Finding Directions in GAN's Latent Space for Neural Face Reenactment

Stella Bounareli[1]
k2033759@kingston.ac.uk
Vasileios Argyriou[1]
vasileios.argyriou@kingston.ac.uk
Georgios Tzimiropoulos[2]
g.tzimiropoulos@qmul.ac.uk

[1] Kingston University, London, UK
[2] Queen Mary University of London, UK

Abstract

This paper is on face/head reenactment where the goal is to transfer the facial pose (3D head orientation and expression) of a target face to a source face. Previous methods focus on learning embedding networks for identity and pose disentanglement which proves to be a rather hard task, degrading the quality of the generated images. We take a different approach, bypassing the training of such networks, by using (fine-tuned) pre-trained GANs which have been shown capable of producing high-quality facial images. Because GANs are characterized by weak controllability, the core of our approach is a method to discover which directions in latent GAN space are responsible for controlling facial pose and expression variations. We present a simple pipeline to learn such directions with the aid of a 3D shape model which, by construction, already captures disentangled directions for facial pose, identity and expression. Moreover, we show that by embedding real images in the GAN latent space, our method can be successfully used for the reenactment of real-world faces. Our method features several favorable properties including using a single source image (one-shot) and enabling cross-person reenactment. Our qualitative and quantitative results show that our approach often produces reenacted faces of significantly higher quality than those produced by state-of-the-art methods for the standard benchmarks of VoxCeleb1 & 2. Source code is available at: https://github.com/StelaBou/stylegan_directions_face_reenactment

1 Introduction

This paper is on face/head reenactment where the goal is to transfer the facial pose, defined here as the rigid 3D face/head orientation *and* the deformable facial expression, of a target facial image to a source facial image. Such technology is the key enabler for creating high-quality digital head avatars which find a multitude of applications in telepresence, Augmented Reality/Virtual Reality (AR/VR), and the creative industries. Recently, and thanks to the advent of Deep Learning, there has been tremendous progress in the so-called neural face reenactment [7, 53, 59]. Despite the progress, synthesizing photorealistic face/head sequences is still considered a hard task with the quality of existing solutions being far from sufficient for the demanding applications mentioned above.

A major challenge that most prior methods [6, 7, 16, 58, 59, 60] have focused so far is how to achieve identity and pose disentanglement to both preserve the appearance and identity characteristics of the source face and successfully transfer to the facial pose of the target face. Training conditional generative models to produce embeddings with such disentanglement properties is known to be a difficult machine learning task [10, 25, 40], and this turns out to be a significant technical impediment for face reenactment too. Additionally, previous methods [58, 59] have approached reenactment using paired data during training. However, under such a paired setting it is not clear how to formulate cross-person reenactment [58].

In this work, we are taking a different path to neural face reenactment. A major motivation for our work is that unconstrained face generation using modern state-of-the-art GANs [21, 22, 23] has reached levels of unprecedented realism to the point that it is often impossible to distinguish real facial images from generated ones. Hence, the research question we would like to address in this paper is: *Can a pretrained GAN* [23] *be adapted for face reenactment?* A key challenge that needs to be addressed to this end, is that GANs come with no semantic parameters to control their output. Hence, in order to alleviate this, the core of our approach is a method to discover which *directions in the latent GAN space* are responsible for controlling facial pose and expression variations. Knowledge of these directions would directly equip the pretrained GAN with the desired reenactment capabilities. Inspired by Voynov and Babenko [60], we present a very simple pipeline to learn such directions with the help of a linear 3D shape model [14]. By construction, such a shape model captures disentangled directions for facial pose, identity and expression which is exactly what is required for reenactment. Moreover, a second key challenge that needs to be addressed is how to use the GAN for the manipulation of real-world images. Capitalizing on [44], we further show that by embedding real images in the GAN latent space, our pipeline can be successfully used for real face reenactment. Overall, we make the following **contributions**:

1. Instead of training conditional generative models [7, 59], we present a different approach to face reenactment by finding the directions in the latent space of a pretrained GAN (StyleGAN2 [23] fine-tuned on the VoxCeleb1 dataset) that are responsible for controlling the facial pose (i.e. rigid head orientation and expression), and show how these directions can be used for neural face reenactment on video datasets.
2. To achieve our goal, we describe *a simple pipeline* that is trained with the aid of a linear 3D shape model which already contains disentangled directions for facial shape in terms of pose, identity and expression. We further show that our pipeline can be trained with real images too by firstly embedding them into the GAN space, enabling the successful reenactment of real-world faces.
3. We show that our method features several favorable properties including using a single source image (one-shot), and enabling cross-person reenactment.
4. We perform several qualitative and quantitative comparisons with recent state-of-the-art reenactment methods, illustrating that our approach often produces reenacted faces of significantly higher quality for the standard benchmarks of VoxCeleb1 & 2 [8, 27].

2 Related work

Semantic face editing: There is a plethora of recent works that investigate the existence of interpretable directions in the GAN's latent space [19, 30, 31, 38, 39, 47, 48, 50, 56, 57]. These methods are able to successfully edit synthetic images, however, most of them do

not allow controllable editing and thus they cannot be applied on the face reenactment task. Voynov and Babenko [50], introduce an unsupervised method that is able to discover disentangled linear directions in the latent GAN space by jointly learning the directions and a classifier that learns to predict which direction is responsible for the image transformation. Our method is inspired by Voynov and Babenko [50], extending it in several ways to make it suitable for neural face reenactment. Additionally, there is a line of work that allows explicit controllable facial image editing [2, 10, 13, 15, 28, 40, 51]. Related to our method is StyleRig [43] which uses 3DMM parameters to control the generated images from a pretrained StyleGAN2. StyleRig's training pipeline is not end-to-end and significantly more complicated than ours, while in order to learn better disentangled directions, StyleRig requires different models for different attributes (e.g. pose, expression). In contrast, we learn all disentangled directions for face reenactment simultaneously and our model can successfully edit all attributes as well as edit only one attribute. Moreover, StyleRig is mainly applied on synthetic images, thus real image editing is not straightforward. Consequently, the aforementioned issues restrict StyleRig's applicability for real-world face reenactment, where various facial attributes change simultaneously. A follow-up work, PIE [42], focuses on inverting real images to enable editing using StyleRig [43]. However, their method is computationally expensive (10 min/image) which is prohibitive for video-based facial reenactment.

GAN inversion: GAN inversion aims to encode real images into the latent space of pretrained GANs [21, 23], which enables their editing using existing methods of synthetic image manipulation. Most of the inversion techniques [3, 9, 11, 34, 44, 52] train encoder-based architectures that focus on predicting the best latent codes that can generate images visually similar with the original ones and allow successful editing. The authors of [63] propose a hybrid approach which consists of learning an encoder followed by an optimization step on the latent space to refine the similarity between the reconstructed and real images. Additionally, Richardson *et al.* [34] introduce a method that tries to solve the editability-perception tradeoff, while recently in [35], the authors propose fine-tuning the generator to better capture appearance features, so that the inverted images resemble the original ones.

Neural face/head reenactment: Face reenactment is a non-trivial task, as it requires wide generalization across identity and facial pose. Many of the proposed methods rely on facial landmark information [16, 18, 45, 46, 54, 58, 59, 61]. The authors of [59] propose a one-shot face reenactment method driven by landmarks, which decomposes an image on pose-dependent and pose-independent components. A limitation of landmark based methods is that landmarks preserve identity information, thus impeding their applicability on cross-subject face reenactment [7]. In [7] the authors perform face reenactment by learning pose and identity embeddings using two different encoders. Additionally, warping-based methods [33, 41, 53, 55] synthesize the reenacted images based on the motion of the driving faces. Those methods produce realistic results, however they suffer from visual artifacts and pose mismatch especially in large head pose variations. Finally, the authors of [26] propose a two-step architecture that aims to disentangle the spatial and style components of an image that leads to better preservation of the source identity, while in [12] the authors present a GAN-based method conditioned on a 3D face representation [64].

To summarize, all the aforementioned methods rely on training *conditional* generative models on large paired datasets in order to learn facial descriptors with disentanglement properties. In this paper, we propose a new approach for face reenactment that learns disentangled directions in the latent space of a pretrained StyleGAN2 on the VoxCeleb dataset. We show that the discovery of meaningful and disentangled directions that are responsible for controlling the facial pose can be used for high quality self- and cross-identity reenactment.

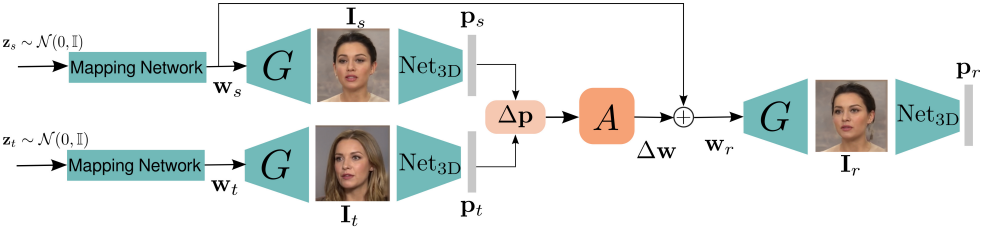

Figure 1: **Overview of our method:** Given a pair of source $\mathbf{I}_s$ and target $\mathbf{I}_t$ images, we calculate the facial pose parameter vectors $\mathbf{p}_s$ and $\mathbf{p}_t$ using the Net_{3D} network, respectively. The matrix of directions $\mathbf{A}$ is trained such that, given the shift $\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p}$, the reenacted image $\mathbf{I}_r$ generated using the latent code $\mathbf{w}_r = \mathbf{w}_s + \Delta\mathbf{w}$, illustrates the facial pose of the target face, while maintaining the identity of the source face.

3 Method

Our method consists of three parts detailed in the following subsections. In Section 3.1, we show how to find the facial pose directions in the latent GAN space and use them for face/head reenactment. In Section 3.2, we describe how to extend our method to handle real facial images. Finally, in Section 3.3, we investigate how better results can be obtained by fine-tuning on paired video data.

3.1 Finding the reenactment latent directions

The generator G takes as input latent codes $\mathbf{z} \sim \mathcal{N}(0,\mathbb{I}) \in \mathbb{R}^{512}$ and generates images $\mathbf{I} = G(\mathbf{z}) \in \mathbb{R}^{3\times256\times256}$. StyleGAN2 firstly maps the latent code $\mathbf{z}$ into the intermediate latent code $\mathbf{w} \in \mathbb{R}^{512}$ using a series of fully connected layers. Then, the latent code $\mathbf{w}$ is fed into each convolution layer of StyleGAN2's generator. This mapping enforces the disentangled representation of StyleGAN2 [23]. In order to fairly compare our work with previous face reenactment methods, we need a StyleGAN2 model that generates synthetic images that resemble the distribution of the VoxCeleb dataset [27]. This dataset is more diverse compared to Flickr-Faces-HQ (FFHQ) dataset [21] in terms of head poses and expressions, providing the ability to find more meaningful directions for face reenactment (e.g. GANs pretrained on FFHQ do not account for roll changes in head pose). Having a pretrained StyleGAN2 generator on FFHQ dataset, we use the method of Karras *et al.* [22] to fine-tune the generator on VoxCeleb. We note that we do not finetune the generator under any reenactment objective. Our generator on VoxCeleb is able to produce synthetic images with random identities (different from the identities of VoxCeleb) that follow the distribution of VoxCeleb dataset in terms of head poses and expressions.

A facial shape $\mathbf{s} \in \mathbb{R}^{3N}$ (N is the number of vertices) can be written in terms of a linear 3D shape model as:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}_i\mathbf{p}_i + \mathbf{S}_e\mathbf{p}_e, \tag{1}$$

where $\bar{\mathbf{s}}$ is the mean 3D shape, $\mathbf{S}_i \in \mathbb{R}^{3N\times m_i}$ and $\mathbf{S}_e \in \mathbb{R}^{3N\times m_e}$ are the PCA bases for identity and expression, and $\mathbf{p}_i$ and $\mathbf{p}_e$ are the corresponding identity and expression coefficients. Moreover, we denote as $\mathbf{p}_\theta \in \mathbb{R}^3$ the rigid head orientation defined by the three Euler angles

(yaw, pitch, roll). For reenactment, we are interested in manipulating head orientation and expression, so our facial pose parameter vector is $\mathbf{p} = [\mathbf{p}_\theta, \mathbf{p}_e] \in \mathbb{R}^{3+m_e}$. We note that all PCA shape bases are orthogonal to each other, and hence they capture disentangled variations of identity and expression. They are calculated in a frontalized reference frame, thus they are also disentangled with head orientation. These bases can be also interpreted as directions in the shape space. We propose to learn similar directions in the GAN latent space.

In particular, we propose to associate a change $\Delta\mathbf{p}$ in facial pose, with a change $\Delta\mathbf{w}$ in the (intermediate) latent GAN space so that the two generated images $G(\mathbf{w})$ and $G(\mathbf{w}+\Delta\mathbf{w})$ differ only in pose by the same amount $\Delta\mathbf{s}$ induced by $\Delta\mathbf{p}$. If the directions sought in the GAN latent space are assumed to be linear [29], this implies the following linear relationship:

$$\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is a matrix, the columns of which (i.e. d_{in}) represent the directions in GAN latent space. In our case, $d_{\text{in}} = (3+m_e)$ and $d_{\text{out}} = N_l \times 512$, where N_l is the number of the generators layers we opt to apply shift changes.

Training pipeline: The matrix $\mathbf{A}$ is unknown so we propose the simple pipeline of Fig. 1 to estimate it: in particular, we sample two random latent codes $\mathbf{z}_s$ and $\mathbf{z}_t$ (s, t for source and target, respectively) and pass them through the generator G. The two generated images $\mathbf{I}_s = G(\mathbf{z}_s)$ and $\mathbf{I}_t = G(\mathbf{z}_t)$ are fed into the pre-trained Net_{3D} which estimates the corresponding pose parameter vectors, $\mathbf{p}_s$ and $\mathbf{p}_t$, respectively. Using Eq. 2, we compute $\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p} = \mathbf{A}(\mathbf{p}_t - \mathbf{p}_s)$ and $\mathbf{w}_r = \mathbf{w}_s + \Delta\mathbf{w}$. From $\mathbf{w}_r$ our pipeline generates the reenacted facial image $\mathbf{I}_r = G(\mathbf{w}_r)$, which depicts the source face in the facial pose of the target. The only trainable quantity in the above pipeline is the matrix $\mathbf{A}$ containing the unknown directions in GAN latent space. We propose to learn it in a self-supervised manner.

Training losses: Our pipeline is trained by minimizing the following total loss:

$$\mathcal{L} = \lambda_r\mathcal{L}_r + \lambda_{id}\mathcal{L}_{id} + \lambda_{per}\mathcal{L}_{per}, \tag{3}$$

where $\lambda_r = 1, \lambda_{id} = 10$ and $\lambda_{per} = 10$. The *reenactment loss* $\mathcal{L}_r$ ensures successful facial pose transfer from target to source and it is defined as: $\mathcal{L}_r = \mathcal{L}_{sh} + \mathcal{L}_{eye} + \mathcal{L}_{mouth}$. $\mathcal{L}_{sh} = \|\mathbf{S}_r - \mathbf{S}_{gt}\|_1$ is the shape loss, where $\mathbf{S}_r$ is the 3D shape of the reenacted image and $\mathbf{S}_{gt}$ is the reconstructed *ground-truth* 3D shape calculated using Eq. 1 with the identity coefficients $\mathbf{p}_i$ of the source face and the coefficients $\mathbf{p}_e$ of the target face. Additionally, to enhance the expression transfer we calculate the eye loss $\mathcal{L}_{eye}$ (the mouth loss $\mathcal{L}_{mouth}$ is computed in a similar fashion) which compares the inner distances between the eye landmark pairs of upper and lower eyelids between the reenacted and reconstructed ground-truth shapes (see supplementary for detailed explanation of eye $\mathcal{L}_{eye}$ and mouth $\mathcal{L}_{mouth}$ losses). Additionally, $\mathcal{L}_{id}$ is an *identity loss* based on the cosine similarity of features extracted from the source $\mathbf{I}_s$ and the reenacted $\mathbf{I}_r$ image using the face recognition network of [9]. This loss imposes that the identity of the source is preserved in the reenacted image. Finally, we also found that better image quality is obtained if we additionally use $\mathcal{L}_{per}$ which is the standard *perceptual loss* of [20].

Training details: We estimate the distribution of each element of the pose parameters $\mathbf{p}$ by randomly generating 10K images and computing their corresponding $\mathbf{p}$ vectors. Using the estimated distributions, during training, we re-scale each element of $\mathbf{p}$ from its original range to a common range $[-a, a]$. Furthermore, to increase the disentanglement of the learned directions of our method, we follow a training strategy where for 50% of the training samples we reenact only one attribute by using $\Delta\mathbf{p} = [0, ..., \varepsilon_i, ...0]$, where ε_i is sampled from a uniform distribution $\mathcal{U}[-a, a]$.

3.2 Real image reenactment

So far our method is able to transfer facial pose from a source facial image to a target only for synthetically generated images. To extend our method to work with real images, in this section, we propose (a) to use a pipeline for inverting the images back to the latent code space of StyleGAN2, and (b) a mixed training approach for discovering the latent directions.
Real image inversion: Ideally, the inversion method should produce latent codes that can generate facial images identical with the original ones *and* enable image editing without producing visual artifacts. Although satisfying both requirements is challenging [3, 34, 44], we found that the following pipeline produces excellent results for the purposes of our goal (i.e. face/head reenactment). During training, we employ an encoder based method [44] to invert the real images into the $\mathcal{W}+$ space [1]. However, directly using the inverted latent codes $\mathbf{w}^{inv}$ does not produce satisfactory reenactment results. This happens because the latent codes obtained from inversion, may present a domain gap from the latent codes of synthetic images. To alleviate this, we propose a mixed data approach for training the pipeline of Section 3.1: specifically, we first invert the extracted frames from the VoxCeleb dataset, and during training, at each iteration (i.e. for each batch) we use 50% random latent codes $\mathbf{w}$ and 50% embedded latent codes $\mathbf{w}^{inv}$.

The inverted images using the encoder based method [44] might still be missing some identity details. To alleviate this, only during inference, we use an additional optimization step [35] that lightly optimizes the generator, so that the newly generated image more closely resembles the original one. Note that this step does not affect the calculation of $\mathbf{w}^{inv}$ and is used only during inference to obtain a higher quality inversion. We perform the optimization for 200 steps and only on the source frame of each video.

3.3 Fine-tuning on paired video data

Our method so far has been trained with unpaired static facial images. This has at least two advantages: (a) it enables training with a very large number of identities, and (b) seems more suitable for cross-person reenactment. However, additional improvements enabled by the optimization of additional losses can be obtained by further training on paired data from VoxCeleb1. Compared to training from scratch on video data, as in most previous methods (e.g. [7, 58, 59]), we believe that our approach offers a more balanced option that combines the best of both worlds: training with unpaired static images and fine-tuning with paired video data. From each video of our training set, we randomly sample a source and a target face that have the same identity but different pose and expression. Consequently, we minimize the following loss function $\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_{id} \mathcal{L}_{id} + \lambda_{per} \mathcal{L}_{per} + \lambda_{pix} \mathcal{L}_{pix}$, where $\mathcal{L}_r$ is the same reenactment loss defined in Section 3.1, $\mathcal{L}_{id}$ and $\mathcal{L}_{per}$ are the identity and perceptual losses, however this time calculated between the reenacted $\mathbf{I}_r$ and the target image $\mathbf{I}_t$ and $\mathcal{L}_{pix}$ is a pixel-wise $L1$ loss between the reenacted and target images.

4 Experiments

In this section, we present qualitative and quantitative results and comparisons of our method with recent state-of-the-art approaches. The bulk of our results and comparisons, reported in Section 4.1, are on self- and cross-person reenactment on the VoxCeleb1 [27] dataset. Comparisons with state-of-the-art on the VoxCeleb2 [8] test set released by [58] are provided

in the supplementary material. Finally, in Section 4.2 we report ablation studies on the various design choices of our method.

Implementation details: We fine-tune StyleGAN2 on the VoxCeleb1 dataset with 256×256 image resolution and we train the encoder of [44] for real image inversion. The 3D shape model we use is DECA [14]. For our training procedure, we only learn a matrix of directions $\mathbf{A} \in \mathbb{R}^{(N_l \times 512) \times k}$ where $k = 3 + m_e, m_e = 12$ and $N_l = 8$. We train three matrices of directions: the first one is on synthetically generated images (Section 3.1), while the second one on mixed real and synthetic data (Section 3.2). Finally, as described in Section 3.3, we obtain a third model by fine-tuning the second one on paired data. For training, we used the Adam optimizer [24] with constant learning rate 10^-4. We train our models for 20K iterations with a batch size of 12 on synthetic and real images. Fine-tuning is performed on real paired images for 150K iterations. All models are implemented in PyTorch [32].

4.1 Comparison with state-of-the-art on VoxCeleb

Herein, we compare the performance of our method against the state-of-the-art in face reenactment on VoxCeleb1 [27]. We conduct two types of experiments, namely self- and cross-person reenactment. For evaluation purposes, we use both the video data provided by [58] and the original test-set of VoxCeleb1. We note that there is no overlap between the train and test identities and videos. Similar comparisons on the VoxCeleb2 [8] test set released by [58] are provided in the supplementary material. We compare our method quantitatively and qualitatively with six methods: X2Face [55], FOMM [41], Fast bi-layer [59], Neural-Head [7], LSR [26] and PIR [33]. For X2Face [55], FOMM [41] and PIR [33], we use the pretrained (by the authors) model on VoxCeleb1. For Fast bi-layer [59], Neural-Head [7] and LSR [26] we also use the pretrained (by the authors) models on VoxCeleb2 [8]. For fair comparison with the methods of Neural-Head [7] and LSR [26], we evaluate their model under the one-shot setting.

Quantitative comparisons: We report seven different metrics. We compute the Learned Perceptual Image Path Similarity (LPIPS) [62] to measure the perceptual similarities, and to quantify identity preservation we compute the cosine similarity (CSIM) of ArcFace [9] features. Moreover, we measure the quality of the reenacted images using the Frechet-Inception Distance (FID) metric [17], while we also report the Fréchet Video Distance (FVD) [49] metric that measures both the video quality and the temporal consistency of the generated videos. To quantify the facial pose transfer, we calculate the normalized mean error (NME) between the predicted landmarks in the reenacted and target images. We use [6] for landmark estimation, and we calculate the NME by normalizing it with the square root of the ground truth face bounding box and multiplying it by 10^3. We further evaluate pose transfer by calculating the mean $L1$ distance of the head pose (Pose) in degrees and the mean $L1$ distance of the expression coefficients $\mathbf{p}_e$ (Exp.).

In Tables 1 and 2, we report the quantitative results for self and cross-subject reenactment, respectively. For self-reenactment, we combine the original test set of VoxCeleb1 [27] and the test set provided by [58]. For cross-subject reenactment, we randomly select 200 video pairs from the small test set of [58]. In self-reenactment, all metrics are calculated between the reenacted and the target faces, while in cross-subject reenactment, CSIM is calculated between the source and the reenacted faces and pose/expression error between the target and the reenacted faces. As a result, the values of CSIM in cross-subject reenactment are expected to be lower. Regarding self-reenactment, X2Face and PIR have a higher value on CSIM, however we argue that this is due to their warping-based technique which enables

better reconstruction of the background and other identity characteristics. Importantly, this quantitative result is accompanied by poor qualitative results (e.g. see Fig. 2). Additionally, regarding pose transfer, we achieve similar results on NME and Pose error with Fast Bi-layer [59] and LSR [26] (their methods are trained on VoxCeleb2 which contains 5× more identities) and we outperform all methods on expression transfer. Finally, our results on FID and FVD metric confirm that the quality of our generated videos resembles the quality of VoxCeleb dataset. Cross-subject reenactment is more challenging, as source and target faces have different identities. In this case, it is important to maintain the source identity characteristics without transferring the target ones. In Table 2, the high CSIM value for FOMM is not accompanied by good qualitative results as shown in Fig. 2, where FOMM, in most cases, is not able to transfer the target head pose (hence their method achieves higher CSIM but poor pose transfer). Additionally, we achieve better head pose and expression transfer compared to all other methods.

Method	CSIM	LPIPS	FID	FVD	NME	Pose	Exp.
X2Face [55]	0.70	0.13	35.5	409	17.8	1.5	0.90
FOMM [41]	0.65	0.14	35.6	402	34.1	5.0	1.3
Fast Bi-layer [59]	0.64	0.23	52.8	634	**13.2**	1.1	0.80
Neural-Head [7]	0.40	0.22	98.4	587	15.5	1.3	0.90
LSR [26]	0.59	0.13	45.7	464	17.8	**1.0**	0.75
PIR [33]	**0.71**	0.12	57.2	414	18.2	1.86	0.94
Ours	0.66	**0.11**	**35.0**	**345**	14.1	1.1	**0.68**

Table 1: Quantitative results on self-reenactment. The results are reported on the combined original test set of VoxCeleb1 [27] and the test set released by [58]. For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓).

Method	CSIM	Pose	Exp.
X2Face [55]	0.57	2.2	1.5
FOMM [41]	**0.73**	7.7	2.0
Fast Bi-layer [59]	0.48	1.5	1.3
Neural-Head [7]	0.36	1.7	1.6
LSR [26]	0.50	1.4	1.2
PIR [33]	0.62	2.2	1.4
Ours	0.63	**1.2**	**1.0**

Table 2: Quantitative results on cross-subject reenactment. The results are reported on 200 video pairs from the test set of [58]. For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓).

Qualitative comparisons: Unfortunately, quantitative comparisons alone are insufficient to capture the quality of reenactment. Hence, we opt for qualitative visual comparisons *in multiple ways*: (a) results in Fig. 2, (b) in the supplementary material, we provide more results in self and cross-subject reenactment both on VoxCeleb1 and VoxCeleb2 datasets, and (c) we also provide *all* self-reenactment videos for the small test set of VoxCeleb1 (and VoxCeleb2) provided in [58] and cross-reenactment videos of *randomly selected* identities (providing all possible pairs is not possible). As we can see from Fig. 2 and the videos provided in the supplementary material, our method provides, for the majority of videos, the highest reenact-

Figure 2: Qualitative results and comparisons for self (top three rows) and cross-subject reenactment (last three rows) on VoxCeleb1. The first and second columns show the source and target faces. Our method preserves the appearance and identity characteristics (e.g. face shape) of the source face significantly better and also better captures fine-grained expression details such as closed eyes (2^{nd} and 5^{th} row).

ment quality including accurate transfer of pose and expression and, significantly enhanced identity preservation compared to all other methods. Importantly one great advantage of our method on cross-subject reenactment, as shown in Fig. 2, is that it is able to reenact the source face with minimal identity leakage (e.g facial shape) from the target face, in contrast to landmark-based methods such as Fast Bi-layer [59]. Finally, to show that our method is able to generalise well on other facial video datasets, we provide additional results on the FaceForensics [36] and 300-VW [37] datasets in the supplementary material.

4.2 Ablation studies

We perform several ablation tests to (a) measure the impact of the identity and perceptual losses, and the additional shape losses for the eyes and mouth, (b) validate our trained models on synthetic, mixed and paired images, and (c) assess the use of optimization in G during inference. For (a), we perform experiments on synthetic images with and without the identity and perceptual losses. To evaluate the models, we randomly generate $5K$ pairs of synthetic images (source and target) and reenact the source image with the pose and expression of the target. As shown in Table 3, the incorporation of the identity and perceptual losses is crucial to isolate the latent space directions that strictly control the head pose and expression characteristics without affecting the identity of the source face. In a similar fashion, in Table 3, we show the impact of the additional shape losses, namely the eye $\mathcal{L}_{eye}$ and mouth $\mathcal{L}_{mouth}$ losses. As shown, omitting these losses leads to higher pose and expression error.

For (b), we evaluate the three different training schemes, namely synthetic only (Section

Method	CSIM ↑	Pose ↓	Exp. ↓
Ours w/ $\mathcal{L}_{id} + \mathcal{L}_{per}$	**0.52**	2.4	1.2
Ours w/o $\mathcal{L}_{id} + \mathcal{L}_{per}$	0.42	2.5	1.2
Ours w/ $\mathcal{L}_{eye} + \mathcal{L}_{mouth}$	0.52	**2.4**	**1.2**
Ours w/o $\mathcal{L}_{eye} + \mathcal{L}_{mouth}$	0.52	2.6	1.5

Table 3: Ablation study on the impact of the identity $\mathcal{L}_{id}$ and perceptual $\mathcal{L}_{per}$ losses, and on the impact of eye $\mathcal{L}_{eye}$ and mouth $\mathcal{L}_{mouth}$ losses. CSIM is calculated between the source and the reenacted images which are on different poses.

Method	CSIM ↑	Pose ↓	Exp. ↓
Ours *synthetic*	0.60	1.7	1.1
Ours *real & synthetic*	0.63	1.6	1.1
Ours *paired*	**0.66**	**1.1**	**0.8**
w/o optim.	0.45	1.4	1.0
w/ optim. in G	**0.66**	**1.1**	**0.8**

Table 4: Ablation studies on self-reenactment using three different models: (a) trained on synthetic images, (b) trained on both synthetic and real images, and (c) fine-tuned on paired data, and on self reenactment with and without optimization of the generator *G*.

3.1), mixed synthetic-real (Section 3.2), and mixed synthetic-real fine-tuned with paired data (Section 3.3) for self-reenactment. The results, shown in Table 4 (first three rows), illustrate the impact of each of these training schemes with the one using paired data providing the best results as expected. Finally, regarding (c), we report results of self-reenactment, without any optimization and with optimization of *G*. As shown in Table 4 (last two rows), the optimization of *G* improves our results (as expected) especially regarding the identity preservation (CSIM). Moreover, with this ablation study we show that our learned directions do not get adversely affected by the optimization step, as both Pose and Expression errors are improving. We note that, to evaluate the different models in Table 4, we use the small test set of [53].

5 Discussion and conclusions

This paper introduces a new approach to neural head/face reenactment using a 3D shape model to learn disentangled directions of facial pose in latent GAN space. The approach comes with specific advantages such as the use of powerful pre-trained GANs and 3D shape models which have been studied and developed for several years in computer vision and machine learning. These advantages however, in some cases, can turn into disadvantages. For example, we observed that in extreme source and target poses the reenacted images have some visual artifacts. We attribute this to the GAN inversion process, which renders the inverted latent codes in extreme head poses less editable. Finally, we acknowledge that although face reenactment can be used in a variety of applications such as art, video conferencing etc., it can also be applied for malicious purposes. However, our work does not amplify any of the potential dangers that already exist.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022.

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[7] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, 2020.

[8] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[10] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

[11] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022.

[12] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: Video-and-audio-driven talking head synthesis. *arXiv preprint arXiv:2012.08261*, 2020.

[13] Ricard Durall Lopez, Jireh Jam, Dominik Strassel, Moi Hoon Yap, and Janis Keuper. Facialgan: Style transfer and attribute manipulation on synthetic faces. In *[32nd British Machine Vision Conference]*, pages 1–14, 2021.

[14] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[15] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J. Black, and Timo Bolkart. GIF: generative interpretable faces. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 868–878. IEEE, 2020.

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

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

[18] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–650, 2022.

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *European Conference on Computer Vision (ECCV)*, 2020.

[26] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13829–13838, 2021.

[27] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[28] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020.

[29] Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. *arXiv preprint arXiv:2107.11186*, 2021.

[30] James Oldfield, Markos Georgopoulos, Yannis Panagakis, Mihalis A. Nicolaou, and Ioannis Patras. Tensor component analysis for interpreting the latent space of gans. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 222, 2021.

[31] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. *arXiv preprint arXiv:2206.00048*, 2022.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[33] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[34] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

[36] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[37] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.

[38] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021.

[39] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[40] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. *arXiv preprint arXiv:2101.02477*, 2021.

[41] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, 2019.

[42] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

[43] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[44] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[45] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3385–3394, 2020.

[46] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2021.

[47] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. Warpedganspace: Finding non-linear rbf paths in gan latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6393–6402, 2021.

[48] Christos Tzelepis, James Oldfield, Georgios Tzimiropoulos, and Ioannis Patras. Contraclip: Interpretable gan generation driven by pairs of contrasting sentences. *arXiv preprint arXiv:2206.02104*, 2022.

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

[50] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020.

[51] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Cross-domain and disentangled face manipulation with 3d guidance. *arXiv preprint arXiv:2104.11228*, 2021.

[52] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022.

[53] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[54] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022.

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

[56] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12177–12185, 2021.

[57] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13789–13798, 2021.

[58] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019.

[59] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020.

[60] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12757–12764, 2020.

[61] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5326–5335, 2020.

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

[63] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European Conference on Computer Vision*, pages 592–608. Springer, 2020.

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