjusting the landmarks. An alternative approach based on Action Units (AUs) is presented in [10], which allows editing without distorting the identity. However, the model was not able to produce high fidelity reenactment results.

In this paper, we propose a Facial Attribute Controllable rEenactment GAN (FACEGAN), which produces high quality source reenactment from various driving pairs even with significant facial structure differences between them. Our model manipulates the source face-landmarks with driving facial attributes to generate a new set of landmarks representing the desired motion with the source identity and structure. This mitigates the identity leakage problem present in many recent works. Furthermore, by representing the motion cues using the action units, we provide a selective editing interface to the reenactment process. The proposed method combines the benefits of the action units (subject agnostic and interpretable) and facial landmarks (strong supervision for image generation) into a single reenactment system. To the best of our knowledge, this is the first work that proposes such a combination for the face reenactment. In addition, FACEGAN decouples and handles facial region and background in two separate branches that help in generating the source background realistically in the final output unlike other models in the literature. Finally, we provide a detailed comparison of our method against the recent state-of-the-art works in a single source image reenactment. The proposed FACEGAN model results in superior performance both in quantitative and qualitative measures.

2. Related work

Face-reenactment has been actively studied over the years, and several different approaches have been presented. The proposed methods can be roughly grouped into the following three categories.

Morphable 3-D face models Parametric 3D face models are popular tools for face animation and several works [3, 19] have adopted them to face reenactment. These methods start by fitting a 3D morphable face model to source and driving faces. Afterward, the motion and pose parameters of the driving model are transferred to the source model, and the output face with the source identity and the driving motion is rendered. The motion parameters can also be controlled using other modalities such as audio [20]. Although these methods can provide high output quality, they are not easily scaled to a large number of identities.

Deep generative models Variational Auto Encoder (VAE) [21] is a popular model for image generation and manipulation. VAE encodes the input image into a latent

representation vector that is subsequently decoded back to the original input image. The VAE models can be utilised for face animation, by manipulating the latent representation of the source face according to the representation of the driving face [22]. If the latent representation is disentangled with respect to the identity and motion, the facial movements can be controlled by manipulating the corresponding vector elements of the encoded source face [23]. However, the disentangled representation is very difficult to obtain, which often leads to compromises in the output quality.

An alternative encoder-decoder based approach was presented in [11]. In this work, an encoder-decoder model is applied to warp the input face into an embedding image. Another encoder-decoder network converts the driving image into a warping field that maps the embedding image to the output face. This approach provides better quality compared to the VAE based models, but the driving identity can still easily leak to the reenacted output. This phenomenon is further emphasized if the source and driving faces have large structural or pose differences.

Recently, another warping based reenactment approach was proposed in [2]. The method learns a warping field using a set of keypoints from the source and driving images. This approach obtains higher output quality compared to [11], but it suffers from similar identity leakage problems. To this end, the authors of [2] proposed an alternative setup where the method is provided by an additional driving image with similar pose and expression as the source. This additional driving image was utilised to obtain better approximation of the identity independent motion of the driving face. Unfortunately, such a driving image would be challenging to obtain.

Deep Generative Adversarial Networks The Generative Adversarial Networks (GANs) are popular models for high quality image generation [15, 16] and manipulation [24, 25]. In [5, 1], a progressively growing GAN architecture was trained to animate a source face based on the landmarks of the driving face. As landmarks preserve the facial structure and, up to some extent, the identity, these models suffer from quality degradation if the driving and source faces are not similar. However, this approach is well suited for applications like telepresence where the driving and source identities are the same.

Landmark transfer models were proposed in [12] and [6] to remove the identity features from the driving landmarks. In [12], the landmark transformer was trained separately for each identity pair, which limited the scalability of the method. A principal component analysis (PCA) based model was used in [6] to separate expression and shape parameters from the landmarks. They utilised shape parameters from the source and expression parameters from the driving face to reconstruct the transferred landmarks. Although the method generalizes better compared to [12],

the proposed linear models are not sufficient to differentiate complex expressions from the shape information.

An action units (AUs) based face representation is used in [7] to manipulate facial expressions (not pose). More recently, in [10], the authors proposed a model that used AUs for the full face reenactment (expression and pose). The AUs represent complex facial expressions by modeling the specific muscle activities [26]. These activations are independent of the facial structure, which makes the action units a potential representation to disentangle driving motion from the identity. Unfortunately, the model presented in [10] was not able to produce similar output quality as the facial landmark based alternatives.

In this paper, we combine the disentanglement properties of the action units with strong supervision from the facial landmarks to produce a high quality reenactment results without significant identity leakage problems.

3. Method

Given an image, I_d with a driving face F_d and an image I_s with a source face F_s, the FACEGAN model aims to transfer the motion (pose and expression) from F_d to F_s while maintaining the identity of F_s. To do so, the model takes the landmarks of F_s and manipulates them to include the motion of F_d. The motion information is extracted from F_d in terms of action units (AUs), which correspond to various face muscle activations [27] and head pose angles. The overall shape features of the source landmarks are not altered during the transformation, which helps to preserve the source identity. The transformed landmarks are passed to reenactment and background mapper modules, which together generate a face image depicting the source identity and background, but with driving facial motion.

The overall architecture of FACEGAN contains three main components, which are illustrated in figure 3. The first component, called landmark-transformer, consisting of a fully connected network L_T. It takes the landmarks l_s from F_s and the action units AU_d of F_d as inputs and generates the transformed landmarks L_t as an output. L_t has pose and expression from F_d, whereas the facial structure from F_s (as the input landmarks are from F_s).

The transformed landmarks L_t are converted to a gray scale image H_t, as shown in Figure 3, and provided to the next component called as face-reenactor. The face-reenactor takes I_s and H_t as inputs for the generator G_r and produces the reenacted face I_{fr} as an output. G_r contains also a segmentation unit G_{rs}, which provides a face, hair and background segmentation maps of I_{fr}.

By removing the background from I_{fr} and the face from I_s, we obtain two images called I_f and I_b. These images are provided to the third module called background mixer. I_f contains the reenacted face information whereas I_b contains the source background. Unlike other contemporary
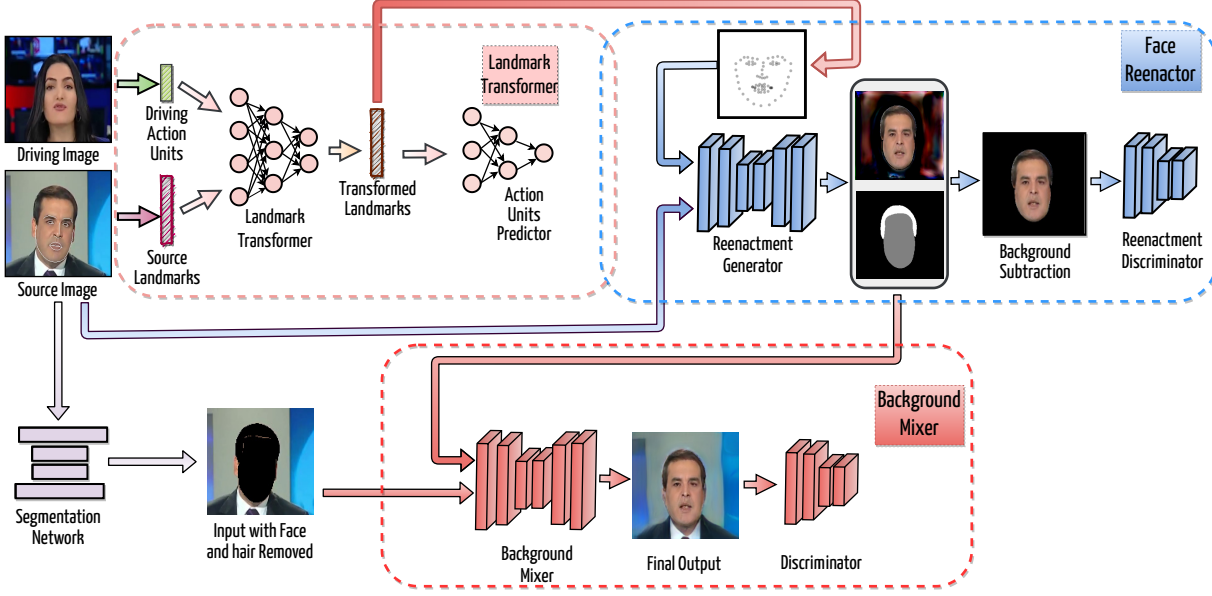
Figure 3. The overall architecture of the proposed FACEGAN contains three major blocks, namely 1. Landmark transformer, 2. Face reenactor, and 3. Background Mixer. The Landmark transformer modifies the source landmarks to include the driving face's expression and pose, which are represented in terms of action units (AUs). The modified landmarks and the source image are provided to the Face reenactor module for producing the reenacted face. Finally, the reenacted face and the source background are combined into the output by the Background mixer module. In the training phase, we use the source and driving images from the same face track, which allows us to use the driving image as a pixel-wise ground truth of the output. However, during the inference, the source and driving may have different identities.

approaches, our model processes the background separately from the face area. In this way, different components will specialize to face reenactment and background manipulation, leading to improved output quality. The background mixer contains a generator G_b producing I_r, which has facial identity and background from I_s but facial pose and expression from I_d.

The training process of our model requires image frames extracted from talking head videos. In the following, we assume that the facial landmarks of the source faces, and the head poses and the action units (AUs) of the driving faces, are already extracted and stored. This can be achieved using publicly available landmark [28] and Pose-AU detectors [29, 30]. For simplicity, we denote the AU and pose combination as AUs throughout the text unless mentioned otherwise. In each step of the training procedure, we extract two frames I_s and I_d from the same video track as described in [11, 10, 5]. In this way, the exact ground truth of our reenacted source is available in terms of I_d. The detailed training steps and loss functions are discussed in the following sections.

3.1. Landmark Transformer

The landmark transformer L_T modifies the source landmarks according to the driving motion, while maintaining the facial shape and identity of the source face. First, *2D*

facial landmarks are extracted from the source image $I_s \in \mathbf{R}^{3 \times H \times W}$ and these are subsequently reshaped to $l_s \in \mathbf{R}^{136 \times 1}$. The facial motion of $I_d \in \mathbf{R}^{3 \times H \times W}$ is extracted in terms of the action units ($\in \mathbf{R}^{17 \times 1}$) and the head pose angles ($\in \mathbf{R}^{3 \times 1}$). These two parameters are concatenated in to a single vector $AU_d \in \mathbf{R}^{20 \times 1}$ that describes the complete motion of F_d. The landmark transformer L_T takes the concatenated l_s and AU_d as an input and predicts the landmark movements δl_s. Finally, the transformed source landmarks are obtained as $l_t = l_s + \delta l_s$.

The landmark transformer is trained using various loss functions. The most straightforward loss is calculated between l_t and the landmarks of I_d denoted as l_d. Since during training, I_s and I_d are from the same video track, the landmarks l_d form the ground truth for l_t. To smooth the predictions δl_s a l2-weighted penalty regularization is added along with the reconstruction loss. The objective function can be written as,

$$\mathcal{L}_{lr} = ||l_t - l_d||_1 + \lambda_{lr}||\delta_s||_2. \qquad (1)$$

Although $\mathcal{L}_{lr}$ encourage the proper pose variations in l_t, they fail to capture the subtle movements due to the expression variations. In order to focus on expression, another fully connected network L_a is introduced to regress the AU parameters from the landmarks. L_a is trained simultane-

ously with l_t using the following loss function

$$\mathcal{L}_{lau} = ||L_a(l_t) - AU_d||_1 + ||L_a(l_d) - AU_d||_1. \quad (2)$$

In order to preserve the facial shape in landmark domain, we include a connectivity loss. The loss preserves the distances between the neighbouring landmark points in l_s and l_t. The loss function is mathematically expressed as,

$$\mathcal{L}_{lc} = ||D_t - D_d||_1, \quad (3)$$

where $D_t \in \mathbf{R}^{d \times 1}$ and $D_d \in \mathbf{R}^{d \times 1}$ are vectors containing the differences between the connected landmark positions of l_t and l_d. Here d is the number of connected landmarks included in the loss function calculation. Finally, the full loss function for the landmark transformer is formed by a linear combination of equations (1), (2), and (3) as

$$\mathcal{L}_l = \lambda_{l1}\mathcal{L}_{lr} + \lambda_{l2}\mathcal{L}_{lau} + \lambda_{l3}\mathcal{L}_{lc}, \quad (4)$$

where λ_{l1}, λ_{l2}, and λ_{l3} are weighting constants.

3.2. Face Reenactor

The transformed landmarks l_t are mapped to single channel image $H_t \in \mathbf{R}^{1 \times H \times W}$ by placing a 2D Gaussian function at each keypoint location as shown in figure 3. H_t is channel-wise concatenated with the source image I_s to form the input for the face reenactor network G_r. G_r produces a reenacted RGB image $I_{fr} \in \mathbf{R}^{3 \times H \times W}$ and a segmentation map S_{fr} $\mathbf{R}^{3 \times H \times W}$ with face, hair and background classes. The segmentation maps are obtained using a CNN head added to the second last layer of the G_r. For training, the pixel wise reconstruction loss, VGG perceptual loss [25], and adversarial loss are applied as given in equations (5), (6), and (7). The segmentation branch is trained with standard cross entropy loss. The full loss function for G_r is provided in Equation (8).

$$\mathcal{L}_{rr} = ||I'_{fr} - F_d||_1 \quad (5)$$

$$\mathcal{L}_{rp} = \sum_i \frac{1}{C_i H_i W_i} ||V_i(I'_{fr}) - V_i(F_d)||_1 \quad (6)$$

$$\mathcal{L}_{radv} = \min_{G_r} \max_{D_{r1}, D_{r2}, D_{r3}} \sum_{r=1}^{3} \mathbb{E}_{F_d}[logD_r(F_d)] \\ + \mathbb{E}_{L_S}\left[log(1 - D_r(I'_{fr}))\right] \quad (7)$$

$$\mathcal{L}_r = \lambda_{r1}\mathcal{L}_{rr} + \lambda_{r2}\mathcal{L}_{rp} + \lambda_{r3}\mathcal{L}_{radv} + \lambda_{r4}\mathcal{L}_{ce}. \quad (8)$$

I'_{fr} and F_d are I_{fr} and I_d with background removed. Each D_r in equation (7) stands for discriminator, used for multiple resolution of images as described in [25]. $\mathcal{L}_{ce}$ is a standard cross entropy loss for S_{fr}. In order to obtain the ground truth for $\mathcal{L}_{ce}$, a pretrained face-segmentation network G_s is utilized as explained in [8].

3.3. Background Mixer

The input to the background mixer network G_b is a channel-wise concatenation of I'_{fr} and I_b, where I_b is I_s with face and hair removed. G_b generates an RGB image I_c and a single channel mask $M \in \mathbf{R}^{HXW}$. Finally, the reenacted output image I_r is obtained using I_c, I_b, and M as

$$I_r = M * I_c + (1 - M) * I_b. \quad (9)$$

In this way, G_b is encouraged to focus on producing the compatible background while directly copying as much information form I_b as possible. G_b is trained using pixel loss and adversarial loss on I_r, and an additional regularization for M. The regularization is obtained by imposing a l_2 weight penalty, smoothing the mixing process in equation 10, and applying a total variation regularization. The full loss on M can be written as,

$$\mathcal{L}_{bm} = \lambda_{b1} \sum_{x,y}^{H,W} [(M^{x+1,y} - M^{x,y})^2 + (M^{x,y+1} - M^{x,y})^2] \\ + \lambda_{b2}||M||_2 \quad (10)$$

Finally, the complete loss function for the background transformer with the regularization parameters can be written as,

$$\mathcal{L}_b = \mathcal{L}_{bm} + \lambda_{b3}\mathcal{L}_{bp} + \lambda_{b4}\mathcal{L}_{badv} + \lambda_{b5}\mathcal{L}_{br}, \quad (11)$$

where $\mathcal{L}_{bp}, \mathcal{L}_{br}$ are perceptual loss and reconstruction loss between I_r and I_d, respectively. $\mathcal{L}_{badv}$ is the adversarial loss of the Equation (7) for the discriminator D_b and G_b.

4. Training Details

We train FACEGAN using a dataset created from IJB-C videos [31]. Most of these videos contain celebrities talking in an unconstrained setup. The faces are first detected using a landmark detector and then tracked using a centroid tracker over the video frames. Around the centroids, a fixed image crop is used to extract the faces in a way that the middle position between the eyes will always be in a fixed position. Small faces are rejected based on landmarks height and width. In total 400k good quality face images are obtained. In addition, we used 400 videos from the Forensic++ dataset [32] to evaluate the performance of our system. Forensic++ was pre-processed in same way as IJB-C, resulting in total $200k$ face images.

The networks G_r and G_b have U-Net like structures and are trained progressively as given in [25]. We first trained the networks to generate images of resolution 128×128 and then increased to the final resolution of 256×256. The landmark transformer L_T is a fully connected network predicting landmark locations in normalized coordinates. These

can be converted to a single channel image of any resolution. All networks are trained separately for a few epochs and then together in and end-to-end manner. For L_T and for other generator networks, the learning rate is kept at 0.0002 whereas the batch-size is 32 for the former and 1 for the latter.

5. Experiments

We assess the proposed approach in a series of experiments and compare the results against the recent state-of-the-art works [11, 1, 2, 10] in quantitative and qualitative terms. We start by evaluating our landmark transformer in isolation and then continue with the full face reenactment comparisons. All experiments are performed using the $200k$ face images obtained from the FaceForensic++ dataset [32] as described in Section 4. We also note that none of the tested models, including ours, are trained with FaceForensic++, which emphasizes the generalization properties of the models. In all cases, we use one source image per identity and the final output resolution is 256×256.

We compare our approach with four recent works, namely X2face [11], FSGAN [1], First-Order-motion-Model(FOM) [2] and ICface [10]. These state-of-the-art models are representatives from a broader set of reenactment frameworks. For instance, FSGAN and ICface [10] utilise similar ideas of interpretable facial attributes like landmarks and AUs in reenactment. In contrast, X2face [11] and First-order-motion-model [2] represent works that learn motion and identity representation in an unsupervised (or self-supervised) setups. By comparing FACEGAN with these representatives from both categories, we obtain better understanding of the drawback in the existing works and the contributions of FACEGAN. We used the original source codes for all comparison methods. We did not consider the few-shot learning models [5] and [6] since they require inference time training and the corresponding implementations are not available.

5.1. Landmark Transformer

In order to evaluate the importance of the landmark transformer (L_T) in our model, we performed self-reenactment and cross-reenactment with and without L_T. The self-reenactment refers to a case where the source and driving images are from the same identity whereas in cross-reenactment these identities are different. In the former case, we take the source and driving images from the same face track, which allows us to use the driving image as a pixel-wise ground truth of the reenacted output. Here we evaluate the result using a pixel-wise with Mean Square euclidean Loss (MSE). In the cross-reenactment case, we do not have access to the ground truth image and cannot use the MSE loss. Instead, inspired by [6, 5, 33], we calculate the following three measures
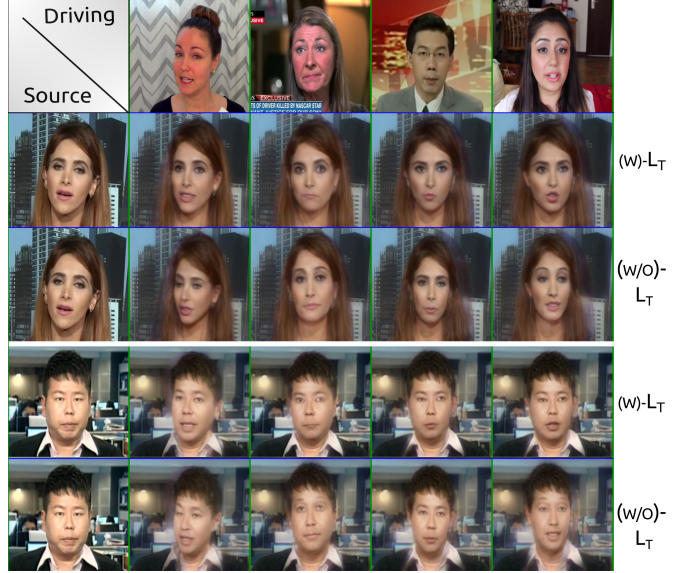

Figure 4. Qualitative comparison of FACEGAN outputs with and without the landmark transformer. The facial shape and identity are clearly distorted if the landmark transformer is not applied.

1. *Cosine Similarity between IMage embeddings (CSIM):* uses a pre-trained face recognition network [34] to obtain embeddings from the source and driving images. Higher score indicates that the source identity is better preserved in the reenactment process. 2. *Pose Cosine Similarity between IMage (PSIM):* is calculated between the head pose angles, estimated using Openface [29], of the driving and the reenacted faces. PSIM measures the ability of the model to retain the driving head poses in the final output. 3. *Expression difference (ED):* is the Euclidean distance between the action units, calculated with Openface [29], of driving and reenacted images. It measures the ability of the model to retain the driving expressions in the final image.

The results of the landmark transformer experiments are presented in Table 1. In the self-reenactment case, the best performance is obtained without using the landmark transformer. The result is expected, since in the self-reenactment setup, the driving landmarks represent the optimal output of the landmark transformer. Nevertheless, the MSE score is only slightly lower if the landmark transformer is applied, which indicates that the expression and pose are faithfully transferred from the driving face.

In the cross-reenactment case, we observe a significant difference in the CSIM score while the PSIM and ED scores are similar. High CSIM score indicates that the source identity is well preserved in the reenactment output. The large difference in terms of CSIM shows that the landmark transformer clearly improves the reenactment quality and reduces the identity leakage problem. The PSIM and ED scores measure pose and expression similarity between the output and the driving face. These measures do not depend

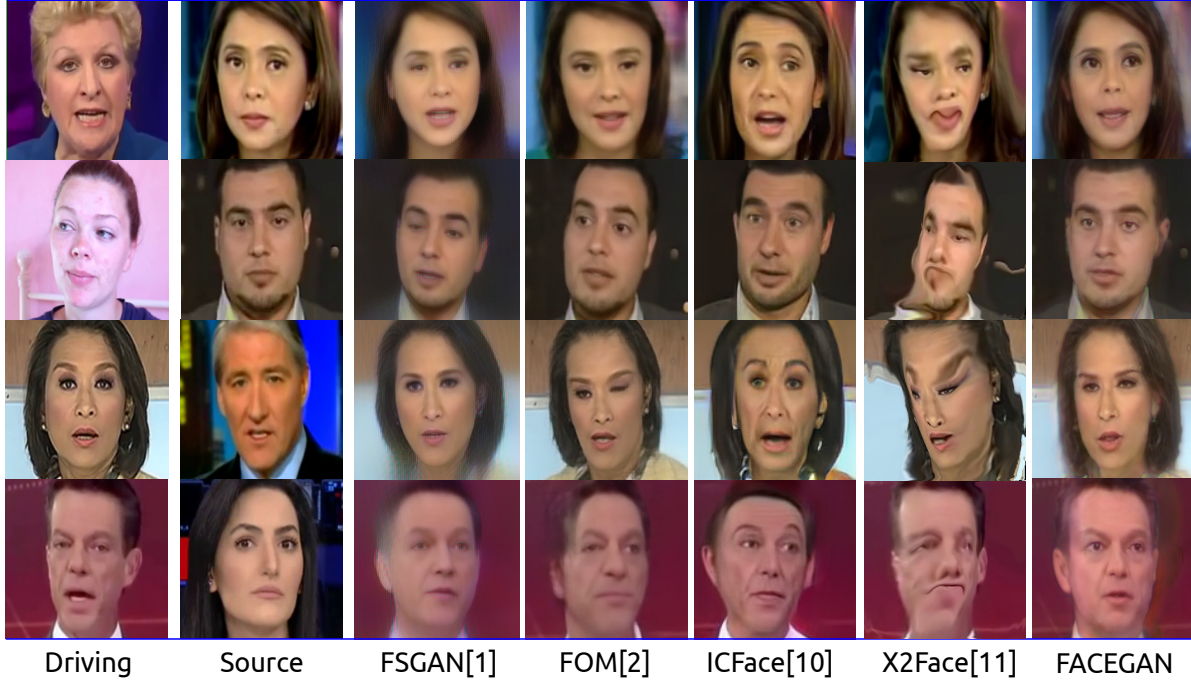| Driving | Source | FSGAN[1] | FOM[2] | ICFace[10] | X2Face[11] | FACEGAN |

Figure 5. Qualitative comparison of FACEGAN and four recent baseline methods: FSGAN [1], First-Order-motion-Model(FOM) [2], ICface [10], and X2face [11]. FACEGAN is clearly able to retain the source face shape and identity better compared to the baselines. Moreover, the driving motion is faithfully reproduced.

	Self-Reenactment	Cross-Reenactment		
Model	MSE↓	CSIM↑	PSIM↑	ED↓
FACEGAN w L_T	0.018	**0.747**	0.811	**0.012**
FACEGAN w/o L_T	**0.015**	0.623	**0.885**	0.014

Table 1. Quantitative evaluation of the impact of the Landmark transformer(L_T). High CSIM and PSIM scores indicate the ability to preserve the source identity and driving head pose, respectively. Low ED score reflects the ability to reproduce the driving expression. The MSE score measures the pixel-wise differences in the self-reenactment case.

on the face identity and the driving landmarks directly provide strong supervision in these terms. However, the landmark transformer results in similar or better scores also according to these measures, which indicates that the pose and expression are faithfully transferred despite substantial improvement in preserving the source identity. Figure 4 shows a few examples of the cross-reenactment results for FACE-GAN with and without the landmark transformer. These illustrate in particular how the face structure is preserved better by using the landmark transformer.

5.2. Qualitative Reenactment Results

In this section, we assess the reenactment performance of the proposed model in qualitative terms and compare it with X2face [11], FSGAN [1], First-Order-motion-Model(FOM) [2] and ICface [10]. The results for versatile driving-source pairs are illustrated in figures 5 and 1. The source identity is well preserved in the FACEGAN results, and there is

considerably less visible structure leakage compared to the baseline methods. The difference is particularly due to the landmark transformer, which helps to preserve the shape of the source face.

In addition, most models like X2Face [11], FSGAN [1], and ICFace [10] operate only on a tight crop around the face area. Such approach completely ignores the untrivial integration of face, background, and other body parts such as hair. These limitations greatly hamper the practical usability of the methods. In contrast, FACEGAN has a dedicated background mixer model that generates the context for the reenacted face by hallucinating high quality background pixels along with the ears and the upper body parts as shown in figures 4 and 6 (Note that the background is cropped out from FACEGAN results in figures 5 and 1 to facilitate comparison). The background mixer enables the reenactment network to focus on the face and hair regions improving the reenactment quality of these parts. In conclusion, our model produces sharper and better reenactment results from a single source and driving images of different identity in comparison to the state-of-the-art models.

5.3. Quantitative Reenactment Results

In this section, we compare the FACEGAN results with the baseline models in quantitative terms. We use similar self and cross-reenactment setups as in Section 5.1 and report the results using MSE, CSIM, PSIM and ED scores in Table 2. In the self-reenactment case, all methods ex-

cept ICFace obtain similar MSE scores. Comparison to IC-Face demonstrates the benefits of using the facial landmark representation instead of pure action units. In the cross-reenactment setup, FACEGAN obtains the highest performance in terms of the CSIM score with a clear margin. This result further illustrates the ability of FACEGAN to preserve the source identity faithfully. In terms of PSIM, the direct landmark based FSGAN model results in the best pose retention capability. However, FACEGAN and FOM are not far behind the FSGAN. ICFace has the lowest PSIM score, which further illustrates the difficulty of producing faithful pose using purely action unit based supervision. In terms of ED metrics, FACEGAN obtains the best results, followed by ICFace.

While the CSIM score reflects the overall ability to preserve the identity, it may be inadequate to reflect small scale structural differences. The difficulty, however, would be to obtain a sufficient ground truth for measuring such details. For this purpose, we propose a new measure called landmark similarity score (LSIM). To calculate the measure, we first randomly choose two source images from a single identity. Then we search the test database for a driving image with different identity but as similar action units as possible to the second source image. Finally, we use the motion information (AUs, landmarks, warping parameters) from the discovered driving image to reenact the face in the first source image. The output is compared with the second source image that acts as a pseudo ground truth for the output. To accommodate for a small differences, we calculate LSIM as a mean square error between the corresponding landmark locations instead of pixel level MSE. If the reenactment process preserves the facial shape and motions accurately then it should result in low LSIM score. Table 2 contains the obtained results. The proposed FACEGAN model achieves clearly the highest performance among the compared approaches.

	Self-Reenactment	Cross-Reenactment			
Model	MSE↓	CSIM↑	PSIM↑	ED↓	LSIM↓
X2Face [11]	0.018	0.564	0.659	0.025	0.040
FSGAN [1]	0.016	0.631	**0.887**	0.016	0.035
ICFace [10]	0.106	0.655	0.699	0.013	0.039
FOM [2]	**0.012**	0.676	0.825	0.015	0.035
FACEGAN	0.018	**0.747**	0.811	**0.012**	**0.021**

Table 2. Quantitative comparison of our model with the state-of-the-arts works. High CSIM scores indicate better ability to preserve the source identity, while low LSIM signifies better landmark shape retention ability. The PSIM and ED measure the head pose and expression reproduction, respectively. FACEGAN obtains clearly the best CSIM and LSIM scores, while obtaining similar PSIM and ED scores. This indicates that our method is capable of preserving the source identity while faithfully reproducing the driving motion.
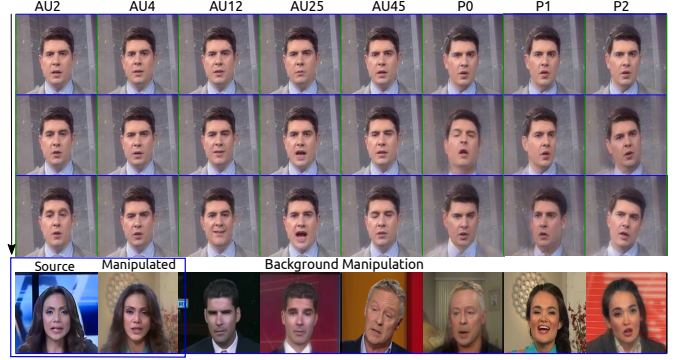

Figure 6. Demonstration of the selective editing capabilities of FACEGAN. The first three rows images are generated by increasing the source AUs directly from minimum to maximum value. The last column illustrate the capability of combining the reenacted face with a non-source background (zoom in to observe the quality better).

5.4. Controllable face reenactment

FACEGAN utilises 20 human interpretable action units (AUs) to represent the pose and expression of the desired output face. The AUs can be obtained from a driving face, but this is not the only option. For instance, one can take the AUs from the source face, manipulate them, and feed them as driving information to the model. In this way, one can selectively edit the source image by, for example, changing the pose or adding an extra smile. Figure 6 contains a few examples of such selective editing procudure.

Moreover, the background mixer can combine face and background from two different sources. This results in a mixed reenactment result, where the face identity is from one source and the background from another. Figure 6 illustrates a few examples of this kind of edits. Such feature would be very useful for relocating the source person into a desired environment. The proposed selective editing properties provide complete freedom and control to generate the desired reenactment video.

6. Conclusion

We proposed a facial animator called FACEGAN that is capable of performing high quality reenactment from a single source image. Unlike many previous works, our model does not pose any restriction on the compatibility of the source and driving pairs. The model combines the best properties of the action unit and facial landmark motion representations for reducing the identity leakage problem and to optimise the reenactment quality. Furthermore, FACE-GAN handles the face and the background separately which improves the output quality and gives additional control of choosing the desired background. We have compared our method with the state-of-the-art approaches and obtained superior results both quantitatively and qualitatively.

References

[1] Nirkin, Y., Keller, Y., Hassner, T.: FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7184–7193

[2] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Conference on Neural Information Processing Systems (NeurIPS). (2019)

[3] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)

[4] Alexander, O., Rogers, M., Lambeth, W., Chiang, M., Debevec, P.: The digital emily project: Photoreal facial modeling and animation. In: ACM SIGGRAPH 2009 Courses. SIGGRAPH '09, New York, NY, USA, Association for Computing Machinery (2009)

[5] Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 9459–9468

[6] Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2020)

[7] Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European conference on computer vision (ECCV). (2018) 818–833

[8] Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE (2018) 98–105

[9] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8789–8797

[10] Tripathy, S., Kannala, J., Rahtu, E.: Icface: Interpretable and controllable face reenactment using gans. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 3385–3394

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

[12] Wu, W., Zhang, Y., Li, C., Qian, C., Change Loy, C.: Reenactgan: Learning to reenact faces via boundary transfer. In: Proceedings of the European conference on computer vision (ECCV). (2018) 603–619

[13] Yao, G., Yuan, Y., Shao, T., Zhou, K.: Mesh Guided One-shot Face Reenactment using Graph Convolutional Networks. arXiv:2008.07783 [cs] (2020) Comment: 9 pages, 8 figures,accepted by MM2020.

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680

[15] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2019) 4401–4410

[16] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: Proceedings of International Conference on Learning Representations (ICLR) 2018. (2018) International Conference on Learning Representations, ICLR ; Conference date: 30-04-2018 Through 03-05-2018.

[17] Zhang, J., Zeng, X., Wang, M., Pan, Y., Liu, L., Liu, Y., Ding, Y., Fan, C.: Freenet: Multi-identity face reenactment. In: CVPR. (2020) 5326–5335

[18] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)

[19] Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Trans. Graph. **32** (2013)

[20] Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36** (2017) 1–12

[21] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat] (2014)

[22] Hou, X., Shen, L., Sun, K., Qiu, G.: Deep feature consistent variational autoencoder. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2017) 1133–1141

[23] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: β-vae: Learning basic visual concepts with a constrained variational framework. (2016)

[24] Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences (2020)

[25] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8798–8807

[26] Ekman, P., Friesen, W.V.: Facial action coding system: A technique for the measurement of facial movement (1978)

[27] Ekman, R.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, USA (1997)

[28] Bulat, A., Tzimiropoulos, G.: How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, IEEE (2017) 1021–1030

[29] Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE (2018) 59–66

[30] Baltrušaitis, T., Mahmoud, M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Volume 6., IEEE (2015) 1–6

[31] Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: IARPA Janus Benchmark - C: Face Dataset and Protocol. In: 2018 International Conference on Biometrics (ICB), Gold Coast, QLD, IEEE (2018) 158–165

[32] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1–11

[33] Xiang, S., Gu, Y., Xiang, P., He, M., Nagano, K., Chen, H., Li, H.: One-Shot Identity-Preserving Portrait Reenactment. arXiv:2004.12452 [cs] (2020) Comment: 29 pages, 14 figures.

[34] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep Face Recognition. In: Procedings of the British Machine Vision Conference 2015, Swansea, British Machine Vision Association (2015) 41.1–41.12