International Journal of Computer Vision (2024) 132:3324–3354
https://doi.org/10.1007/s11263-024-02018-6

One-Shot Neural Face Reenactment via Finding Directions in GAN's Latent Space

Stella Bounareli[1] · Christos Tzelepis[2] · Vasileios Argyriou[1] · Ioannis Patras[3] · Georgios Tzimiropoulos[3]

Received: 31 March 2023 / Accepted: 27 January 2024 / Published online: 13 March 2024

Abstract

In this paper, we present our framework for neural face/head reenactment whose goal is to transfer the 3D head orientation and expression of a target face to a source face. Previous methods focus on learning embedding networks for identity and head pose/expression 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 head pose and expression variations. We present a simple pipeline to learn such directions with the aid of a 3D shape model which, by construction, inherently captures disentangled directions for head 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. Extensive qualitative and quantitative results show that our approach typically produces reenacted faces of notably higher quality than those produced by state-of-the-art methods for the standard benchmarks of VoxCeleb1 & 2.

Keywords Neural face reenactment · Generative adversarial networks (GANs) · Image synthesis · Image editing

1 Introduction

Neural face reenactment aims to transfer 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 that find a multitude of applications in telepresence, Augmented Reality/Virtual Reality (AR/VR), and the creative industries. Recently, thanks to the advent of Deep Learning, Neural Face Reenactment has seen remarkable progress (Burkov et al., 2020; Meshry et al., 2021; Wang et al., 2021b; Zakharov et al., 2020). In spite of this, synthesizing photorealistic face/head sequences remains a challenging problem with the quality of existing solutions being far from sufficient for the demanding aforementioned applications.

A major challenge that most prior works (Bao et al., 2018; Burkov et al., 2020; Ha et al., 2020; Zakharov et al., 2019, 2020; Zeng et al., 2020) have focused on is how to achieve identity and head pose/expression disentanglement to both preserve the appearance and identity characteristics of the source face and successfully transfer the head pose and the expression of the target face. A recent line of research relies on training conditional Generative Adversarial Networks (GANs) (Deng et al., 2020; Kowalski et al., 2020; Shoshan et al., 2021) in order to produce disentangled embeddings and control the generation process. However, such methods mainly focus on synthetic image generation, rendering reen-

Communicated by Guang Yang.

✉ Stella Bounareli
k2033759@kingston.ac.uk

Christos Tzelepis
Christos.Tzelepis@city.ac.uk

Vasileios Argyriou
vasileios.argyriou@kingston.ac.uk

Ioannis Patras
i.patras@qmul.ac.uk

Georgios Tzimiropoulos
g.tzimiropoulos@qmul.ac.uk

[1] School of Computer Science and Mathematics, Kingston University, 55-59 Penrhyn Road, London KT1 2EE, UK

[2] Department of Computer Science, University of London, Northampton Square, EC1 0HB London, UK

[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK

actment on real faces challenging. Another line of works (Zakharov et al., 2019, 2020) rely on training with paired data (i.e., source and target facial images of the same identity), leading to poor cross-person face reenactment.

In this work, we propose a neural face reenactment framework that addresses the aforementioned limitations of state-of-the-art (SOTA), motivated by the remarkable ability of modern pre-trained GANs (e.g., StyleGAN Karras et al. 2019; 2020a; 2020b) in generating realistic and aesthetically pleasing faces, often indistinguishable from real ones. The research question we address in this paper is: *Can a pre-trained GAN be adapted for face reenactment?* A key challenge that needs to be addressed to this end is the absence of any inherent semantic structure in the latent space of GANs. In order to gain control over the generative process, inspired by Voynov and Babenko (2020), we propose to learn a set of latent direction (i.e., direction vectors in the GAN's latent space) that are responsible for controlling head pose and expression variations in the generated facial images. Knowledge of these directions directly equips the pre-trained GAN with the ability of controllable generation in terms of head pose and expression, allowing for effective face reenactment. Specifically, in this work we present a simple pipeline to learn such directions leveraging the ability of a linear 3D shape model (Feng et al., 2021) in capturing disentangled directions for head pose, identity, and expression, which is crucial towards effective neural face reenactment. Moreover, another key challenge that needs to be addressed is how to use the GAN for the manipulation of real-world images. Capitalizing on Tov et al. (2021), 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 from-scratch conditional generative models (Burkov et al., 2020; Zakharov et al., 2020), we present a novel approach to face reenactment by finding the directions in the latent space of a pre-trained GAN (i.e., StyleGAN2 Karras et al. Karras et al. (2020b) fine-tuned on the VoxCeleb1 dataset) that are responsible for controlling the rigid head orientation and expression, and show how these directions can be used for neural face reenactment on video datasets.
2. We present *a simple pipeline* that is trained with the aid of a linear 3D shape model (Feng et al., 2021), that is inherently equipped with disentangled directions for facial shape in terms of head pose, identity and expression. We further show that our pipeline can be trained with real images by firstly embedding them into the GAN space, allowing for effective reenactment of real-world faces.
3. We show that our method features several favorable properties including requiring 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 typically produces reenacted faces of notably higher quality for the standard benchmarks of VoxCeleb1 & 2 (Chung et al., 2018; Nagrani et al., 2017).

Compared to our previous work in Bounareli et al. (2022), this paper further investigates the real image inversion step and proposes a joint training scheme (Sect. 3.4) that eliminates the need for the optimization step during inference, described in Sect. 3.2, resulting in a more efficient inference process and better quantitative and qualitative results. The proposed joint training scheme efficiently addresses existing visual artifacts on the reenacted images caused by large head pose variations between the source and target faces, resulting in improved overall image quality. We qualitatively and quantitatively show that by jointly learning the real image inversion encoder and the directions, our method achieves compelling results without the need of one-shot fine-tuning during inference. Finally, to further improve the visual quality of the reenacted images in terms of crucial (for the purpose of face reenactment) background and identity characteristics, we propose to further fine-tune the feature space $\mathcal{F}$ of StyleGAN2 (Sect. 3.5).

2 Related Work

2.1 Semantic Face Editing

Several recent works (Härkönen et al., 2020; Oldfield et al., 2021, 2023; Shen & Zhou, 2021; Shen et al., 2020; Voynov & Babenko, 2020; Tzelepis et al., 2021, 2022; Yang et al., 2021; Yao et al., 2021) study the existence of directions/paths in the latent space of a pre-trained GAN in order to perform editing (i.e., with respect to specific facial attributes) on the generated facial images. Voynov and Babenko (2020) introduced an unsupervised method that optimizes a set of vectors in the GAN's latent space by learning to distinguish (using a "reconstructor" network) the image transformations caused by distinct latent directions. This leads to the discovery of a set of "interpretable", but not "controllable", directions—i.e., the optimized latent directions cannot be used for controllable (in terms of head pose and expression) facial editing and, thus, for face reenactment. Our method is inspired by the work of Voynov and Babenko (2020), extending it in several ways to make it suitable for neural face reenactment. Another line of recent works allows for explicit controllable facial image editing (Abdal et al., 2021; Deng et al., 2020;

Durall Lopez et al., 2021; Ghosh et al., 2020; Nitzan et al., 2020; Shoshan et al., 2021; Wang et al., 2021a). However, these methods mostly rely on synthetic image editing rather than performing face reenactment on real video data. A work that is related to our framework is StyleRig (Tewari et al., 2020b), which uses 3D Morphable Model's (3DMM) (Blanz & Vetter, 1999) parameters to control the generated images from a pre-trained StyleGAN2 (Karras et al., 2020b). However, by contrast to our method, StyleRig's training pipeline is not end-to-end and is significantly more complicated than ours, while in order to learn better disentangled directions, StyleRig requires the training of distinct models for different attributes (e.g., head pose and expression). This, along with the fact that StyleRig operates mainly on synthetic images, poses a notable restriction towards real-world face reenactment, where various facial attributes change simultaneously. By contrast, we propose to learn all disentangled directions for face reenactment simultaneously, allowing in this way for the effective editing of all, a subset, or a single attribute, whilst we optimize our framework on real faces as well. A follow-up work, PIE (Tewari et al., 2020a), focuses on inverting real images to enable editing using StyleRig (Tewari et al., 2020b). However, their method is computationally expensive (10 min/image) which is prohibitive for video-based facial reenactment. By contrast, we propose a framework that effectively and efficiently performs face reenactment (0.13 sec/image).

2.2 GAN Inversion

GAN inversion methods aim to encode real images into the latent space of pre-trained GANs (Karras et al., 2019, 2020b), allowing for subsequent editing using existing methods of synthetic image manipulation. The major challenge in the GAN inversion problem comprises of the so called "editability-perception" trade-off; that is, finding a sweet spot between faithful reconstruction of the real image and the editability of the corresponding latent code. The majority of recent inversion methods (Alaluf et al., 2021, 2022; Dinh et al., 2022; Richardson et al., 2021; Tov et al., 2021; Wang et al., 2022a) train encoder-based architectures that focus on predicting the latent codes **w** that best reconstruct the original (real) images and that allow for subsequent editing. Zhu et al. (2020) 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. Richardson et al. (2021) introduce a method that aims to improve the "editability-perception" trade-off, while recently (Roich et al., 2021) propose to fine-tune the generator to better capture/transfer appearance features.

The aforementioned works typically perform inversion onto the $\mathcal{W}+$ latent space of StyleGAN2. However, Parmar et al. (2022) have shown that $\mathcal{W}+$ is not capable of fully reconstructing the real images. Specifically, details such as the background, the hair style or facial accessories i.e., hats and glasses, cannot be inverted with high fidelity. A recent line of works (Alaluf et al., 2022; Bai et al., 2022; Wang et al., 2022a; Yao et al., 2022a) propose to mitigate this by investigating more expressive spaces of StyleGAN2 (such as the feature space $\mathcal{F} \in \mathbb{R}^{h\times w\times c}$ Kang et al. 2021) to perform real image inversion. Although such methods are able to produce high quality reconstructions, their ability to accurately edit the inverted images is limited. Especially when changing the head pose, such methods tend to produce many visual artifacts (Fig. 25). In order to balance between expressive invertibility and editing performance, the authors of Parmar et al. (2022) (SAM) propose to fuse different spaces, i.e., the $\mathcal{W}+$ latent space and the feature space $\mathcal{F} = \{\mathcal{F}_4, \mathcal{F}_6, \mathcal{F}_8, \mathcal{F}_{10}\}$, where each one corresponds to a different feature layer of StyleGAN2 (Karras et al., 2020b). In more detail, they propose to break the facial images into different segments (background, hat, glasses etc.) and choose the most suitable space to invert each segment, leveraging the editing capabilities of the $\mathcal{W}+$ latent space and the reconstruction quality of the feature space $\mathcal{F}$. However, when performing global editings, i.e., changing the head pose orientation, SAM (Parmar et al., 2022) results in notable visual artifacts, in contrast to our method, as will be shown in the experimental section.

2.3 Neural Face Reenactment

Neural face reenactment poses a challenging problem that requires strong generalization ability across many different identities and a large range of head poses and expressions. Many of the proposed methods rely on facial landmark information (Ha et al., 2020; Hsu et al., 2022; Tripathy et al., 2020, 2021; Wang et al., 2022b; Zakharov et al., 2019, 2020; Zhang et al., 2020). Specifically, Zakharov et al. (2020) propose an 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 (Burkov et al., 2020). In order to mitigate this limitation (Hsu et al., 2022) propose to use an ID-preserving Shape Generator (IDSG) that transforms the target facial landmarks so that they preserve the identity, i.e. facial shape, of the source image. Additionally, several methods (Doukas et al., 2021; Ren et al., 2021; Yang et al., 2022; Yao et al., 2020) rely on 3D shape models to remove the identity details of the driving images. Warping-based methods (Doukas et al., 2021; Ren et al., 2021; Siarohin et al., 2019; Wang et al., 2021b; Wiles et al., 2018; Yang et al., 2022) synthesize the reenacted images based on the motion of the driving faces. Specifically, HeadGAN (Doukas et al., 2021) and Face2Face (Yang et al., 2022) are warping-based methods conditioned on 3D Morphable Models. Whilst such methods produce realistic results, they suffer from several visual artifacts and head pose mismatch, especially in large head pose variations. Finally, Meshry et al. (2021) 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.

By contrast to the methods discussed above, which rely on the *training of conditional generative models* on large *paired* datasets in order to learn facial descriptors with disentanglement properties, in this paper we propose a novel and simple face reenactment framework that learns disentangled directions in the latent space of a StyleGAN2 (Karras et al., 2020b) pre-trained on the VoxCeleb (Nagrani et al., 2017) dataset. We show that the discovery of meaningful and disentangled directions that are responsible for controlling the head pose and the facial expression can be used for high quality self- and cross-identity reenactment.

3 Proposed Method

In this section, we present the proposed framework for one-shot neural face reenactment via finding directions in the latent space of StyleGAN2. More specifically, we begin with the most basic variant of our framework for finding reenactment latent directions using unpaired synthetic images in Sect. 3.1—an overview of this is shown in Fig. 1. Next, in Sect. 3.2 we extend this methodology for handling real images along with synthetic ones (i.e., towards real face reenactment), while in Sect. 3.3 we investigate the incorporation of paired video data. In Sect. 3.4 we introduce a joint training scheme that allows for optimization-free reenactment, leading to efficient and consistent neural reenactment. Finally, in Sect. 3.5, on top of the previously introduced variants of our method, we propose the refinement of crucial visual details (i.e., background, hair style) by leveraging the impressive reconstruction capability of StyleGAN2's feature space $\mathcal{F}$.

3.1 Finding Reenactment Latent Directions on Unpaired Synthetic Images

3.1.1 StyleGAN2 Background

Let $\mathcal{G}$ denote the generator of StyleGAN2 (Karras et al., 2020b), as shown in Fig. 1. Specifically, $\mathcal{G}$ takes as input a latent code $\mathbf{w} \in \mathcal{W} \subset \mathbb{R}^{512}$, which is typically the output of StyleGAN2's input MLP-based Mapping Network f that acts on samples $\mathbf{z} \in \mathbb{R}^{512}$ drawn from the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. That is, given a latent code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the generator produces an image $\mathcal{G}(f(\mathbf{z})) \in \mathbb{R}^{3\times256\times256}$.

StyleGAN2 is typically pre-trained on the Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019), which exhibits poor diversity in terms of head pose and facial expression; for instance, FFHQ does not typically account for roll changes in head pose. In order to compare our method with other state-of-the-art methods, we fine-tune StyleGAN2's generator $\mathcal{G}$ on the VoxCeleb dataset (Nagrani et al., 2017), which provides a much wider range of head poses and facial expressions, rendering it very useful for the task of neural face reenactment by finding the appropriate latent directions as will be discussed in the following sections. We note that we fine-tune the StyleGAN2's generator on VoxCeleb dataset using the method provided by Karras et al. (2020a), while we do not impose any reenactment objectives. That is, the fine-tuned generator can 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.

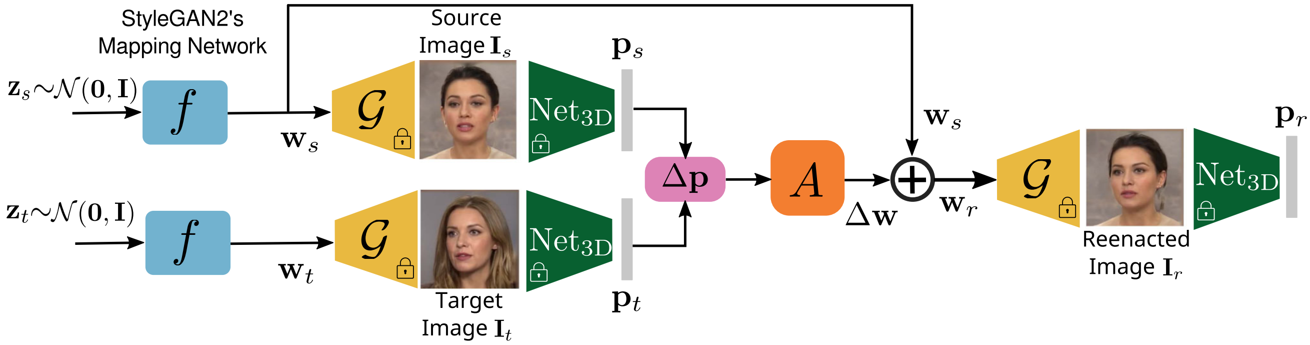


Fig. 1 Overview of the proposed framework: Given a pair of source $\mathbf{I}_s$ and target $\mathbf{I}_t$ images, we calculate the head pose/expression parameter vectors $\mathbf{p}_s$ and $\mathbf{p}_t$ using the Net_{3D} network, respectively. The matrix of directions $\mathbf{A}$ is trained so as, given the shift $\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p}$, the reenacted image $\mathbf{I}_r$ generated using the latent code $\mathbf{w}_r = \mathbf{w}_s + \Delta\mathbf{w}$, transfers the head pose and the expression of the target face, maintaining at the same time the identity of the source face

3.1.2 3D Morphable Model (Net3D)

Given an image, Net3D (Feng et al., 2021) encodes the depicted face's pose into a facial shape vector $\mathbf{s} \in \mathbb{R}^{3N}$, where N denotes the number of vertices, which can be decomposed in terms of a linear 3D facial shape as

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

where $\bar{\mathbf{s}}$ denotes the mean 3D facial shape, $\mathbf{S}_i \in \mathbb{R}^{3N\times m_i}$, $\mathbf{S}_\theta \in \mathbb{R}^{3N\times m_\theta}$ and $\mathbf{S}_e \in \mathbb{R}^{3N\times m_e}$ denote the PCA bases for identity, head orientation and expression, and $\mathbf{p}_i$, $\mathbf{p}_\theta$ and $\mathbf{p}_e$ denote the corresponding identity, head orientation and expression coefficients, respectively. The variables m_i, m_θ and m_e correspond to the number of identity, head pose and expression coefficients. For reenactment, we are interested in manipulating head orientation and expression, thus, our head pose/expression parameter vector is given as $\mathbf{p} = [\mathbf{p}_\theta, \mathbf{p}_e] \in \mathbb{R}^{3+m_e}$. We note that all PCA shape bases are orthogonal to each other, and hence they capture disentangled variations of identity and expression. Finally, we note that they are calculated in a frontalized reference frame, thus, they are also disentangled from 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 as described in detail in the following section.

3.1.3 Reenactment Latent Directions

In particular, we propose to associate a change $\Delta\mathbf{p}$ in the head pose orientation and expression, 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 head pose and expression 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 (Nitzan et al., 2021), 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 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 generator's layers we opt to apply shift changes.

3.1.4 Training Pipeline

In order to optimize the matrix of controllable latent directions $\mathbf{A}$, we propose a simple pipeline, shown in Fig. 1. Specifically, during training, a pair of a source ($\mathbf{z}_s$) and a target ($\mathbf{z}_t$) latent codes are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, giving rise to a pair of a source ($\mathbf{I}_s = \mathcal{G}(f(\mathbf{z}_s))$) and a target ($\mathbf{I}_t = \mathcal{G}(f(\mathbf{z}_t))$) images, as shown in the left part of Fig. 1. The pair of images $\mathbf{I}_s$ and $\mathbf{I}_t$ are then encoded by the pre-trained Net3D into the head pose/expression parameter vectors $\mathbf{p}_s$ and $\mathbf{p}_t$, respectively. Using (2), we calculate the shift $\Delta\mathbf{w}$ in the intermediate latent space of StyleGAN2 as $\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p} = \mathbf{A}(\mathbf{p}_t - \mathbf{p}_s)$ and the *reenactment latent code* $\mathbf{w}_r = \mathbf{w}_s + \Delta\mathbf{w}$. Using the latter we arrive at the reenacted image $\mathbf{I}_r = \mathcal{G}(\mathbf{w}_r)$.

It is worth noting that the only trainable module of the proposed framework is the matrix $\mathbf{A} \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$—i.e., the number of trainable parameters of the proposed framework is $65K$. We also note that, before training, we estimate the distribution of each element of the head pose/expression parameters $\mathbf{p}$ by randomly generating 10K images and calculating using the pre-trained Net3D (Feng et al., 2021) 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 $[-a, a]$ (a being a hyperparameter empirically set to 6). In the appendices (Sect. A.1.1) we further discuss the re-scaling of each element of $\mathbf{p}$. To further encourage disentanglement in the optimized latent directions matrix $\mathbf{A}$, we follow a training strategy where for 50% of the training samples we reenact only one attribute by using $\Delta\mathbf{p} = [0, \ldots, \varepsilon, \ldots, 0]$, where ε is uniformly sampled from $\mathcal{U}[-a, a]$. In the appendices (Sect. A.1.3) we show that the above training strategy improves the disentanglement between the learned directions.

3.1.5 Losses

We train our framework 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 $\mathcal{L}_r$, $\mathcal{L}_{id}$, and $\mathcal{L}_{per}$ denote respectively the *reenactment*, *identity*, and *perceptual* losses with λ_r, λ_{id}, and λ_{per} being weighting hyperparameters empirically set to $\lambda_r = 1$, $\lambda_{id} = 10$ and $\lambda_{per} = 10$. We detail each loss term below.

Reenactment loss $\mathcal{L}_r$ We define the reenactment loss as

$$\mathcal{L}_r = \mathcal{L}_{sh} + \mathcal{L}_{eye} + \mathcal{L}_{mouth},$$

where the *shape* loss term $\mathcal{L}_{sh} = \|\mathbf{S}_r - \mathbf{S}_{gt}\|_1$ imposes head pose and expression transfer from target to source, 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 (1). Specifically, the ground-truth 3D facial shape $\mathbf{S}_{gt}$ should have the identity, i.e., facial shape, of the source image and the facial expression and head pose of the target image, either on the task of self reenactment or on cross-subject reenactment. On self reenactment $\mathbf{S}_{gt}$ is the same with $\mathbf{S}_t$, where $\mathbf{S}_t$ is the facial shape of the target image. On cross-subject reenactment, we calculate $\mathbf{S}_{gt}$ using the identity coefficients

$\mathbf{p}_i^s$ of the source face and the facial expression and head pose coefficients $\mathbf{p}_e^t$, $\mathbf{p}_\theta^t$ of the target face as:

$$\mathbf{S}_{gt} = \bar{\mathbf{S}} + \mathbf{S}_i \mathbf{p}_i^s + \mathbf{S}_\theta \mathbf{p}_\theta^t + \mathbf{S}_e \mathbf{p}_e^t, \quad (4)$$

To enhance the expression transfer, we calculate the *eye* ($\mathcal{L}_{eye}$) and the *mouth* ($\mathcal{L}_{mouth}$) losses. The eye loss $\mathcal{L}_{eye}$ (the mouth loss $\mathcal{L}_{mouth}$ is computed in a similar fashion) compares the inner distances between the eye landmark pairs of upper and lower eyelids between the reenacted and reconstructed ground-truth shapes. In Appendix A.2, we provide a detailed discussion on $\mathcal{L}_{eye}$ and $\mathcal{L}_{mouth}$.

Identity loss $\mathcal{L}_{id}$ We define the identity loss as the cosine similarity between feature representations extracted from the source $\mathbf{I}_s$ and the reenacted $\mathbf{I}_r$ image using ArcFace (Deng et al., 2019). The identity loss imposes the identity preservation between the source and the reenacted image.

Perceptual loss $\mathcal{L}_{per}$ We defined the perceptual loss similarly to Johnson et al. (2016) in order to improve the quality of the reenactment face images.

3.2 Fine-Tuning on Unpaired Real Images

In this section, we extend the basic pipeline of the proposed framework, described in the previous section, in order to learn from both synthetic and real images. For doing so, we propose to (a) use a pipeline for inverting the images back to the latent code space of StyleGAN2, and (b) adopt a mixed training approach (using both synthetic and inverted latent codes) for discovering the latent directions (Sect. 3.1.3).

As discussed in previous sections, the main challenge in the GAN inversion problem is finding a good trade-off between faithful reconstruction of the real image and effective editability using the inverted latent code. Although satisfying both requirements is challenging (Alaluf et al., 2021; Richardson et al., 2021; Tov et al., 2021), we found that the following pipeline produces compelling results for the purposes of our goal (i.e., face/head reenactment). During training, we employ an encoder based method (e4e) (Tov et al., 2021) to invert the real images into the $\mathcal{W}+$ latent space of StyleGAN2 (Abdal et al., 2019). However, directly using the inverted $\mathcal{W}+$ latent codes performs poorly in face reenactment due to the domain gap between the synthetic and inverted latent codes. To alleviate this, we propose a mixed-data approach (i.e., using both synthetic and real images) for training the pipeline presented in Sect. 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 $\mathbf{w}^{inv}$.

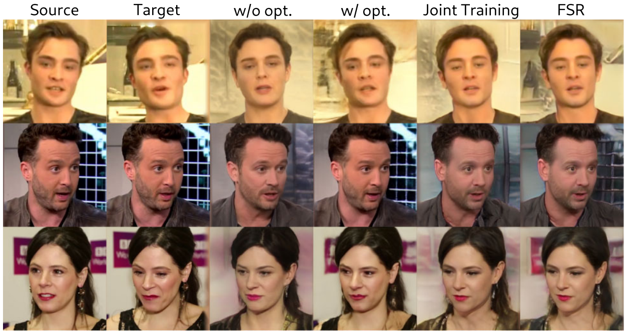


Fig. 2 Examples of face reenactment without ("w/o opt.") and with ("w/ opt.") the generator's optimization. We additionally show results using our proposed joint training scheme ("Joint Training") and the refinement of StyleGAN2's feature space ("FSR") described in Sect. 3.4 and 3.5, respectively

Since the inverted images using e4e (Tov et al., 2021) might still be missing some crucial identity details, we propose to use an additional optimization step (*only during inference*), similarly to Roich et al. (2021), in order to slightly update the generator $\mathcal{G}$ and arrive at better reenacted images in terms of identity preservation. 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. In Fig. 2 we illustrate examples of neural face reenactment without optimizing the generator's weights (w/o opt.—third column) and with optimization (w/ opt.—fourth column), where we observe that, clearly, the reenacted images without the additional optimization step are not able to faithfully reconstruct the real images, while the reenacted images after optimizing the generator weights resembles the real ones more closely.

3.3 Fine-Tuning on Paired Real Images (Video Data)

In the previous sections, we presented the proposed framework for learning from unpaired synthetic and real images. Whilst this provides the benefit of learning from a very large number of identities, making it useful for cross-person reenactment, we show that we can achieve additional improvements by optimizing novel losses introduced by further training on paired data from the VoxCeleb1 (Nagrani et al., 2017) video dataset.

Compared to training from scratch on video data, as in most previous methods (e.g. Zakharov et al. 2020; 2019, Burkov et al. 2020), we argue that our approach offers a more balanced strategy that combines the best of both worlds; that is, training with unpaired 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 head pose/expression. Consequently, we minimize the following loss function

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_{id} \mathcal{L}_{id} + \lambda_{per} \mathcal{L}_{per} + \lambda_{pix} \mathcal{L}_{pix}, \tag{5}$$

where $\mathcal{L}_r$ is the same reenactment loss defined in Sect. 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.

3.4 Joint Training of the Real Image Inversion Encoder $\mathcal{E}_w$ and the Directions Matrix A

As discussed in Sect. 3.2, the encoder-based e4e (Tov et al., 2021) inversion method often fails to faithfully reconstruct real images by typically failing to preserve crucial identity characteristics, as shown in the third column ("w/o opt.") of Fig. 2. Clearly, this poses a certain limitation to the face reenactment methodology presented in Sect. 3.1.4. Optimizing the generator's weights leads to notable improvements (Sect. 3.2), as shown in the fourth column ("w/ opt.") of Fig. 2, albeit, this comes at a significant cost for the task of face reenactment (that is, the optimization of $\mathcal{G}$ takes approximately 20 sec. per frame).

In this section, we propose to *jointly* train the real image inversion encoder $\mathcal{E}_w$ and the directions matrix $\mathbf{A}$, which leads to optimization-free face reenactment at inference time. For doing so, we use paired data as described in Sect. 3.3. An overview of this approach is shown in Fig. 3. Specifically, we first sample a source ($\mathbf{I}_s$) and a target ($\mathbf{I}_t$) image from the same video of VoxCeleb1 (Nagrani et al., 2017) training set, that have the same identity but different head pose/expression. Those images are then fed into the inversion encoder $\mathcal{E}_w$ to predict the corresponding source ($\mathbf{w}_s$) and target ($\mathbf{w}_t$) latent codes. Then, the pre-trained Net$_{3D}$ network extracts the corresponding source ($\mathbf{p}_s$) and target ($\mathbf{p}_t$) parameter vectors. Finally, as described in Sect. 3.1, we generate the reenacted image using the latent code $\mathbf{w}_r = \mathbf{w}_s + \Delta\mathbf{w}$, where $\Delta\mathbf{w} = \mathbf{A}(\mathbf{p}_t - \mathbf{p}_s)$.

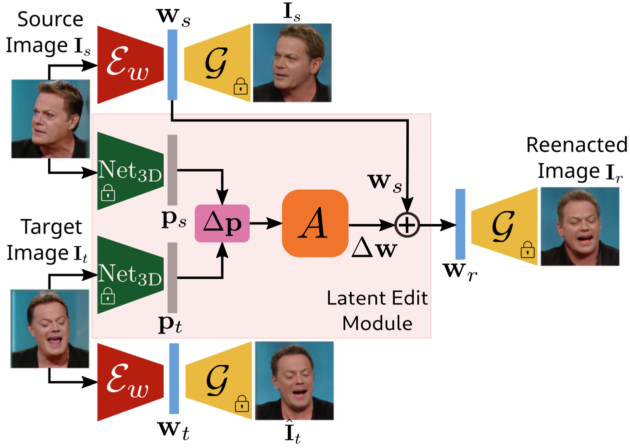


Fig. 3 To eliminate the need for the optimization step during inference, we propose to jointly train the real image inversion encoder $\mathcal{E}_w$ and the directions matrix $\mathbf{A}$. We note that during training both the generator $\mathcal{G}$ and the Net$_{3D}$ network are frozen

3.4.1 Real Image Encoder $\mathcal{E}_w$ Optimization Objective

In order to train the real image encoder $\mathcal{E}_w$ we minimize the following loss:

$$\begin{aligned}\mathcal{L}_{\mathcal{E}_w} = \lambda_{id}(\mathcal{L}_{id}(\mathbf{I}_s, \hat{\mathbf{I}}_s) + \mathcal{L}_{id}(\mathbf{I}_t, \hat{\mathbf{I}}_t))\\ +\lambda_{per}(\mathcal{L}_{per}(\mathbf{I}_s, \hat{\mathbf{I}}_s) + \mathcal{L}_{per}(\mathbf{I}_t, \hat{\mathbf{I}}_t))\\ +\lambda_{pix}(\mathcal{L}_{pix}(\mathbf{I}_s, \hat{\mathbf{I}}_s) + \mathcal{L}_{pix}(\mathbf{I}_t, \hat{\mathbf{I}}_t))\\ +\lambda_{style}(\mathcal{L}_{style}(\mathbf{I}_s, \hat{\mathbf{I}}_s) + \mathcal{L}_{style}(\mathbf{I}_t, \hat{\mathbf{I}}_t)),\end{aligned} \tag{6}$$

where $\mathcal{L}_{id}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{pix}$ denote the identity, perceptual, and pixel-wise losses described in the previous sections.

Additionally, to further improve the style and the quality of the reconstructed images we propose to use a style loss $\mathcal{L}_{style}$ similarly to Barattin et al. (2023). Specifically, we use FaRL (Zheng et al., 2022), a method for general facial representation learning that leverages contrastive learning between images and text pairs to learn meaningful feature representations of facial images. In our method, we use the image Transformer-based encoder, $\mathcal{E}_{FaRL}$, to extract a 512-dimensional feature vector from each image. The proposed style loss is defined as:

$$\begin{aligned}\mathcal{L}_{style} = \|\mathcal{E}_{FaRL}(\mathbf{I}_s) - \mathcal{E}_{FaRL}(\hat{\mathbf{I}}_s)\|_1\\ +\|\mathcal{E}_{FaRL}(\mathbf{I}_t) - \mathcal{E}_{FaRL}(\hat{\mathbf{I}}_t)\|_1.\end{aligned} \tag{7}$$

3.4.2 Directions Matrix A Optimization Objective

In order to train the directions matrix $\mathbf{A}$ we minimize the following loss:

$$\begin{aligned}\mathcal{L}_{\mathbf{A}} = \lambda_r \mathcal{L}_r + \lambda_{id} \mathcal{L}_{id} + \lambda_{per} \mathcal{L}_{per}\\ +\lambda_{pix} \mathcal{L}_{pix} + \lambda_{style} \mathcal{L}_{style},\end{aligned} \tag{8}$$

where $\mathcal{L}_r$, $\mathcal{L}_{id}$, $\mathcal{L}_{per}$, $\mathcal{L}_{pix}$, and $\mathcal{L}_{style}$ denote respectively the reenactment loss defined in Sect. 3.1, the identity, the perceptual, the pixel-wise, and the style losses calculated between the reenacted $\mathbf{I}_r$ and the target images $\mathbf{I}_t$.

Moreover, to further improve the reenactment results we propose an additional *cycle loss* term $\mathcal{L}_{cycle}$ (Bounareli et al., 2023; Sanchez & Valstar, 2020). Specifically, as shown in Fig. 4, given an image pair of a source ($\mathbf{I}_s^1$) and a target ($\mathbf{I}_t^1$) images, we calculate the corresponding reenacted image $\mathbf{I}_r^1 \equiv \mathbf{I}_t^1$. Having as source image the reenacted image $\mathbf{I}_r^1$ and as target the source image $\mathbf{I}_s^1$, we calculate a new reenacted

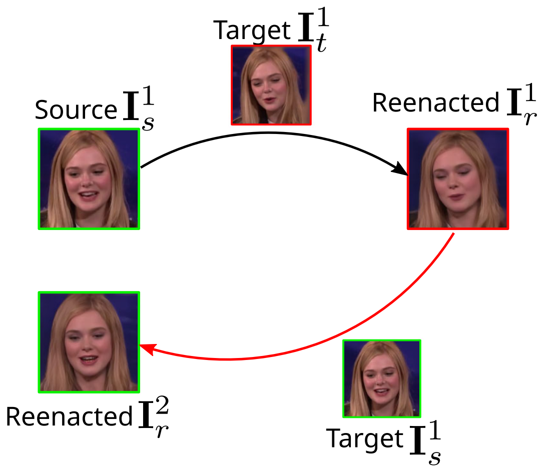


Fig. 4 Cycle loss: Given a pair of source ($\mathbf{I}_s^1$) and target ($\mathbf{I}_t^1$) images, we calculate the corresponding reenacted image $\mathbf{I}_r^1$. We then use this image as source and as target the source image from the first image pair and we calculate the second reenacted image $\mathbf{I}_r^2$, which is imposed to be similar with $\mathbf{I}_s^1$

image $\mathbf{I}_r^2$ that is imposed to be similar to $\mathbf{I}_s^1$. Consequently, we calculate all reconstruction losses, i.e. $\mathcal{L}_{id}$, $\mathcal{L}_{per}$, $\mathcal{L}_{pix}$, and $\mathcal{L}_{style}$, between the source image $\mathbf{I}_s^1$ and the reenacted image $\mathbf{I}_r^2$. In our ablation studies (Sect. 4.2), we show that using the proposed cycle loss improves the face reenactment performance.

3.4.3 Joint Optimization Objective

Overall, the objective of the joint optimization is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathbf{A}} + \mathcal{L}_{E_w} + \mathcal{L}_{cycle}. \tag{9}$$

We note that, in this training phase, we fine-tune the matrix $\mathbf{A}$ and the real image inversion encoder $\mathcal{E}_w$, trained as described in Sect. 3.2. As demonstrated in Fig. 2, using the proposed joint training scheme (Joint Training) our method is able to reconstruct the identity details of the real faces without performing any optimization step. In Sect. 4, we quantitatively demonstrate that our proposed method produces similar results on self reenactment with our method when optimizing the generator's weights. Nevertheless, on the more challenging tasks of cross-subject reenactment and on large head pose differences between the source and target faces, the joint training scheme outperforms our results with optimization, producing more realistic images with less visual artifacts.

3.5 Feature Space $\mathcal{F}$ Refinement

In this section, we propose an additional module for our face reenactment framework that refines the feature space $\mathcal{F}$ of the StyleGAN2's generator; taking advantage from its exceptional expressiveness (e.g., in terms of background, hair style/color, or hair accessories). In order to mitigate the limited editability of $\mathcal{F}$ (Kang et al., 2021; Parmar et al., 2022), we propose a two-step training procedure, which we illustrate in Fig. 5. Specifically, we first train a feature space encoder $\mathcal{E}_{\mathcal{F}}$, using the ResNet-18 (He et al., 2016) architecture, in the task of real image inversion. $\mathcal{E}_{\mathcal{F}}$ takes as input a real image and predicts the shift Δf_4 that updates the feature map f_4 as:

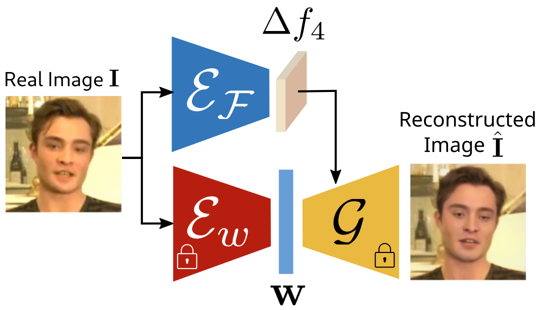


Fig. 5 Training of feature space encoder $\mathcal{E}_{\mathcal{F}}$ in the real image inversion task. $\mathcal{E}_{\mathcal{F}}$ takes as input a real image and predicts the shift Δf_4 that updates the feature map f_4 of the 4^{th} feature layer of StyleGAN2's generator

$$\hat{f}_4 = f_4 + \Delta f_4, \tag{10}$$

where f_4 is the feature map calculated using the inverted latent code $\mathbf{w}$. The training objective in this step consists of the reconstruction losses, namely identity, perceptual, pixel-wise, and style, calculated between the reconstructed $\hat{\mathbf{I}}$ and the real images $\mathbf{I}$ as described in (Eq. 6). It is worth nothing that we only refine the 4^{th} feature layer of StyleGAN2's generator $\mathcal{G}$ that we found to be in particular beneficial to the face reenactment task, in contrast to later feature layers that, despite their capability in reconstructing almost perfectly the real images, they suffer from poor semantic editability (as shown by Yao et al. (2022b)).

As discussed above, using the updated feature map $\hat{f}_4$ to refine details on the edited images leads to visual artifacts. To address this, we propose a framework that efficiently learns to predict the updated feature map of the edited image $\hat{f}_4^r$ using the refined source feature map $\hat{f}_4^s$. We illustrate this in Fig. 6, where, given a source and a target image pair, we first calculate the reenacted latent code $\mathbf{w}_r$ as described in Sect. 3.4. We note that the directions matrix $\mathbf{A}$ and the real image inversion encoder $\mathcal{E}_w$ are frozen during training. Then, using the feature encoder $\mathcal{E}_{\mathcal{F}}$, we calculate the source refined feature map $\hat{f}_4^s$ using (10). In order to calculate the refined feature map of the reenacted image $\hat{f}_4^r$, we introduce the *Feature Transformation (FT)* module, that takes as input the difference of the source refined feature map $\hat{f}_4^s$ and the reenacted feature map f_4^r, and outputs the shift Δf_4^r, which can be used to calculate the updated feature map $\hat{f}_4^r$ given by

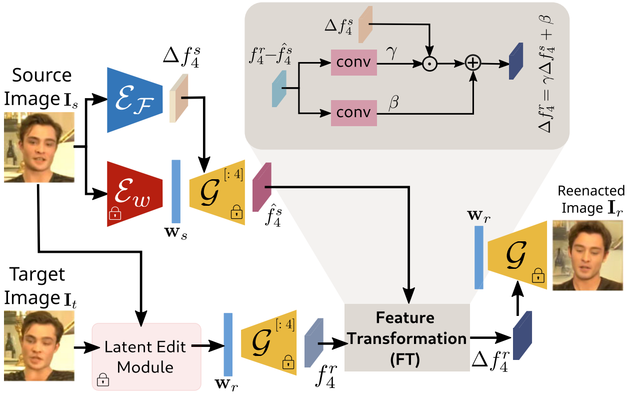


Fig. 6 Training of the feature space encoder $\mathcal{E}_{\mathcal{F}}$ and the Feature Transformation (FT) module to efficiently refine the feature map f_4^r of the reenacted images

(10). As shown in Fig. 6, the proposed Feature Transformation (FT) module learns two modulation parameters, namely γ and β, that efficiently transform the shift Δf_4^s of the source feature map into the shift Δf_4^r of the reenacted feature map as:

$$\Delta f_4^r = \gamma \odot \Delta f_4^s + \beta. \tag{11}$$

As illustrated in Fig. 6, the FT module consists of two convolutional blocks with 2 convolutional layers each. We note that in this training step we train both the FT module and the feature space encoder $\mathcal{E}_{\mathcal{F}}$. Our training objective consists of the reconstruction losses, namely identity, perceptual, pixel-wise and style, calculated between the reenacted and the target images (described in detail in Sect. 3.4).

Finally, in Fig. 7 we give some indicative results of the proposed reenactment variant of our method that learns to optimize the feature space $\mathcal{F}$ ("FSR") in comparison to the variant of our method described in Sect. 3.4 ("Joint Training") and Parmar et al. (2022) ("SAM"). We note that using the $\mathcal{W}+$ latent space (Joint Training / Sect. 3.4) leads to relatively faithful reconstruction performance, albeit, without being able to reconstruct every detail on the background or the hair styles. As we will show in the experimental section, qualitatively and quantitatively, but also in the conducted user study, such level of detail is crucial for the task of face reenactment. Similarly, SAM (Parmar et al., 2022) is able to better reconstruct the background however the reenacted images suffer from visual artifacts (marked with red arrows in Fig. 7) and, thus, look unrealistic, especially around the face area. By contrast, the proposed framework that learns to optimize the feature space $\mathcal{F}$ ("FSR") leads to both notably more faithful face reenactment exhibiting less artifacts.

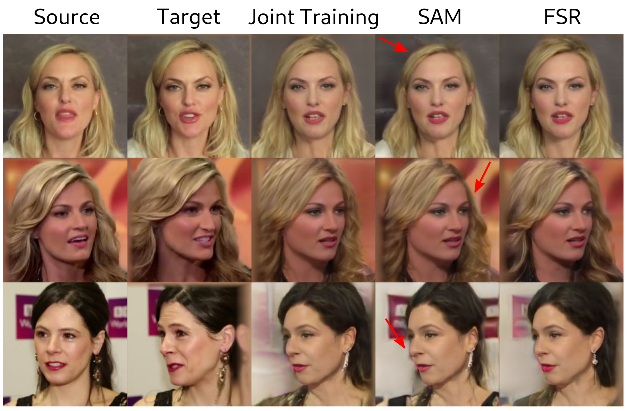


Fig. 7 Face reenactment examples using only the $\mathcal{W}+$ latent space ("Joint Training"), SAM method (Parmar et al., 2022) and our proposed method for feature space refinement ("FSR")

4 Experiments

In this section, we present qualitative and quantitative results, along with a user study, in order to evaluate the proposed framework (all its variants) in the task of neural face reenactment and compare with several recent state-of-the-art (SOTA) approaches. The bulk of our results and comparisons, reported in Sect. 4.1, are on self- and cross-person reenactment on the VoxCeleb1 (Nagrani et al., 2017) dataset. Comparisons with state-of-the-art on the VoxCeleb2 (Chung et al., 2018) test set are provided in the appendices. Finally, in Sect. 4.2 we report ablation studies on the various design choices of our method and in Sect. 4.3 we discuss its limitations.

Implementation details We fine-tune StyleGAN2 on the VoxCeleb1 dataset with 256 × 256 image resolution and we train the e4e encoder of Tov et al. (2021) for real image inversion. The 3D shape model we use (i.e., the Net3D module shown in Figs. 1, 3) is DECA (Feng et al., 2021). For our training procedure described in Sects. 3.1 3.2, and 3.3, we only learn the directions matrix $\mathbf{A} \in \mathbb{R}^{(N_l \times 512) \times k}$ where $k = 3 + m_e$, $m_e = 12$ and $N_l = 8$. We train three matrices of directions: (i) the first one is on synthetically generated images (Sect. 3.1), (ii) the second one is on mixed real and synthetic data (Sect. 3.2), and (iii) the third one is fine-tuning (ii) on paired data (Sect. 3.3). Additionally, on the proposed joint training scheme (Sect. 3.4), we fine-tune both the directions matrix $\mathbf{A}$ and the real image inversion encoder $\mathcal{E}_w$. Finally, in the feature space refinement variant (Sect. 3.5) we train both the feature space encoder $\mathcal{E}_{\mathcal{F}}$ and the propose Feature Transformation (FT) module. It is worth noting that during the first and second training phases, we perform cross-subject training, i.e., the source and target faces have different identities. This approach enables our model to generalize effectively across various identities, resulting in improved performance on the challenging task of cross-subject reenactment. On the rest training phases we perform self reenactment, where the source and target faces are sampled from the same video. For training, we used the Adam optimizer (Kingma & Ba, 2015) with constant learning rate 10^{-4}. All models are implemented in PyTorch (Paszke et al., 2019).

4.1 Comparison with State-of-the-Art on VoxCeleb

In this section, we compare the performance of our method against the state-of-the-art in face reenactment on VoxCeleb1 (Nagrani et al., 2017). We conduct two types of experiments, namely self- and cross-person reenactment. For evaluation purposes, we use both the video data provided by Zakharov et al. (2019) and the original test-set of VoxCeleb1. We note that there is no overlap between the train and test identities and videos. We compare our method quantitatively and qualitatively with nine methods: X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022). For X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021) and Face2Face (Yang et al., 2022), we use the pre-trained (by the authors) model on VoxCeleb1. For Fast bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020) and LSR (Meshry et al., 2021) we also use the pre-trained (by the authors) models on VoxCeleb2 (Chung et al., 2018). Regarding Dual (Hsu et al., 2022), we use the pre-trained by the authors model on both VoxCeleb (Chung et al., 2018; Nagrani et al., 2017) and MPIE (Gross et al., 2010) datasets. For fair comparison with the methods of Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021) and Dual (Hsu et al., 2022), we evaluate their model under the one-shot setting. We note that we will be referring to our method that optimizes the generator's weights during inference as Latent Optimization Reenactment (LOR), whereas LOR+ will be referring to our final model with joint training and feature space refinement. We note that in the LOR+ model, we do not optimize the generator weights.

4.1.1 Quantitative Comparisons

We report eight different metrics. We compute the Learned Perceptual Image Path Similarity (LPIPS) (Zhang et al., 2018) to measure the perceptual similarities, and to quantify identity preservation we compute the cosine similarity (CSIM) of ArcFace (Deng et al., 2019) features. Moreover, we measure the quality of the reenacted images using the Fréchet-Inception Distance (FID) metric (Heusel et al., 2017), while we also report the Fréchet Video Distance (FVD) (Skorokhodov et al., 2022; Unterthiner et al., 2018) metric that measures both the video quality and the temporal consistency of the generated videos. To quantify the head pose/expression transfer, we calculate the normalized mean error (NME) between the predicted landmarks in the reenacted and target images. We use (Bulat & Tzimiropoulos, 2017) for landmark estimation, and we calculate the NME by normalizing it with the square root of the ground truth face bounding box and scaled by a factor of 10^3. We further evaluate the head pose transfer by calculating the average $\mathcal{L}_1$ distance of the head pose orientation (Average Rotation Distance, ARD) in degrees, and the expression transfer by calculating the average $\mathcal{L}_1$ distance of the expression coefficients $\mathbf{p}_e$ (Average Expression Distance, AED) and the Action Units Hamming distance (AU-H) computed as in Doukas et al. (2021).

In Table 1 we report quantitative results on self reenactment, using the original test set of VoxCeleb1 (Nagrani et al., 2017) and the test set provided by Zakharov et al. (2019). Additionally, in Table 2 we report results on a more challenging condition on self reenactment where the source and target faces have large head pose difference. Specifically, we randomly selected from the test set of VoxCeleb1 1,000 image pairs with head pose distance larger than 10°. The head pose distance is calculated as the average of the absolute differences of the three Euler angles (i.e., yaw, pitch, and roll) between the source and target faces. In the appendices (Sect. A.4), we provide additional details regarding our benchmark dataset. We note that in self reenactment, all metrics are calculated between the reenacted and the target faces. As shown in Table 1, the warping-based methods, namely X2Face, PIR, HeadGAN and Face2Face have high values 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, these results are accompanied by poor quantitative and qualitative results when there is a significant change on the head pose (e.g., see Fig. 8 and Table 2). Additionally, regarding head pose/expression transfer, our method (LOR+) achieves similar results on NME with Fast Bi-layer (Zakharov et al., 2020), while on ARD and AED metrics we outperform all methods. Finally, our results on FID and FVD metrics confirm that the quality of our generated videos resembles the quality of VoxCeleb dataset. Nevertheless, our method (LOR+) on the challenging condition with large head pose differences between the source and target faces (Table 2) outperforms all methods.

Cross-subject reenactment is more challenging compared to self reenactment, as source and target faces have different identities, and in this case it is important to maintain the source identity characteristics without transferring the tar-

Table 1 Quantitative results on self-reenactment. The results are reported on the combined original test set of VoxCeleb1 (Nagrani et al., 2017) and the test set released by Zakharov et al. (2019). For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓)

Method	CSIM	LPIPS	FID	FVD	NME	ARD	AED	AU-H
X2Face (Wiles et al., 2018)	0.70	0.13	35.5	490	17.8	1.5	0.90	0.22
FOMM (Siarohin et al., 2019)	0.65	0.14	35.6	523	34.1	5.0	1.30	0.28
Fast Bi-layer (Zakharov et al., 2020)	0.64	0.23	52.8	706	**13.2**	1.1	0.80	**0.21**
Neural-Head (Burkov et al., 2020)	0.40	0.22	98.4	617	15.5	1.3	0.90	0.23
LSR (Meshry et al., 2021)	0.59	0.13	45.7	484	17.8	1.0	0.75	0.22
PIR (Ren et al., 2021)	0.71	0.12	57.2	545	18.2	1.8	0.94	0.24
HeadGAN (Doukas et al., 2021)	0.68	0.13	52.5	518	15.6	1.8	1.30	0.26
Dual (Hsu et al., 2022)	0.26	0.21	75.0	602	35.0	3.7	1.20	0.27
Face2Face (Yang et al., 2022)	**0.72**	0.12	55.3	682	16.0	1.5	0.93	0.24
LOR (Ours)	0.66	0.11	**35.0**	**400**	14.1	1.1	0.68	**0.21**
LOR+ (Ours)	0.67	**0.10**	36.0	440	13.6	**0.7**	**0.60**	**0.21**

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

Table 2 Quantitative comparisons on the benchmark set (Benchmark-L) with image pairs from VoxCeleb1 dataset, where the average head pose distance is larger than 10°. For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓)

Method	CSIM	LPIPS	FID	ARD	AED	AU-H
X2Face (Wiles et al., 2018)	0.60	0.14	57.4	1.8	1.1	0.25
FOMM (Siarohin et al., 2019)	0.60	0.15	65.2	2.2	1.1	0.25
Fast Bi-layer (Zakharov et al., 2020)	0.58	0.20	96.2	1.2	0.8	**0.22**
Neural-Head (Burkov et al., 2020)	0.40	0.18	94.2	1.2	0.9	0.23
LSR (Meshry et al., 2021)	0.55	**0.12**	56.0	1.2	0.8	0.23
PIR (Ren et al., 2021)	0.57	0.15	67.6	2.4	1.4	0.25
HeadGAN (Doukas et al., 2021)	0.38	0.26	66.2	3.6	1.4	0.29
Dual (Hsu et al., 2022)	0.25	0.22	83.6	4.0	1.3	0.28
Face2Face (Yang et al., 2022)	0.47	0.28	**35.3**	1.6	1.2	0.27
LOR (Ours)	0.51	0.13	47.0	1.6	0.8	0.35
LOR+ (Ours)	**0.62**	**0.12**	46.7	**0.8**	**0.6**	**0.22**

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

get ones. In Table 3, we report the quantitative results for cross-subject reenactment, where we randomly select 200 video pairs from the small test set of Zakharov et al. (2019). In this task, CSIM metric is calculated between the source and the reenacted faces while ARD, AED and AU-H metrics between the target and the reenacted faces. As depicted in Table 3, our method (LOR+) achieves the best results on head pose and expression transfer, while we achieve high score in CSIM metric. It is worth noting that the high CSIM value for FOMM, HeadGAN and Face2Face is not accompanied by good qualitative results as shown in Figs. 8 and 27, where in most cases, those methods are not able to generate realistic images.

To further evaluate the performance of reenactment methods we conduct a user study, where we ask 30 users to select the method that best reenacts the source frame on self and cross-subject reenactment tasks. For the purposes of the user study we utilise only our final model (LOR+). The results are reported in Table 14 and as shown our method is the most preferable (by a large margin—52.1% versus the 19.2% second best method), which also validates our quantitative results.

Additionally, in Table 4 we report comparisons on inference time required to generate a video of 200 frames. As shown, X2Face (Wiles et al., 2018) and FOMM (Siarohin et al., 2019) are the fastest methods, however their overall performance (quantitative and qualitative results) is unsatisfactory (i.e., visual artifacts). Nevertheless, our proposed method (LOR+) is able to generate compelling reenacted images, while also being competitive in terms of inference time. Notably, our final model (LOR+) outperforms our model that requires the optimization step (LOR), which is a time consuming operation.

4.1.2 Qualitative Comparisons

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. 8, (b) in the appendices, we provide more results in self and cross-subject

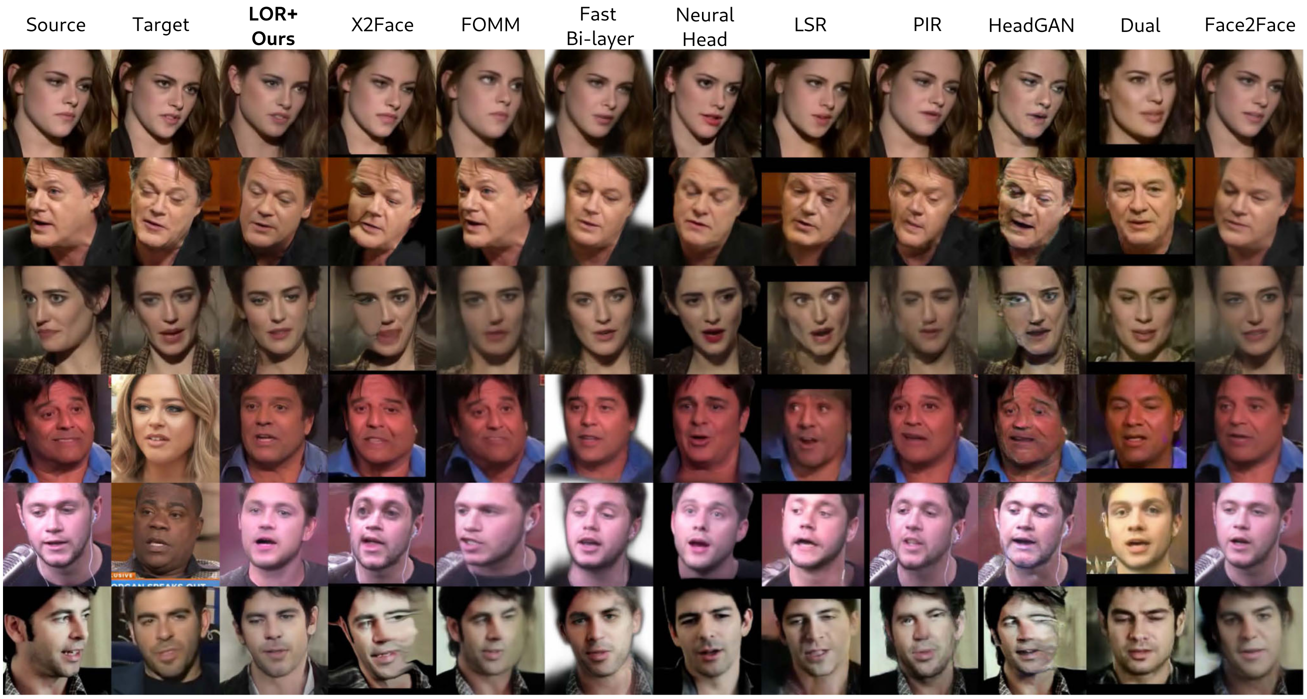


Fig. 8 Qualitative results and comparisons for self (top three rows) and cross-subject reenactment (last three rows) on VoxCeleb1. The first and second columns show the source and target faces. Our method preserves the appearance and identity characteristics (e.g., face shape) of the source face significantly better and also faithfully transfer the target head pose/expression without producing visual artifacts

Table 3 Quantitative results on cross-subject reenactment. For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓)

Method	CSIM	FID	ARD	AED	AU-H
X2Face (Wiles et al., 2018)	0.57	89.0	2.2	1.5	0.31
FOMM (Siarohin et al., 2019)	0.73	116.3	7.7	2.0	0.41
Fast Bi-layer (Zakharov et al., 2020)	0.48	116.0	1.5	1.3	**0.29**
Neural-Head (Burkov et al., 2020)	0.36	124.1	1.7	1.6	0.30
LSR (Meshry et al., 2021)	0.50	84.2	1.4	1.2	0.30
PIR (Ren et al., 2021)	0.62	110.5	2.2	1.4	0.33
HeadGAN (Doukas et al., 2021)	0.75	122.2	2.1	1.7	0.33
Dual (Hsu et al., 2022)	0.22	107.0	3.5	1.5	0.33
Face2Face (Yang et al., 2022)	**0.76**	124.2	1.8	1.5	0.32
LOR (Ours)	0.63	86.2	1.2	1.0	0.31
LOR+ (Ours)	0.68	**78.4**	**0.7**	**0.8**	0.30

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

reenactment both on VoxCeleb1 and VoxCeleb2 datasets (Figs. 23, 26, 27, 28, 29), and (c) we also provide a supplementary video with self and cross-subject reeenactment results from VoxCeleb1 and VoxCeleb2 datasets. As we can see from Fig. 8 and the videos provided in the supplementary material, our method provides, for the majority of videos, the highest reenactment quality including accurate transfer of head 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. 8, 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 (Zakharov et al., 2020). Finally, to show that our method is able to generalise well on other facial video datasets, we provide additional results on the FaceForensics (Rössler et al., 2018) and 300-VW (Shen et al., 2015) datasets in the appendices (Fig. 30).

Table 4 Quantitative comparisons on inference time required to generate a video of 200 frames

Method	Inf. time (sec)
X2Face (Wiles et al., 2018)	**11.0**
FOMM (Siarohin et al., 2019)	**11.0**
Fast Bi-layer (Zakharov et al., 2020)	61.0
Neural-Head (Burkov et al., 2020)	115.0
LSR (Meshry et al., 2021)	110.0
PIR (Ren et al., 2021)	54.0
HeadGAN (Doukas et al., 2021)	84
Dual (Hsu et al., 2022)	27.0
Face2Face (Yang et al., 2022)	146
LOR (Ours)	40.0
LOR+ (Ours)	27.0

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

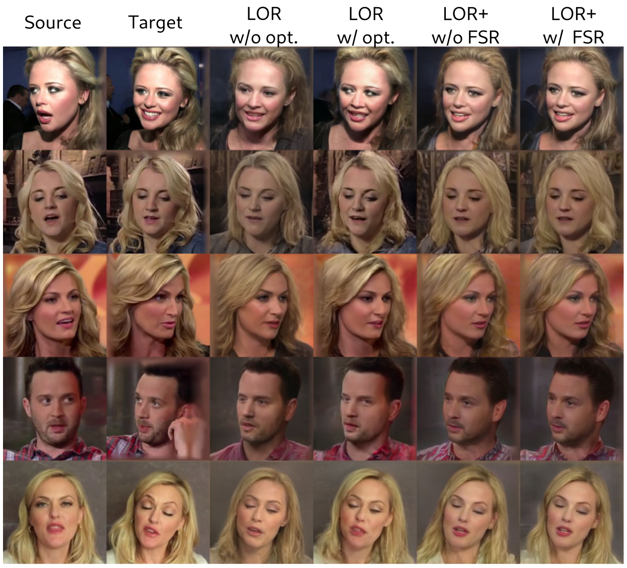


Fig. 9 Qualitative comparisons of the various models of our work on self reenactment

4.2 Ablation Studies

In this section, we perform several ablation tests to (a) assess the different variants of our method, i.e., the optimization of generator $\mathcal{G}$ during inference (Sect. 3.3), the proposed joint training scheme (Sect. 3.4) and the refinement of the feature space (Sect. 3.5), (b) measure the impact of the identity and perceptual losses, and the additional shape losses for the eyes and mouth (Sect. 3.1), (c) validate our trained models on synthetic, mixed and paired images, and (d) measure the impact of the style and cycle losses (introduced in Sect. 3.4).

For (a), we report results of our method on self and cross-subject reenactment, with our model (LOR) described in Sect. 3.3 without performing optimization (w/o opt.) and with optimization (w/ opt.) of the generator $\mathcal{G}$ during inference. We also report results of our final model (LOR+) without the additional feature space refinement (FSR) (Sect. 3.4) and with feature space refinement (Sect. 3.5). As shown in Table 5, the optimization of $\mathcal{G}$ during inference improves our results (as expected) especially regarding the identity preservation (CSIM) compared to our model without performing optimization. Nevertheless, our proposed joint training scheme (LOR+ *w/o FSR*) achieves the same results on image reconstruction metrics (CSIM and LPIPS), and improves our results on head pose/expression transfer (ARD, AED) without performing any optimization of the generator. Additionally, the proposed refinement on the feature space of StyleGAN2 (LOR+ *w/ FSR*) improves all quantitative results. It is worth mentioning that the new proposed components (Joint Training and Feature Space Refinement) compared to our previous work (Bounareli et al., 2022) improve our results especially on the challenging tasks of self reenactment with large head pose differences between the source and target faces and on cross-subject reenactment. Figure 9 illustrates results on self reenactment using the above described models. As shown LOR without optimization cannot accurately reconstruct the identity of the source face, while with optimization the identity details are better reconstructed but the reenacted images contain noticeable visual artifacts. On the contrary, the proposed joint training scheme (LOR+ w/o FSR) is able to accurately reconstruct the identity of the source faces and produce artifact-free images without performing any subject fine-tuning. Finally, the proposed feature space refinement (LOR+ w/ FSR) improves our qualitative results by producing more realistic images (i.e., better background and hair style reconstruction).

Table 5 Quantitative results of the various models of our work on self reenactment (SR), self reenactment with image pairs that have large head pose difference (SR - large head pose) and on cross-subject reenactment (CR)

Method	SR				SR-large head pose			CR		
	CSIM	LPIPS	ARD	AED	CSIM	ARD	AED	CSIM	ARD	AED
LOR *w/o opt.*	0.37	0.12	1.4	0.90	0.34	1.7	0.9	0.43	1.5	1.0
LOR *w/ opt.*	0.66	0.11	1.1	0.68	0.51	1.6	0.8	0.63	1.2	1.0
LOR+ *w/o FSR*	0.66	0.11	0.8	**0.60**	0.60	0.9	**0.6**	0.67	**0.7**	**0.8**
LOR+ *w/ FSR*	**0.67**	**0.10**	**0.7**	**0.60**	**0.62**	**0.8**	**0.6**	**0.68**	**0.7**	**0.8**

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

Table 6 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 head pose and expression

Method	CSIM	ARD	AED
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

Bold values indicate the best values for each metric between the different methods

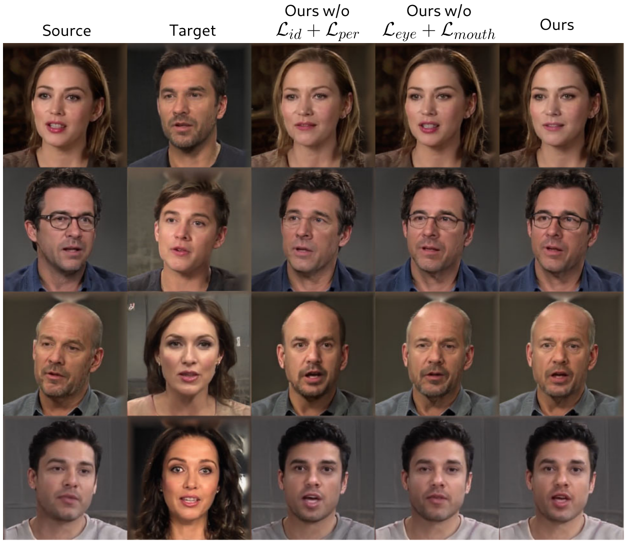

Source
Target
Ours w/o $\mathcal{L}_{id} + \mathcal{L}_{per}$
Ours w/o $\mathcal{L}_{eye} + \mathcal{L}_{mouth}$
Ours

Fig. 10 Qualitative comparisons 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

For (b), 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 head pose and expression of the target. As shown in Table 6, 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 6, 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 head pose and expression error. The impact of those losses is also obvious on our qualitative comparisons in Fig. 10. As shown, when we exclude the identity and perceptual losses from the training process, the generated images lack several appearance details, while omitting the eye and mouth losses leads to less accurate facial expression transfer.

For (c), we evaluate the three different training schemes, namely synthetic only (Sect. 3.1), mixed synthetic-real (Sect. 3.2), and mixed synthetic-real fine-tuned with paired data (Sect. 3.3) for self-reenactment. The results, reported in Table 7 and in Fig. 11, illustrate the impact of each of these training schemes with the one using paired data providing the best results as expected. As shown in Fig. 11, our final model trained with paired data produces more realistic images with less artifacts.

Table 7 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

Method	CSIM	ARD	AED
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**

Bold values indicate the best values for each metric between the different methods

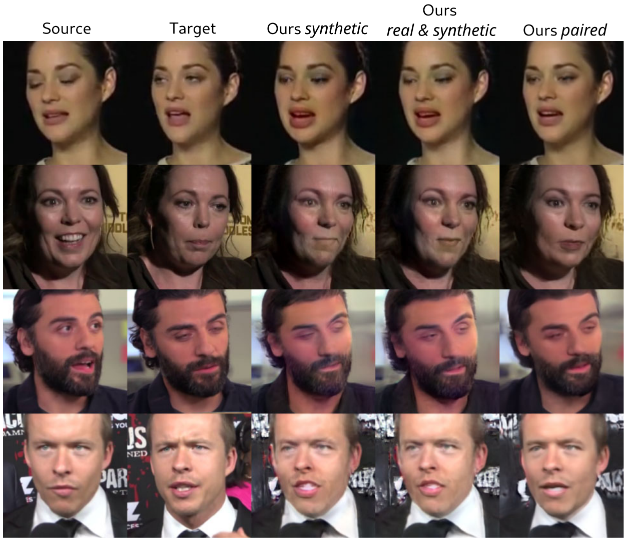


Fig. 11 Qualitative results of the three different models trained on synthetic images, on both synthetic and real images and on paired data

Table 8 Ablation study on the impact of style $\mathcal{L}_{style}$ and cycle $\mathcal{L}_{cycle}$ losses

Method	CSIM	ARD	AED
Ours w/o $\mathcal{L}_{style}$	0.64	0.9	0.7
Ours w/o $\mathcal{L}_{cycle}$	0.62	1.0	0.7
Ours w/ $\mathcal{L}_{style} + \mathcal{L}_{cycle}$	**0.66**	**0.8**	**0.60**

Bold values indicate the best values for each metric between the different methods

Finally, for (d) we perform experiments on self reenactment using our model with joint training scheme, without using the style loss $\mathcal{L}_{style}$ and without the cycle loss $\mathcal{L}_{cycle}$. As shown in Table 8 our final model with both those losses has better results both on identity preservation and on head pose/expression transfer. Additionally, as illustrated in Fig. 12, our final model using both the style and the cycle loss

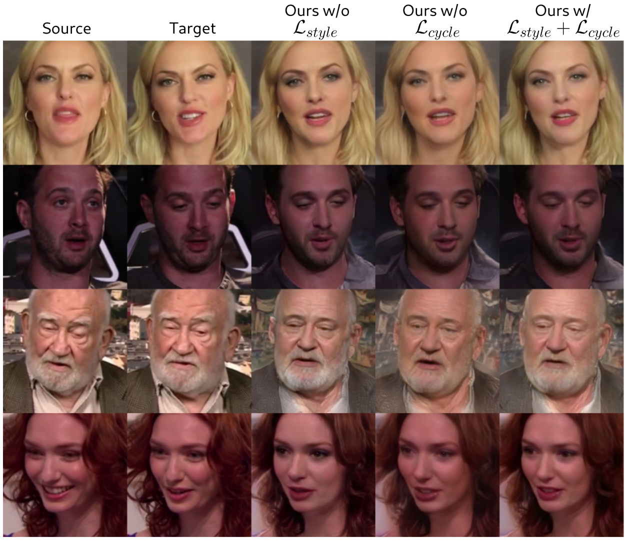

Source
Target
Ours w/o $\mathcal{L}_{style}$
Ours w/o $\mathcal{L}_{cycle}$
Ours w/ $\mathcal{L}_{style} + \mathcal{L}_{cycle}$

Fig. 12 Qualitative comparisons on the impact of the style $\mathcal{L}_{style}$ and cycle $\mathcal{L}_{cycle}$ losses

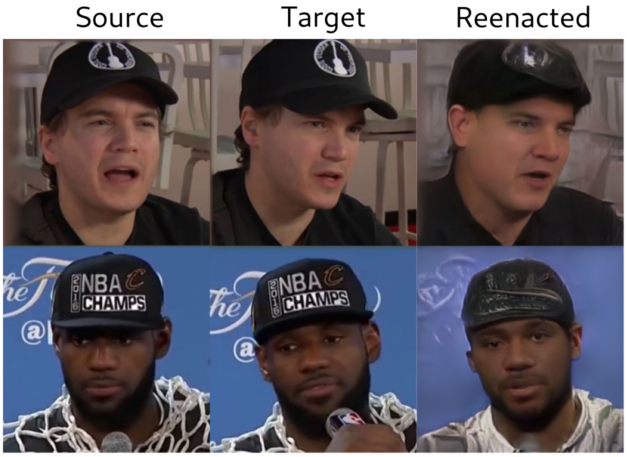


Fig. 13 Cases where the reconstruction of facial accessories like hats fails. The first two columns show the source and target images, while the reenacted images are presented on the last column

has improved results in terms of identity/appearance preservation.

4.3 Limitations

As shown both in our quantitative and qualitative results, our method is able to efficiently reenact the source faces, by preserving the source identity characteristics and by faithfully transferring the target head pose and expression. Our proposed method, which is based on a pre-trained StyleGAN2 model, enables both self and cross-subject reenactment using only one source frame and without any further subject fine-tuning. The proposed joint training scheme of the real image encoder $\mathcal{E}_w$ and the direction matrix $\mathbf{A}$ enables more accurate identity reconstruction and facial image editing without many visual artifacts, especially on the challenging task of extreme head poses. Additionally, the refinement of StyleGAN2's feature space $\mathcal{F}$ enables better reconstruction of various image details including background, hair style/color and facial accessories, resulting in visually more realistic images. Nevertheless, in Fig. 13 we observe that especially on hair accessories, such as hats that are underrepresented on the training dataset, our method is not able to faithfully reconstruct every detail when editing the head pose orientation.

5 Conclusions

In this paper, we presented a novel approach towards neural head/face reenactment using a 3D shape model to learn disentangled directions of head pose and expression in the latent GAN space. This approach comes with specific advantages, such as the use of powerful pre-trained GANs and 3D shape models, which have been thoroughly developed and studied by the research community during the past years. Our method is able to successfully disentangle the facial movements and the appearance of the input images leveraging the disentangled properties of the pre-trained StyleGAN2 model. Consequently, our framework effectively mimics the target head pose and expression without transferring identity details from the driving images. Additionally, our method features several favorable properties including one-shot face reenactment without the need for further subject-specific fine-tuning. It also allows for improved cross-subject reenactment through the proposed upaired data training with synthetic and real images. While our method demonstrates compelling results, it relies on the capabilities of StyleGAN2 model, which is bounded by the distribution of the training dataset. If the dataset lacks diversity in terms of complex backgrounds, facial accessories like hats, glasses e.t.c, this can affect our model's ability to generalize well to more complex datasets. This limitation highlights the importance of using more diverse video datasets during the training of the generative models.

Finally, we acknowledge that although face reenactment can be used in a variety of applications such as art, entertainment, video conferencing etc., it can also be applied for malicious purposes, including deepfake creation, that could potentially harm individuals and the society. It is important for the researchers on our field to be aware of the potential risks and promote the responsible use of this technology.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02018-6.

Acknowledgements This work has been supported by the EU H2020 AI4Media No. 951911 project.

Data Availibility Statement The VoxCeleb1 and VoxCeleb2 video datasets are publicly available at https://www.robots.ox.ac.uk/\nobreakspace{}vgg/data/voxceleb/index.html. One possible issue of using these two datasets is that videos might be missing or taken down from YouTube. The FaceForensics and the 300-VW datasets are available upon the acceptance of End User License forms at https://github.com/ondyari/FaceForensic and https://ibug.doc.ic.ac.uk/resources/300-VW/, respectively.

Appendix A

In this appendix, we first provide an analysis of the discovered directions in the latent space in App. A.1 and we describe in detail the calculation of the shape losses in App. A.2. Additionally, we show results of our method on the task of facial attribute editing in App. A.3. In App. A.4, we provide details about the benchmark datasets used to evaluate our method on large head pose variations. Finally, in App. A.5, we compare the proposed framework with state-of-the-art methods for synthetic image editing on FFHQ dataset (Karras et al., 2019) and we show comparisons on real image editing against five methods that perform real image inversion using the feature space of StyleGAN2 (Karras et al., 2020b). Moreover, we provide additional quantitative and qualitative results both on the VoxCeleb1 (Nagrani et al., 2017) and the VoxCeleb2 (Chung et al., 2018) datasets and we show additional results on the FaceForensics (Rössler et al., 2018) and the 300-VW (Shen et al., 2015) datasets.

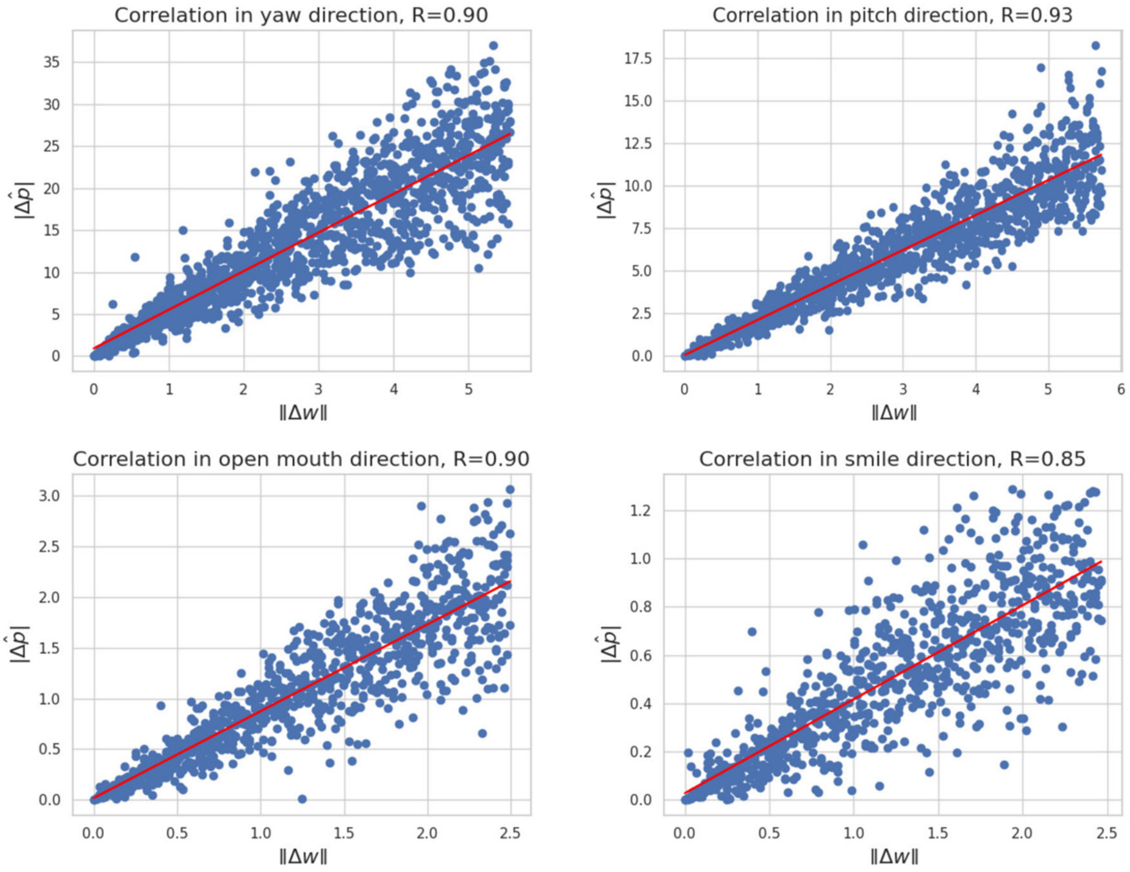


Fig. 14 Analysis of the correlation between shifts $\|\Delta\mathbf{w}\|$ in the latent space and the predicted changes $|\hat{\Delta\mathbf{p}}|$ in the parameters space. We show results of four different attributes (yaw and pitch angles, smile, and open mouth). In all attributes the correlation is high, indicating strong linear relationship

A.1 Analysis of the Learned Directions

A.1.1 Head Pose/Expression Parameter Vector

The elements of $\mathbf{p} = [\mathbf{p}_\theta, \mathbf{p}_e]$, i.e., the head pose $\mathbf{p}_\theta$ and the expression $\mathbf{p}_e$ coefficients, are typically in different ranges of values. That is, head pose $\mathbf{p}_\theta$ is given in terms of the three Euler angles (i.e., yaw, pitch, and roll) in degrees (i.e., in the range $[-90, 90]$), while the expression coefficients $\mathbf{p}_e$ are given in the range of $[-2, 2]$ with the vast majority (99%) of samples in VoxCeleb1 dataset being within the range of $[-1, 1]$. In order to bring each element of $\mathbf{p} = [\mathbf{p}_\theta, \mathbf{p}_e]$ into a common range of values $[-a, a]$, we sampled 10,000 synthetic facial images and calculated the corresponding values for $\mathbf{p}_\theta$ and $\mathbf{p}_e$ using the pre-trained DECA (Feng et al., 2021) network. We then re-scaled each element x of $\mathbf{p}$ in $[-a, a]$ using min-max scaling; i.e., $\hat{x} = \frac{x - x_{min}}{x - x_{max}} \times 2a - a$. This way, we guarantee that each component contributes evenly to the overall facial representation, regardless of its original range, providing stability in the training process. The specific re-scaling range, i.e., $[-6, 6]$, is practically imposed by the StyleGAN's latent space, as Voynov and Babenko (2020) originally pointed out, meaning that traversing the latent space outside this range, often leads to severe degradation in the quality of the generated images, since latent codes lie in regions of low density.

A.1.2 Linearity

In this work, we discover the disentangled directions that control the head pose and the expression by optimising a matrix $\mathbf{A}$ so that:

$$\Delta\mathbf{w} = \mathbf{A}\Delta\mathbf{p}, \tag{12}$$

where $\Delta\mathbf{w}$ denotes a shift in the latent space and $\Delta\mathbf{p}$ denotes the corresponding change in the parameters space. That is, independently of the source attributes, we assume linearity between a shift $\Delta\mathbf{w}$ that is applied to an arbitrary code $\mathbf{w}$ and the induced change $\Delta\mathbf{p}$ in the parameter space—i.e., the change between the source and the reenacted attributes.

Several recent methods propose to learn linear directions in the latent space of StyleGAN (Shen & Zhou, 2021; Shen et al., 2020; Voynov & Babenko, 2020) in order to perform synthetic image editing, based on the fact that the $\mathcal{W}$ latent space of StyleGAN (Karras et al., 2019) has been designed to be linear and disentangled. Furthermore, Nitzan et al. (2021) provide a comprehensive analysis on the existence of linear relations between the magnitude of change in the semantic attributes (e.g., head orientation, smile, etc) and the traversal distance along the corresponding *linear* latent paths. In order to further support our hypothesis (i.e., Eq. 12), we perform a similar analysis by examining the correlation between random shifts in the latent space, $\Delta\mathbf{w}$, and the predicted shifts in the parameters space, $\hat{\Delta\mathbf{p}}$. Specifically, given a *known* change $\Delta\mathbf{p}$, we calculate the corresponding $\Delta\mathbf{w}$ using Eq. 12 and we apply this change (i.e., $\Delta\mathbf{w}$) onto random latent codes of images with different attributes. Then, we calculate the *predicted* change $\hat{\Delta\mathbf{p}}$ between the source and the reenacted images. In Fig. 14 we demonstrate the results of our analysis in four different attributes, namely, yaw angle, pitch angle, smile, and open mouth. In all attributes, the calculated correlation is close to 0.9 indicating strong linear relationship. Finally, additional visual results of two different subjects in

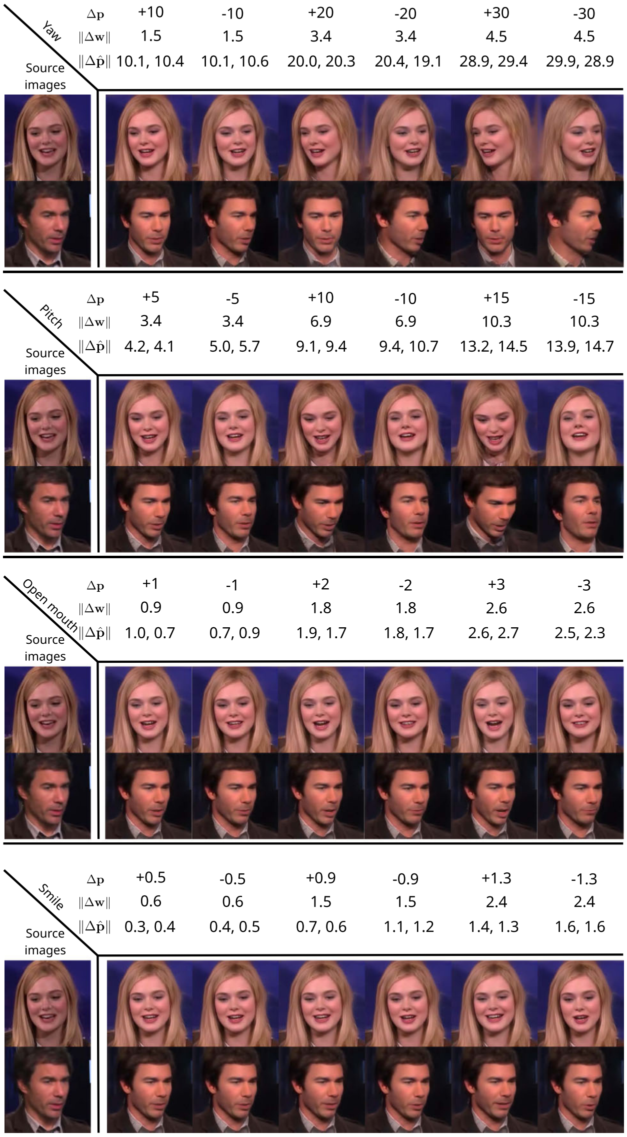
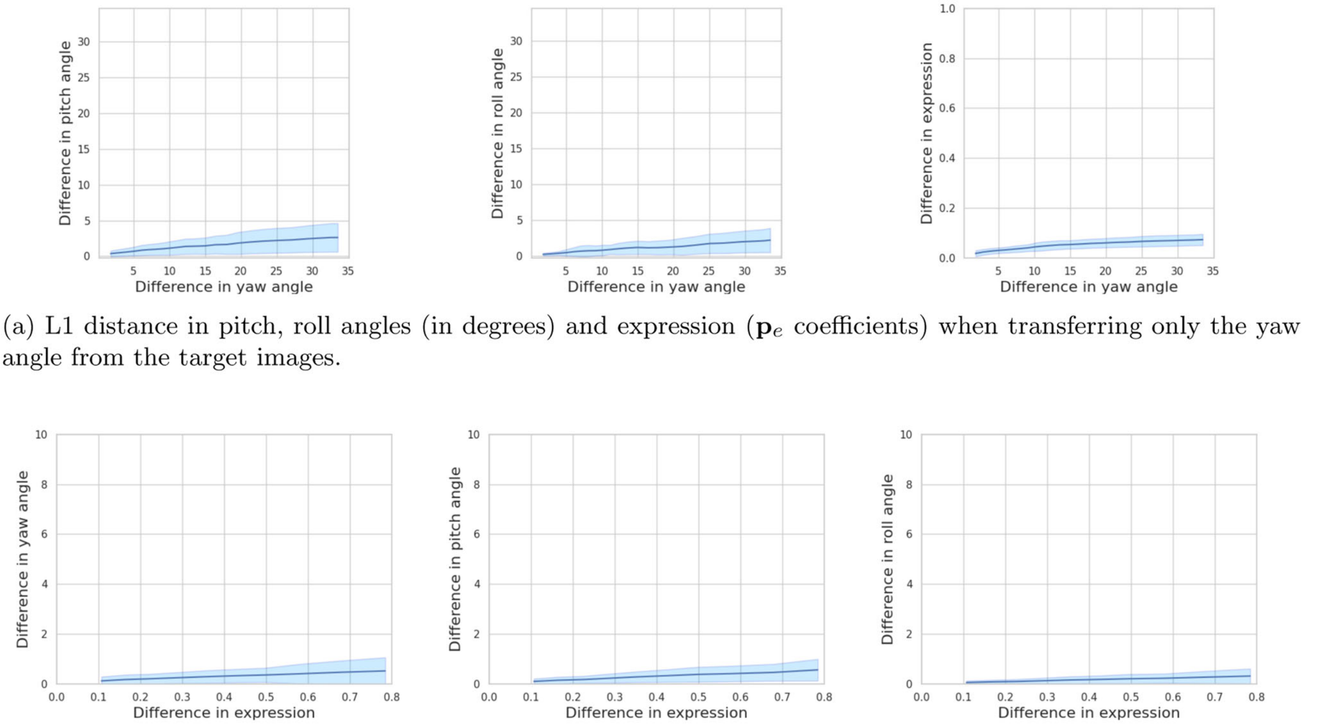

Yaw
Source images
Δp +10 -10 +20 -20 +30 -30
‖Δw‖ 1.5 1.5 3.4 3.4 4.5 4.5
‖Δp̂‖ 10.1, 10.4 10.1, 10.6 20.0, 20.3 20.4, 19.1 28.9, 29.4 29.9, 28.9
Pitch
Source images
Δp +5 -5 +10 -10 +15 -15
‖Δw‖ 3.4 3.4 6.9 6.9 10.3 10.3
‖Δp̂‖ 4.2, 4.1 5.0, 5.7 9.1, 9.4 9.4, 10.7 13.2, 14.5 13.9, 14.7
Open mouth
Source images
Δp +1 -1 +2 -2 +3 -3
‖Δw‖ 0.9 0.9 1.8 1.8 2.6 2.6
‖Δp̂‖ 1.0, 0.7 0.7, 0.9 1.9, 1.7 1.8, 1.7 2.6, 2.7 2.5, 2.3
Smile
Source images
Δp +0.5 -0.5 +0.9 -0.9 +1.3 -1.3
‖Δw‖ 0.6 0.6 1.5 1.5 2.4 2.4
‖Δp̂‖ 0.3, 0.4 0.4, 0.5 0.7, 0.6 1.1, 1.2 1.4, 1.3 1.6, 1.6

Fig. 15 Visual results illustrating the strongly linear relationship between $\|\Delta\mathbf{p}\|$ and $\|\Delta\mathbf{w}\|$. Specifically, given two different input images and ground truth changes $\|\Delta\mathbf{p}\|$ in the parameter space, we calculate the corresponding $\|\Delta\mathbf{w}\|$ shift in the latent space and the predicted changes $\|\Delta\hat{\mathbf{p}}\|$ between the source and shifted images. We note that a similar shift $\|\Delta\mathbf{w}\|$ corresponds to a similar change in the parameter space independently of the facial attributes of the source images



(a) L1 distance in pitch, roll angles (in degrees) and expression ($\mathbf{p}_e$ coefficients) when transferring only the yaw angle from the target images.

(b) L1 distance in yaw, pitch and roll angles (in degrees) when transferring only the smile expression from the target images.

Fig. 16 Difference between the source and reenacted facial attributes when transferring only one facial attribute (e.g., yaw angle and smile expression) from the target images

different head poses and expressions are depicted in Fig. 15. Specifically, we show the ground truth change $\|\Delta \mathbf{p}\|$ in the parameter space, the corresponding $\|\Delta \mathbf{w}\|$, and the predicted changes $\|\Delta \hat{\mathbf{p}}\|$ between the source and shifted images. Above the presented images in the row where we report $\|\Delta \hat{\mathbf{p}}\|$ the two values separated by commas correspond to the subjects depicted in the first and second row. As shown, a change $\|\Delta \mathbf{w}\|$ corresponds to a similar change in the parameter space $\|\Delta \hat{\mathbf{p}}\|$ independently of the facial attributes of the source images.

A.1.3 Disentanglement

Following the common understanding of disentanglement in the area of GANs (Chen et al., 2016; Deng et al., 2020; Karras et al., 2019), we refer to a disentangled latent direction when travelling across it leads to image generations where only a single attribute changes. To assess the directions learnt by our method in terms of disentanglement, in Fig. 16 we illustrate the differences between the source and reenacted attributes when changing a single attribute. In Fig. 16a, we only transfer the yaw angle from the target image, while in Fig. 16b we only transfer the smile expression from the target image. We observe that the differences between the rest of the attributes (i.e., pitch, roll, and expression in Fig. 16a and yaw, pitch, and

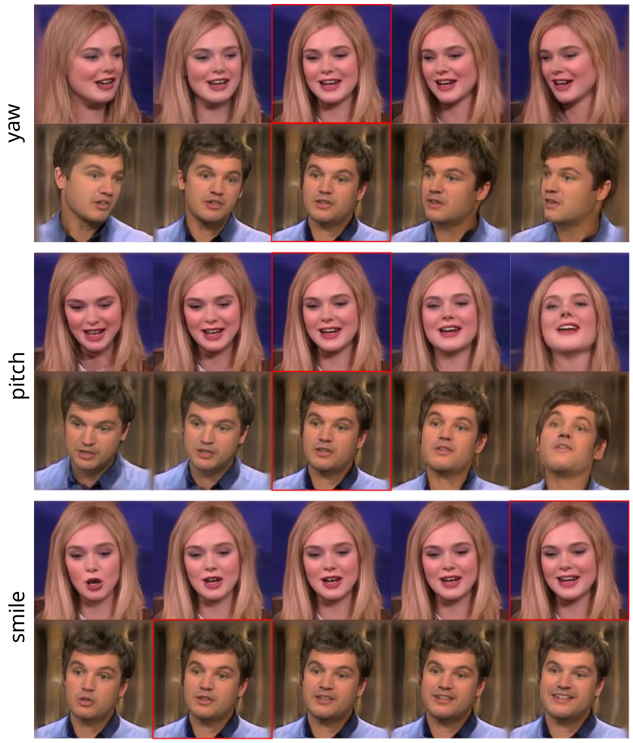


Fig. 17 Visual examples of editing only one facial attribute, namely yaw and pitch angles, and smile. The source images are depicted inside the red boxes

Table 9 Ablation on the impact of single attribute change during training

Head pose changing		Single attribute change during training	Absolute Error (↓)			
			Yaw	Pitch	Roll	$\frac{1}{12}\sum_{t=1}^{12} e_i$
Yaw		True	–	**1.2**	**1.3**	**0.85**
		False	–	1.3	1.4	0.88
Pitch		True	**1.0**	–	**0.7**	**0.58**
		False	1.2	–	0.9	0.65
Roll		True	**1.5**	**0.9**	–	**0.53**
		False	1.6	**0.9**	–	0.57
Expression changing		Single expression change during training	Yaw	Pitch	Roll	$\frac{1}{12}\sum_{t=0, t\neq i}^{12} e_i$
i	e_i					
1	"Expression_1"	True	**0.3**	**0.2**	**0.2**	**0.49**
		False	0.7	0.5	0.4	0.53
2	"Expression_2"	True	**0.4**	**0.2**	**0.1**	**0.27**
		False	0.8	0.5	0.3	0.40
3	"Expression_3"	True	**0.3**	**0.2**	**0.1**	**0.18**
		False	0.7	0.4	0.3	0.30
4	"Expression_4"	True	**0.4**	**0.2**	**0.2**	**0.16**
		False	0.7	0.4	0.5	0.30
5	"Expression_5"	True	**0.3**	**0.2**	**0.1**	**0.25**
		False	1.0	0.4	0.5	0.50
6	"Expression_6"	True	**0.4**	**0.2**	**0.1**	**0.35**
		False	0.7	1.0	0.4	0.50
7	"Expression_7"	True	**0.4**	**0.2**	**0.1**	**0.26**
		False	0.7	1.0	0.4	0.36
8	"Expression_8"	True	**0.3**	**0.2**	**0.1**	**0.11**
		False	0.7	1.0	0.4	0.25
9	"Expression_9"	True	**0.3**	**0.2**	**0.1**	**0.23**
		False	0.7	1.0	0.4	0.37
10	"Expression_10"	True	**0.4**	**0.2**	**0.1**	**0.18**
		False	0.7	1.0	0.4	0.30
11	"Expression_11"	True	**0.3**	**0.2**	**0.1**	**0.18**
		False	0.7	1.0	0.4	0.30
12	"Expression_12"	True	**0.4**	**0.2**	**0.2**	**0.16**
		False	0.7	1.0	0.4	0.35

Bold values indicate the best values for each metric between the different methods

roll in Fig. 16b) are clearly small, which indicates that the discovered directions are disentangled. We note that these plots were calculated using 2000 random image pairs. In Fig. 16a, we show the differences in yaw angle that were calculated as the absolute difference between the source and the target yaw angles (measured in degrees), while the differences in the *unchanged* attributes were calculated between the source and reenacted images. In a similar way, in Fig. 16b we show the differences in expression that were calculated as the absolute difference between the source and the target expression ($\mathbf{p}_e$ coefficients). Moreover, in Fig. 17 we demonstrate visual results of editing only one direction, namely yaw, pitch angles and smile. As shown, when altering the head pose, i.e., yaw and pitch angles, all other facial attributes, i.e., facial expressions, remain unchanged. Additionally, when altering the smile expression, we observe changes only around the mouth area, while head orientation remains the same. In more detail, in the first subject where smile is controlled (row 5), the eyes remain closed despite editing the smile expression, while in the second subject (row 6) the raised brows remain unaffected.

Finally, in order to encourage better disentanglement between the facial attributes that we control, during training we propose to change only one attribute on 50% of the training samples within each batch. To validate the effectiveness of the above training choice, in Table 9 we compare two models trained on synthetic images and report results indicating with “True" the model trained with single attribute change and “False" the model without the single attribute change. Specifically, we change only one attribute, namely the yaw, pitch, or roll head rotation angles, or one of the expression coefficients ($e_i, i = 1, \ldots, 12$). We then calculate and report the error, i.e., the $l1$-distance between the source and the reenacted attributes that should remain unchanged. For instance, when changing only the yaw angle, then both the pitch and the roll angles, as well as the expressions should remain the same as those of the source image. We note that regarding the expression error we report the mean error across all expressions. When we alter a specific expression e_i, we calculate the expression error by excluding that particular expression, as denoted by the last column of Table 9. As shown, adopting this training strategy leads to better disentanglement with respect to all the 3 Euler angles and the 12 facial expressions.

A.2 Shape Losses

In order to transfer the target head pose and expression to the source face, we calculate the *reenactment loss* as:

$$\mathcal{L}_r = \mathcal{L}_{sh} + \mathcal{L}_{eye} + \mathcal{L}_{mouth}, \tag{13}$$

where $\mathcal{L}_{sh}$ is the shape loss and $\mathcal{L}_{eye}$, $\mathcal{L}_{mouth}$ the eye and mouth loss, respectively. As shown in our ablation studies (Sect. 4.2), $\mathcal{L}_{eye}$, $\mathcal{L}_{mouth}$ losses contribute to more accurate expression transfer from the target face to the source face. Specifically, eye loss compares the inner distances $\mathrm{d} = \|(\cdot,\cdot)\|_1$ of the eye landmark pairs (defined as P_{eye}) of upper and lower eyelids between the reenacted ($\mathbf{S}_r$) and reconstructed ground-truth ($\mathbf{S}_{gt}$) shapes:

$$\mathcal{L}_{eye} = \sum_{(i,j)\in P_{eye}} \left\| \mathrm{d}\big(\mathbf{S}_r(i), \mathbf{S}_r(j)\big) - \mathrm{d}\big(\mathbf{S}_{gt}(i), \mathbf{S}_{gt}(j)\right\|, \tag{14}$$

Similarly, mouth loss is computed between the mouth landmark pairs. In Fig. 18, we show the landmark pairs that are used to calculate these losses. In more detail, P_{eye} and P_{mouth} are defined as:

$$\begin{aligned} P_{eye} = \big[&(37,40),(38,42),(39,41),\\ &(43,46),(44,48),(45,47)\big],\\ P_{mouth} = \big[&(49,55),(50,60),(51,59),(52,58),\\ &(53,57),(54,56),\\ &(61,65),(62,68),(63,67),(64,66)\big] \end{aligned}$$

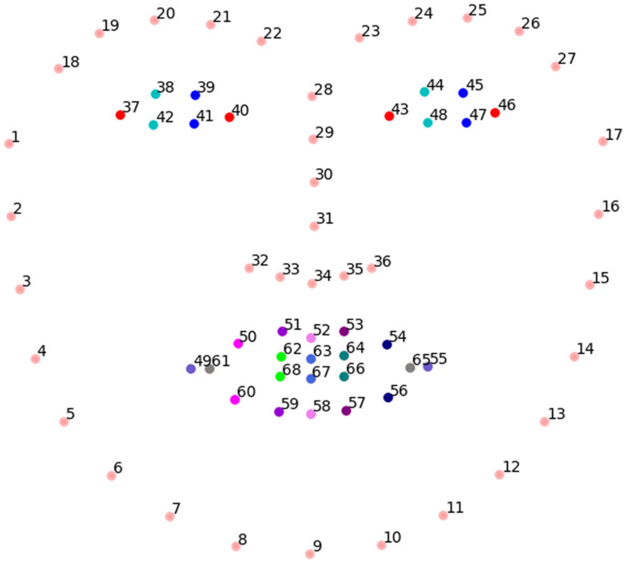


Fig. 18 Depiction of the landmark pairs P_{eye} and P_{mouth} that contribute to the corresponding losses $\mathcal{L}_{eye}$ and $\mathcal{L}_{mouth}$. The landmarks of each pair are drawn with the same color

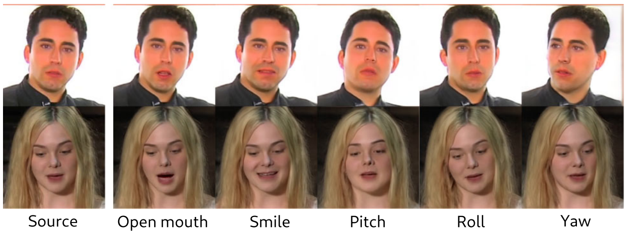


Fig. 19 Our method can perform head pose and expression editing on real images. Specifically, we are able to edit an attribute by keeping all other attributes unchanged. The first column shows the source images, while the rest columns show editings of different expressions and head poses

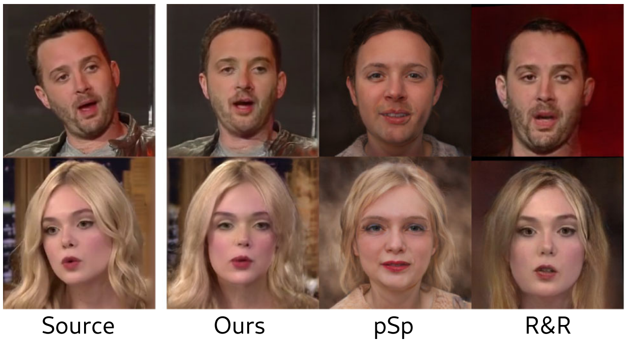


Fig. 20 Face frontalization examples. We perform comparisons with pSp (Richardson et al., 2021) and R&R (Zhou et al., 2020) and we show that our method successfully perform face frontalization by preserving the identity of the source face

Table 10 Quantitative results on frontalization task. We compare our method with pSp (Richardson et al., 2021) and R&R (Zhou et al., 2020) by evaluating the identity preservation (CSIM) and the Average Rotation Distance (ARD) between the source and the frontalized images

Method	CSIM	ARD
pSp (Richardson et al., 2021)	0.40	3.0
R&R (Zhou et al., 2020)	0.45	3.5
Ours	**0.60**	**1.2**

Bold values indicate the best values for each metric between the different methods

A.3 Image Editing

Our method is able to discover the disentangled directions of head pose and expression in the latent space of StyleGAN2. Consequently, except from face reenactment, our model can perform head pose and expression editing by simply setting the desired head pose or expression. Figure 19 illustrates results of per attribute editing. As shown, our model can alter the head pose (i.e., yaw, pitch, and roll) or the expression (e.g., open mouth, smile) by maintaining all other attributes unchanged. Similarly, our method can be used in the frontalization task. We compare our model with the methods of pSp (Richardson et al., 2021) and R&R (Zhou et al., 2020) and we report both qualitative (Fig. 20) and quantitative (Table 10) results. Specifically, we randomly select 250 frames of different identities from the VoxCeleb test set and we perform frontalization. In Table 10, we evaluate the identity preservation (CSIM) and the Average Rotation Distance (ARD) between the source and the frontalized images.

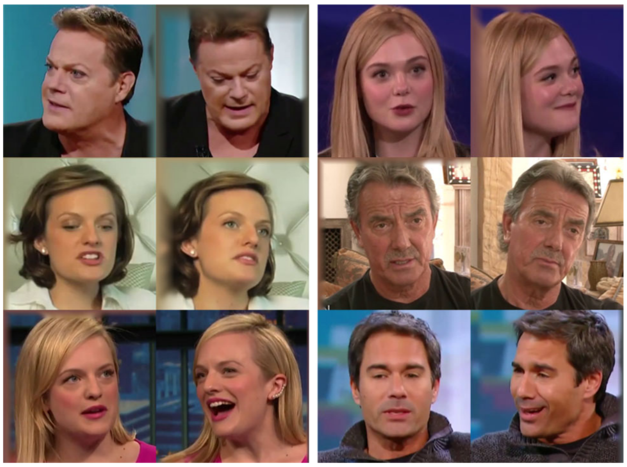


Fig. 21 Indicative examples of source-target image pairs from our benchmark set (Benchmark-L), where the average head pose distance is larger than 10°

A.4 Benchmark Datasets with Large Head Pose Variations

As mentioned in Sect. 4.1.1, the benchmark used in Table 2 (Benchmark-L) in order to evaluate our method on large head pose reenactment contains 1,000 image pairs from the VoxCeleb1 dataset with head pose distance larger than 10°, calculated as the average $L1$ distance of the three Euler angles (yaw, pitch, roll). In Fig. 22, we present a comparison of the distributions of the three Euler angles (yaw, pitch and roll) and the average head pose distance between the VoxCeleb1 dataset and the aforementioned benchmark dataset (Benchmark-L). As shown Benchmark-L comprises image pairs that have larger head pose distances compared to the average head pose distance observed in the VoxCeleb1 dataset. Additionally, Fig. 21, illustrates some indicative example image pairs from the benchmark dataset, where there is a wide range on the head pose variations across all three Euler angles.

To further validate our method on larger head pose differences we generate a new benchmark dataset (Benchmark-XL) using images from both the VoxCeleb1 and the VoxCeleb2 datasets. Specifically, we randomly select $1,000$ image pairs where the distance on the yaw angle is larger than 30° and on the pitch or roll angles larger than 20°. As shown in Fig. 22, Benchmark-XL consists of image pairs with “extreme" head pose differences compared to the distribution of the overall dataset. In Tables 11, 12 and in Fig. 23, we demonstrate the quantitative and qualitative comparisons on Benchmark-XL both on VoxCeleb1 and VoxCeleb2, respectively. As shown our method is able to better preserve the identity of the source faces, successfully transfer the target head pose and expression and generate realistic images without many visual artifacts compared to the other state-of-the-art methods.

A.5 Additional Results

A.5.1 Comparisons with Synthetic Image Editing Methods

In order to show the superiority of our method against methods for synthetic image editing, we compare against two state-of-the-art methods, namely ID-disentanglement (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021). The authors of ID-disentanglement (Nitzan et al., 2020) introduce a method that learns to disentangle the head pose/expression and the identity characteristics using a pre-trained StyleGAN2 on FFHQ dataset. Additionally, StyleFlow (Abdal et al., 2021) is a state-of-the-art method that finds meaningful non-linear directions in the latent space of StyleGAN2 using supervision from multiple attribute classifiers and regressors. Both ID-disentanglement (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021) provide pre-trained models using the

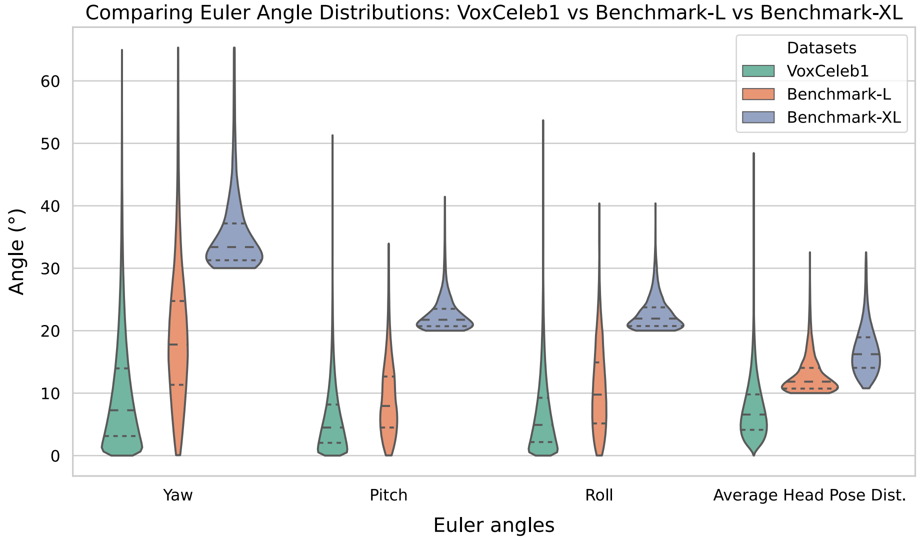


Fig. 22 Comparison of the distributions of the three Euler angles (yaw, pitch and roll) and the average head pose distance between the VoxCeleb1 dataset, the benchmark set, called here Benchmark-L (average head pose distance larger than 10°), and the new benchmark set, called here Benchmark-XL (yaw larger than 30°, pitch/roll larger than 20°)

Table 11 Quantitative comparisons on the Benchmark-XL with image pairs from VoxCeleb1 dataset, where the distance on the yaw angle is larger than 30° and on the pitch or roll angles larger than 20°

Method	CSIM	LPIPS	FID	ARD	AED	AU-H
X2Face (Wiles et al., 2018)	0.55	0.13	91.2	2.2	1.1	0.27
FOMM (Siarohin et al., 2019)	0.56	0.14	92.6	2.7	1.2	0.27
Fast Bi-layer (Zakharov et al., 2020)	0.53	0.19	113.5	1.4	0.8	0.23
Neural-Head (Burkov et al., 2020)	0.40	0.17	109.5	2.0	0.8	0.25
LSR (Meshry et al., 2021)	0.53	**0.12**	78.1	1.2	0.8	0.23
PIR (Ren et al., 2021)	0.53	0.14	95.5	3.0	1.1	0.27
HeadGAN (Doukas et al., 2021)	0.30	0.27	92.0	3.7	1.4	0.30
Dual (Hsu et al., 2022)	0.25	0.20	101.3	4.6	1.2	0.28
Face2Face (Yang et al., 2022)	0.38	0.28	**60.0**	3.4	1.1	0.27
LOR (Ours)	0.42	0.15	72.0	2.0	1.0	0.24
LOR+ (Ours)	**0.57**	0.13	65.6	**0.9**	**0.6**	**0.22**

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

Table 12 Quantitative comparisons on the Benchmark-XL with image pairs from VoxCeleb2 dataset, where the distance on the yaw angle is larger than 30° and on the pitch or roll angles larger than 20°

Method	CSIM	LPIPS	FID	ARD	AED	AU-H
X2Face (Wiles et al., 2018)	0.45	0.20	161.4	8.6	1.4	0.31
FOMM (Siarohin et al., 2019)	0.49	0.18	175.3	6.2	1.2	0.28
Fast Bi-layer (Zakharov et al., 2020)	0.47	0.22	172.2	1.7	0.9	0.27
Neural-Head (Burkov et al., 2020)	0.36	0.18	160.1	1.9	1.0	0.25
LSR (Meshry et al., 2021)	0.51	**0.15**	146.6	1.4	0.8	**0.24**
PIR (Ren et al., 2021)	0.42	0.19	173.1	4.5	1.2	0.27
HeadGAN (Doukas et al., 2021)	0.28	0.32	170.2	2.5	1.4	0.33
Dual (Hsu et al., 2022)	0.22	0.30	146.3	4.8	1.3	0.27
Face2Face (Yang et al., 2022)	0.29	0.30	151.2	2.5	1.1	0.29
LOR (Ours)	0.39	0.17	155.5	2.5	1.0	0.29
LOR+ (Ours)	**0.53**	0.18	**136.0**	**1.3**	**0.7**	**0.24**

Bold values indicate the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

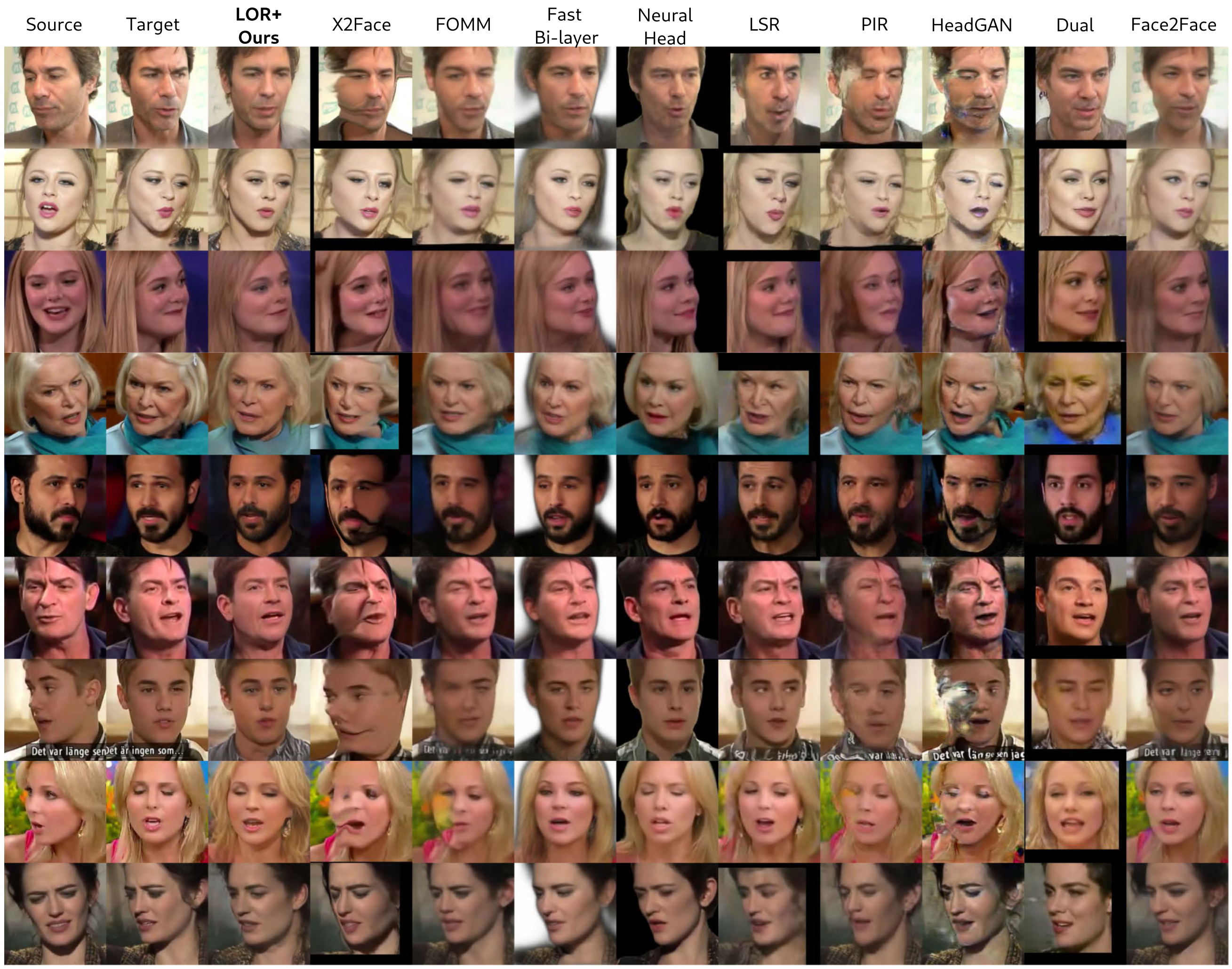


Fig. 23 Qualitative comparisons on the Benchmark-XL with image pairs from VoxCeleb1 and VoxCeleb2 datasets, where the distance on the yaw angle is larger than 30° and on the pitch or roll angles larger than 20°

StyleGAN2 generator trained on FFHQ dataset (Karras et al., 2019). Consequently, in order to fairly compare against these methods, we train our model using synthetically generated images from StyleGAN2 generator trained on FFHQ. We compare against ID-disentanglement (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021) on cross-subject reenactment using synthetic images. Specifically, we use the small test set (1000 images) provided by the authors of StyleFlow (Abdal et al., 2021) and we randomly select 500 image pairs (source and target faces) to perform face reenactment. In Table 13 and in Fig. 24, we show quantitative and qualitative results of our method against ID-disentanglement (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021). As shown in Table 13 our method outperforms all other method both on identity preservation (CSIM) and on head pose/expression transfer metrics, namely ARD, AED and NME. Additionally, as illustrated in Fig. 24, our method can successfully edit the source image given the target head pose/expression, without altering the source identity. On the contrary, ID-disentanglement (ID-dis) method (Nitzan et al., 2020) is not able to preserve the source identity, while StyleFlow (Abdal et al., 2021) fails to faithfully transfer the target head pose and expression.

Table 13 Quantitative comparisons against two state-of-the-art methods for synthetic image editing, namely ID-dis (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021). For CSIM metric, higher is better (↑), while in all other metrics lower is better (↓)

Method	CSIM	ARD	AED	NME
ID-dis (Nitzan et al., 2020)	0.56	2.0	0.12	12.0
StyleFlow (Abdal et al., 2021)	0.67	2.6	0.13	16.0
Ours	**0.80**	**1.1**	**0.09**	**10.1**

Bold values indicate the best values for each metric between the different methods

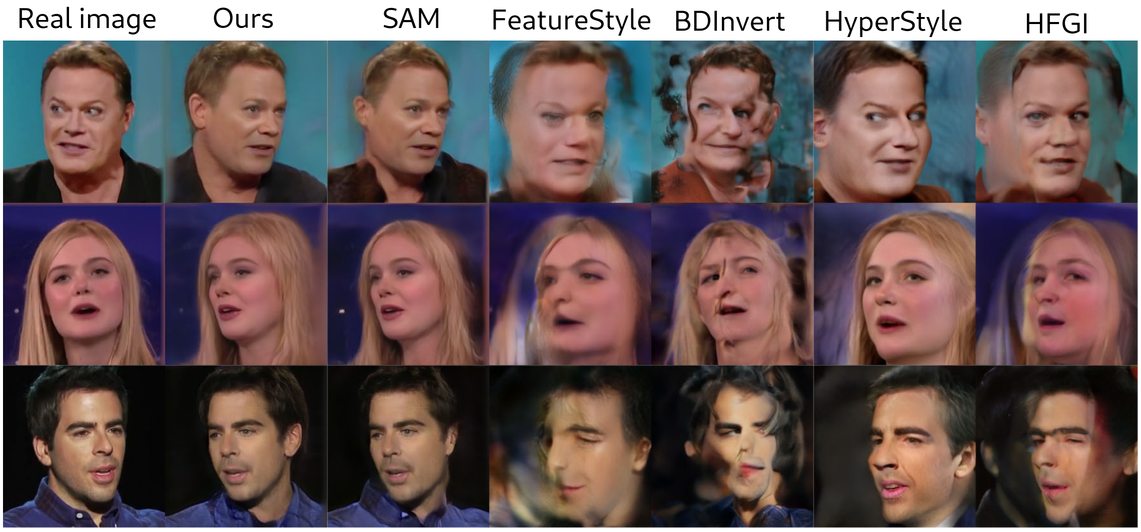


Fig. 25 Qualitative comparison of the proposed framework (Ours) against SAM (Parmar et al., 2022), FeatureStyle (Yao et al., 2022a), BDInvert (Kang et al., 2021), HyperStyle (Alaluf et al., 2022), and HFGI (Wang et al., 2022a) on the task of real image head pose editing

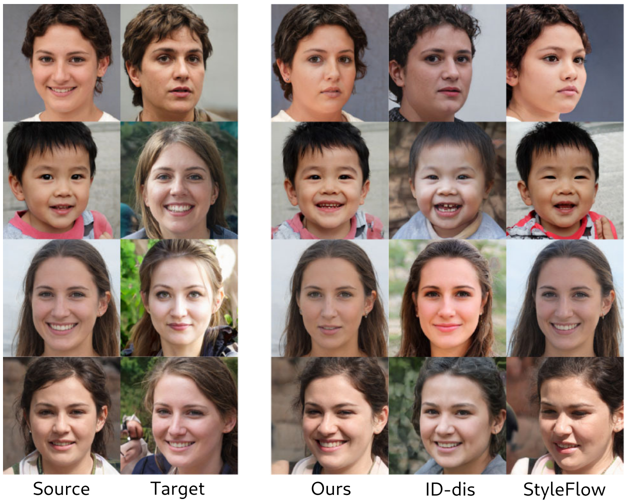


Fig. 24 Qualitative comparisons against ID-disentanglement (ID-dis) (Nitzan et al., 2020) and StyleFlow (Abdal et al., 2021) using random source-target pairs from the small test set provided by the authors of StyleFlow (Abdal et al., 2021)

Table 14 Results of a user study that we conduct to evaluate the user preference (Pref. (%)) on the generated images of state-of-the-art methods

Method	Pref. (%)
X2Face (Wiles et al., 2018)	1.3
FOMM (Siarohin et al., 2019)	5.0
Fast Bi-layer (Zakharov et al., 2020)	9.4
Neural-Head (Burkov et al., 2020)	<u>19.2</u>
LSR (Meshry et al., 2021)	10.7
PIR (Ren et al., 2021)	2.3
LOR+ (Ours)	**52.1**

Bold value indicates the best values for each metric between the different methods
Underlined values indicate the second best results for each metric between the different methods

A.5.2 Comparisons with Real Image Inversion Methods

Additionally, in order to validate that our proposed Feature Transformation module is necessary to perform image editing without producing visual artifacts when altering the feature space of StyleGAN2, we compare our method against four methods that perform real image inversion using the feature space and one method that learns to alter the weights of the StyleGAN2 generator. Specifically, we compare against SAM (Parmar et al., 2022), FeatureStyle (Yao et al., 2022b), BDInvert (Kang et al., 2021), HFGI (Wang et al., 2022a) and HyperStyle (Alaluf et al., 2022). Both SAM (Parmar et al., 2022) and BDInvert (Kang et al., 2021) are optimization-based approaches that refine the feature space of StyleGAN2 to perform real image inversion with better reconstruction quality. Additionally, FeatureStyle (Yao et al., 2022b) is an encoder-based method that simultaneously predicts the inverted latent code $\mathbf{w}$ and feature map F_K at Kth convolution layer of StyleGAN2. Similarly, HFGI (Wang et al., 2022a) predicts both the latent code $\mathbf{w}$ and the spatial feature of StyleGAN2 generator to improve the inversion quality. Finally, HyperStyle (Alaluf et al., 2022) proposes to alter the generator's weights using a hypernetwork. In Fig. 25, we demonstrate results of editing the head pose using our direction matrix $\mathbf{A}$ by first inverting the real images using the above methods. As shown, our method is the only one without visual artifacts when editing the head pose orienta-

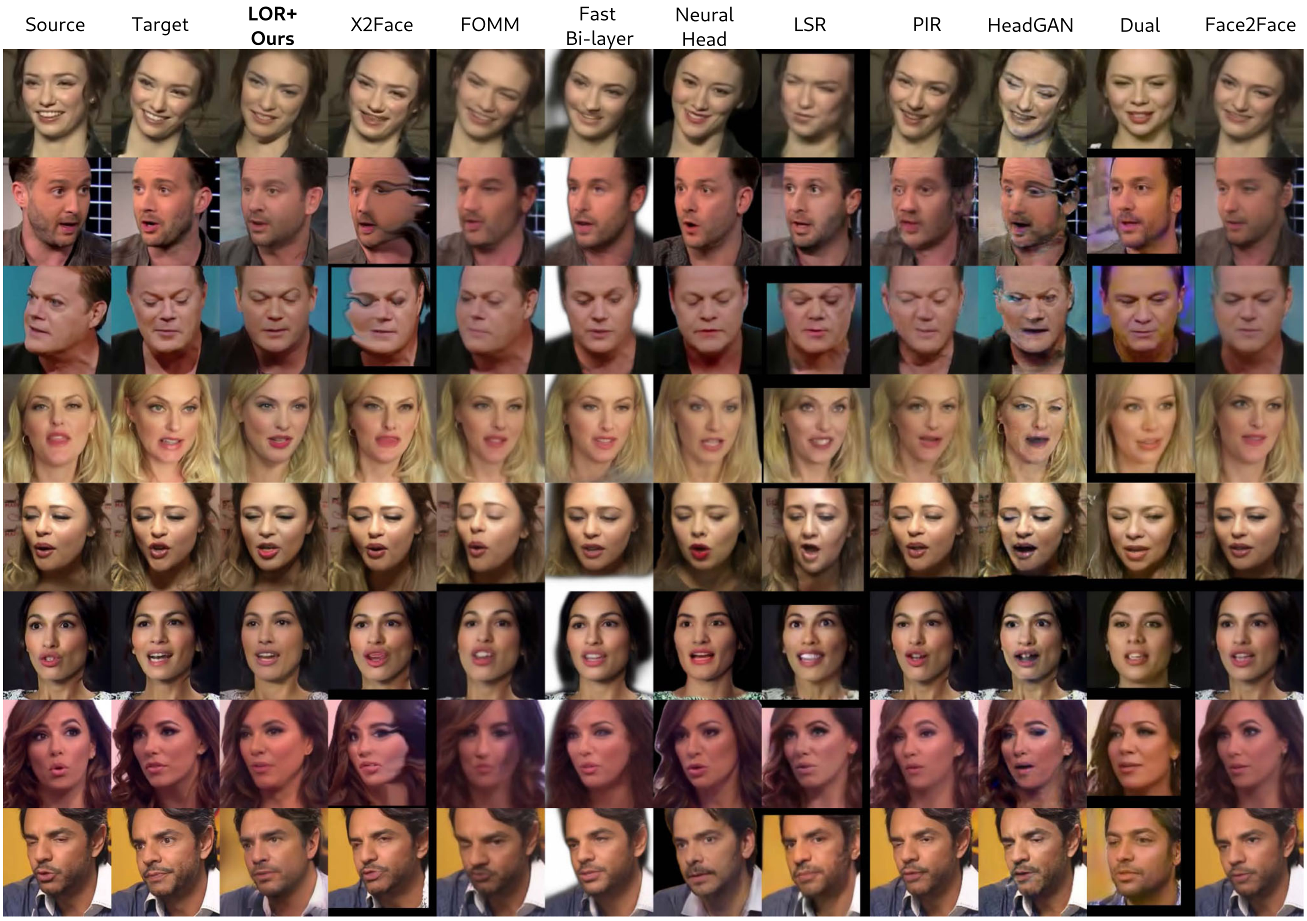


Fig. 26 Qualitative results and comparisons for self-reenactment on VoxCeleb1 (Nagrani et al., 2017) dataset. The first and second columns show the source and target faces. We compare our method against X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast Bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022)

tion. All the aforementioned methods are able to faithfully reconstruct the real images but fail on editing.

A.5.3 Additional Comparisons

In Table 14, we report the results of our user study. Specifically, we ask 30 users to select the method that best reenacts the source frame on self and cross-subject reenactment tasks. For the purposes of the user study we utilise only our final model (LOR+) and we compare against X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021) and PIR (Ren et al., 2021). As shown our method is the most preferable, by a large margin—52.1% versus the 19.2% of the second best method.

We provide additional results on self (Fig. 26) and cross-subject (Figs. 27, 28) reenactment on VoxCeleb1 (Nagrani et al., 2017) dataset and we compare our method with X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022). Moreover, in Fig. 29 we show additional comparisons on VoxCeleb2 (Chung et al., 2018) dataset both on self and on cross-subject reenactment. Additionally, we provide a supplementary video with randomly selected videos on self-reenactment and randomly selected pairs on cross-subject reenactment from the

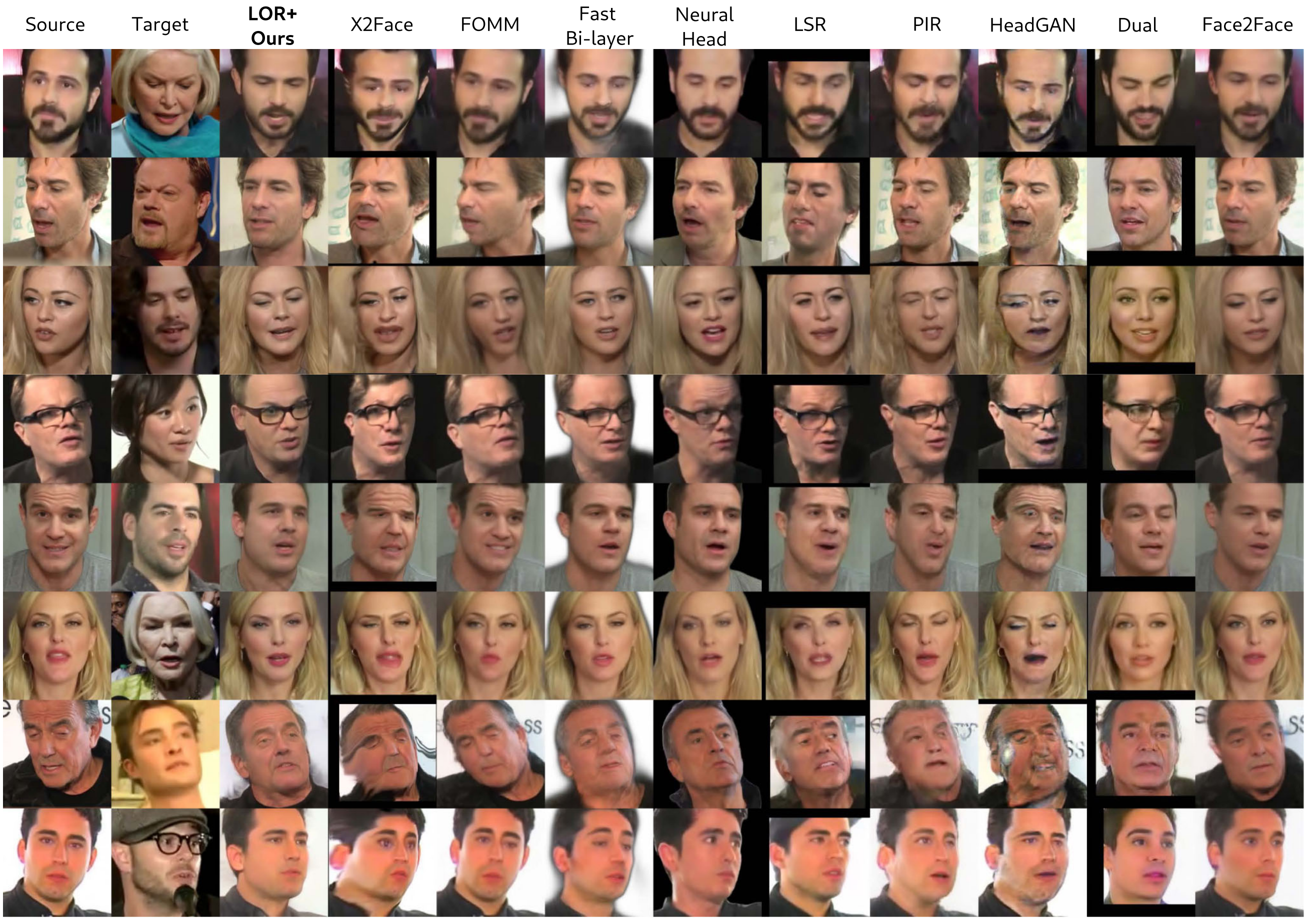


Fig. 27 Qualitative results and comparisons for cross-subject reenactment on VoxCeleb1 (Nagrani et al., 2017) dataset. The first and second columns show the source and target faces. We compare our method against X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast Bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022)

test sets of VoxCeleb1 and VoxCeleb2 datasets. Finally, we show that our method is able to generalise well on other facial video datasets. In Fig. 30 we provide results on Face-Forensics (Rössler et al., 2018) and 300-VW (Shen et al., 2015) datasets both on self (Fig. 30a) and on cross-subject (Fig. 30b) reenactment.

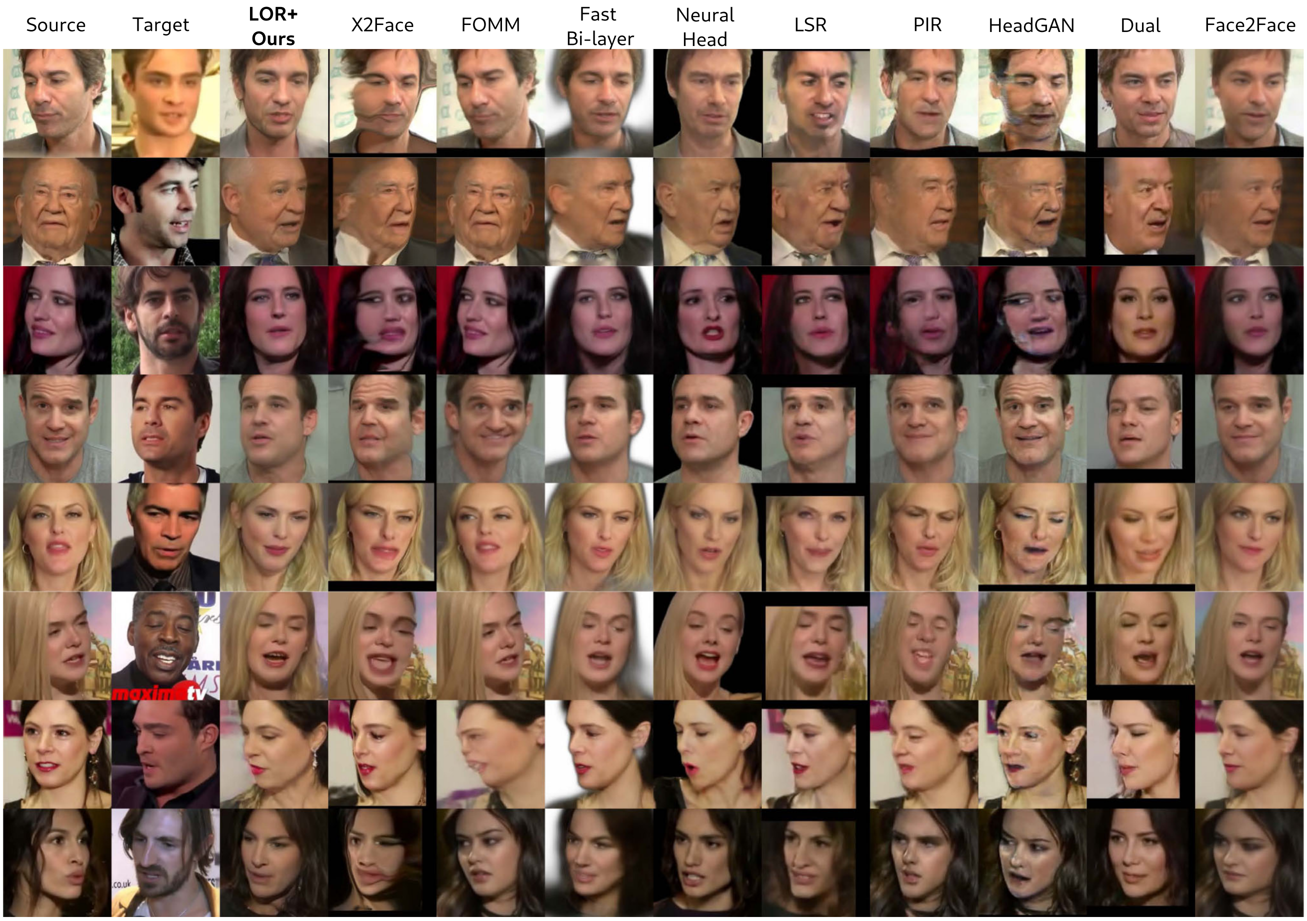


Fig. 28 Qualitative results and comparisons for cross-subject reenactment on VoxCeleb1 (Nagrani et al., 2017) dataset. The first and second columns show the source and target faces. We compare our method against X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast Bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022)

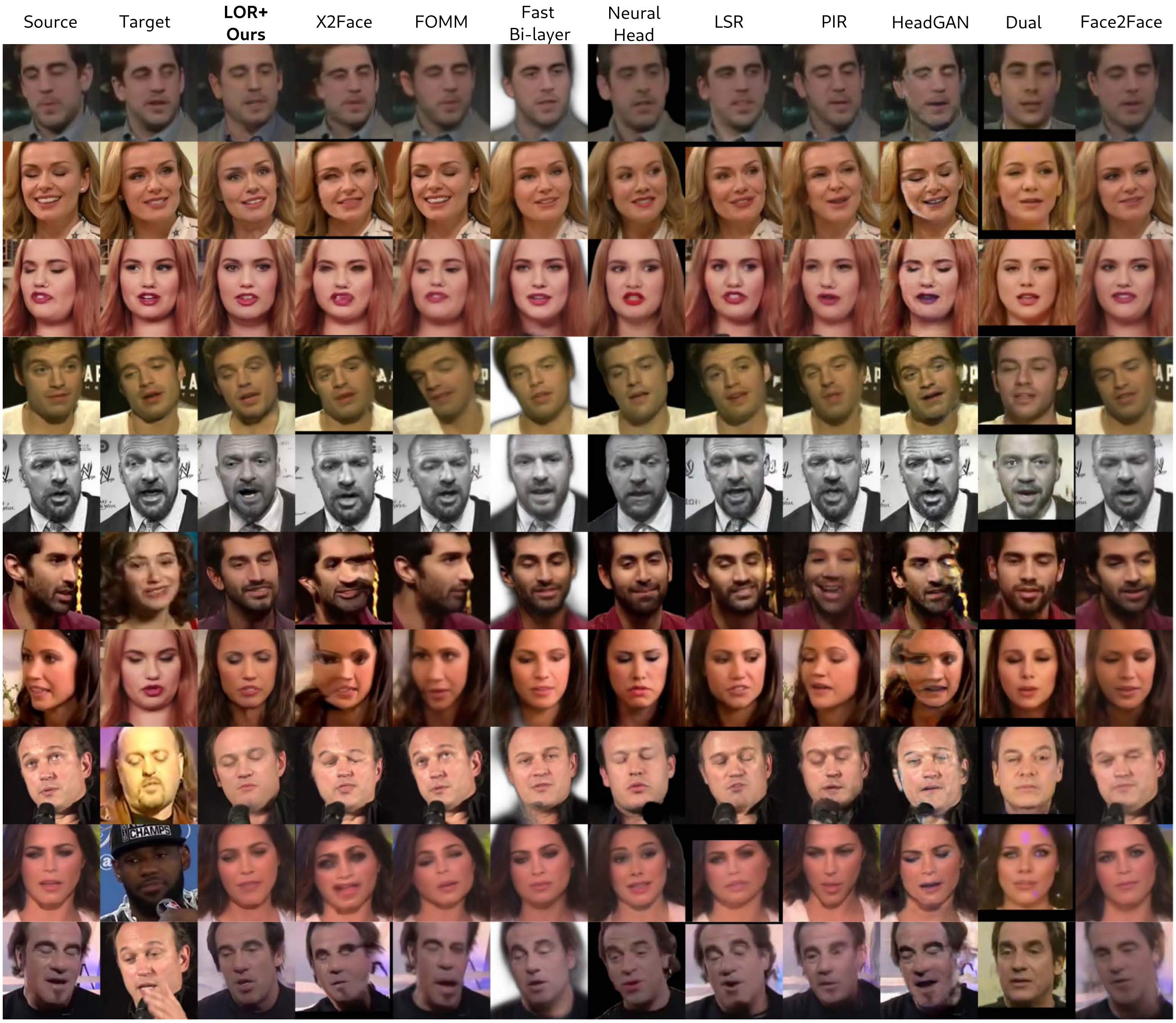


Fig. 29 Qualitative results and comparisons for the tasks of self (first 5 rows) and cross-subject (last 5 rows) reenactment on VoxCeleb2 (Chung et al., 2018) dataset. The first and second columns show the source and target faces. We compare our method against X2Face (Wiles et al., 2018), FOMM (Siarohin et al., 2019), Fast Bi-layer (Zakharov et al., 2020), Neural-Head (Burkov et al., 2020), LSR (Meshry et al., 2021), PIR (Ren et al., 2021), HeadGAN (Doukas et al., 2021), Dual (Hsu et al., 2022) and Face2Face (Yang et al., 2022)

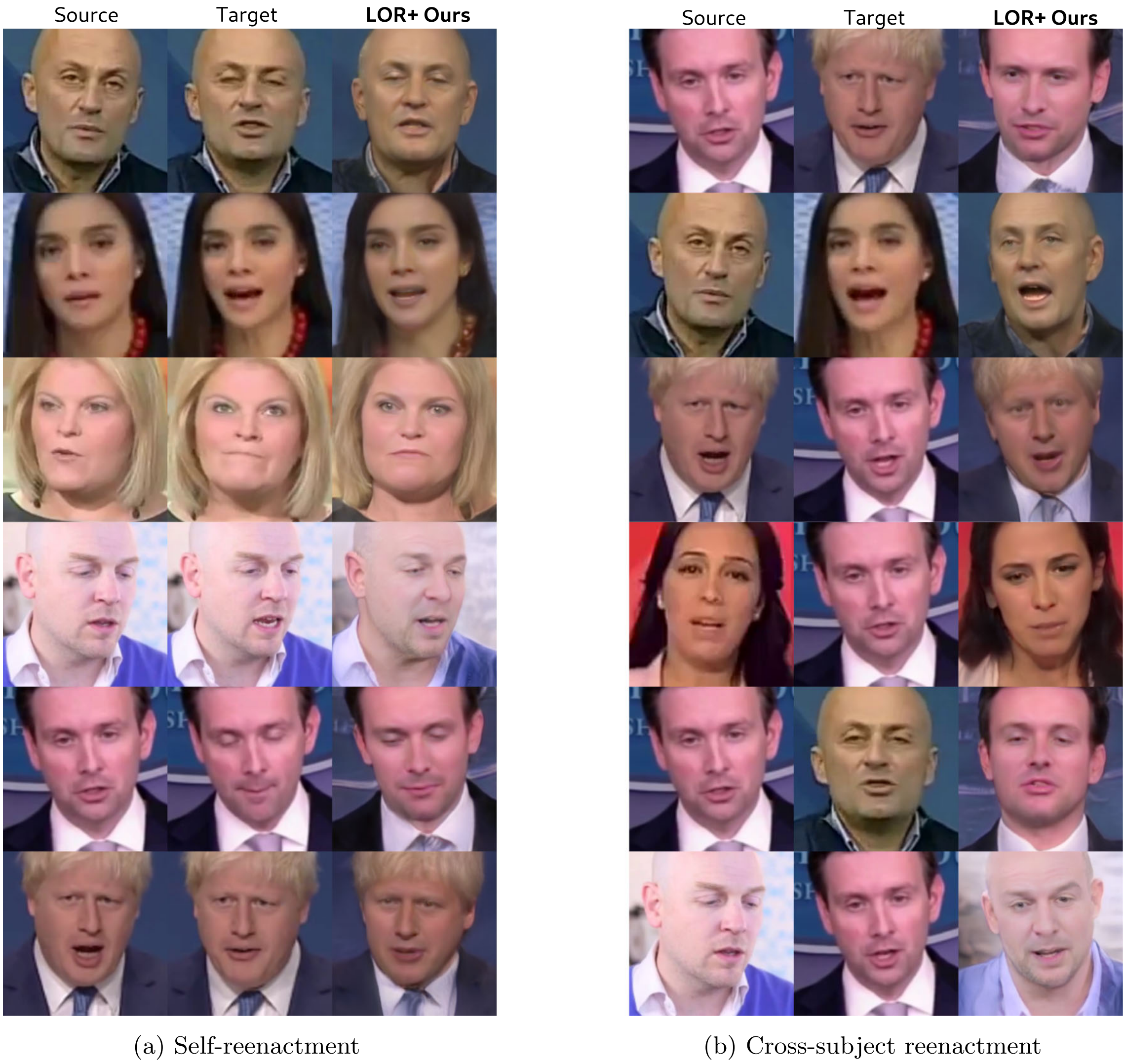


Fig. 30 Qualitative results of our method for self (**a**) and cross-subject (**b**) reenactment on FaceForensics (Rössler et al., 2018) and 300-VW (Shen et al., 2015) datasets

References

Abdal, R., Qin, Y., & Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4432–4441).

Abdal, R., Zhu, P., Mitra, N. J., & Wonka, P. (2021). Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG), 40*(3), 1–21.

Alaluf, Y., Patashnik, O., & Cohen-Or, D. (2021). Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6711–6720).

Alaluf, Y., Tov, O., Mokady, R., Gal, R., & Bermano, A. (2022). Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18511–18521).

Bai, Q., Xu, Y., Zhu, J., Xia, W., Yang, Y., & Shen, Y. (2022). High-fidelity GAN inversion with padding space. In X. V. Part (Ed.), *Computer Vision-ECCV 2022: 17th European conference* (pp. 36–53). Springer.

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (2018). Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6713–6722).

Barattin, S., Tzelepis, C., Patras, I., & Sebe, N. (2023). Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8001–8010).

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Bounareli, S., Argyriou, V., & Tzimiropoulos, G. (2022). Finding directions in GAN's latent space for neural face reenactment. In *British Machine vision conference (BMVC)*

Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I., & Tzimiropoulos, G. (2023). StyleMask: Disentangling the style space of StyleGAN2 for neural face reenactment. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)* (pp. 1–8). IEEE.

Bulat, A., & Tzimiropoulos, G. (2017). 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* (pp. 1021–1030).

Burkov, E., Pasechnik, I., Grigorev, A., & Lempitsky, V. (2020). Neural head reenactment with latent pose descriptors. In: CVPR.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural Information Processing Systems*, *29*.

Chung, J.S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690–4699).

Deng, Y., Yang, J., Chen, D., Wen, F., & Tong, X. (2020). Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5154–5163).

Dinh, T. M., Tran, A. T., Nguyen, R., & Hua, B. S. (2022). Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11389–11398).

Doukas, M.C., Zafeiriou, S., & Sharmanska, V. (2021). Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14398–14407).

Durall, R., Jam, J., Strassel, D., Yap, M. H., & Keuper, J. (2021). Facialgan: Style transfer and attribute manipulation on synthetic faces. In *32nd British machine vision conference* (pp. 1–14).

Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics, 40*(4), 1–13.

Ghosh, P., Gupta, P. S., Uziel, R., Ranjan, A., Black, M. J., & Bolkart, T. (2020) GIF: Generative interpretable faces. In *8th international conference on 3D vision, 3DV 2020, Virtual Event* (pp. 868–878). IEEE.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multipie. *Image and Vision Computing, 28*(5), 807–813.

Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D. (2020). Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10893–10900).

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Hsu, G.S., Tsai, C.H., & Wu, H.Y. (2022). Dual-generator face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 642–650).

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711). Springer.

Kang, K., Kim, S., & Cho, S. (2021). Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13941–13949).

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020a). Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020b). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

Kingma, D.P., Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, Y. LeCun (Eds.) *3rd International conference on learning representations*, ICLR 2015, Conference Track Proceedings.

Kowalski, M., Garbin, S.J., Estellers, V., Johnson, M., & Shotton, J. (2020). Config: Controllable neural face image generation. In *European conference on computer vision (ECCV)*.

Meshry, M., Suri, S., Davis, L.S., & Shrivastava, A. (2021). Learned spatial representations for few-shot talking-head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13829–13838).

Nagrani, A., Chung, J.S., & Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*.

Nitzan, Y., Bermano, A., & Li, Y., & Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. Preprint at arXiv:2005.07728.

Nitzan, Y., Gal, R., & Brenner, O., & Cohen-Or, D. (2021). Large: Latent-based regression through GAN semantics. Preprint at arXiv:2107.11186.

Oldfield, J., Georgopoulos, M., Panagakis, Y., Nicolaou, M. A., & Patras, I. (2021). Tensor component analysis for interpreting the latent space of GANs. In *32nd British machine vision conference 2021*, BMVC 2021 (p. 222).

Oldfield, J., Tzelepis, C., & Panagakis, Y., Nicolaou, M. A., & Patras, I. (2023). PandA: Unsupervised learning of parts and appearances in the feature maps of GANs. In *The eleventh international conference on learning representations, ICLR 2023*, OpenReview.net. https://openreview.net/pdf?id=iUdSB2kK9GY.

Parmar, G., Li, Y., Lu, J., Zhang, R., Zhu, J. Y., & Singh, K. K. (2022). Spatially-adaptive multilayer selection for GAN inversion and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11399–11409).

Paszke, A., Gross, S., Massa, F., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems, 32*, 8026–8037.

Ren, Y., Li, G., Chen, Y., Li, T. H., & Liu, S. (2021). Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13759–13768).

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: A stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2287–2296).

Roich, D., Mokady, R., Bermano, A. H., & Cohen-Or, D. (2021). Pivotal tuning for latent-based editing of real images. Preprint arXiv:2106.05744.

Rössler, A., Cozzolino, D., & Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A large-scale video dataset for forgery detection in human faces.

Sanchez, E., & Valstar, M. (2020). A recurrent cycle consistency loss for progressive face-to-face synthesis. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)* (pp. 53–60). IEEE.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 50–58).

Shen, Y., Yang, C., Tang, X., & Zhou, B. (2020). Interfacegan: Interpreting the disentangled face representation learned by GANs. In *IEEE transactions on pattern analysis and machine intelligence*.

Shen, Y., & Zhou, B. (2021). Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1532–1540).

Shoshan, A., Bhonker, N., Kviatkovsky, I., & Medioni, G. (2021). Gancontrol: Explicitly controllable GANs. Preprint arXiv:2101.02477.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems, 32*, 7137–7147.

Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3626–3636).

Tewari, A., Elgharib, M., Bernard, F., Seidel, H. P., Pérez, P., Zollhöfer, M., & Theobalt, C. (2020). Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics, 39*(6), 1–14.

Tewari A, Elgharib M, Bharaj G, Bernard F, Seidel HP, Pérez P, Zollhofer M, Theobalt C(2020b). Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6142–6151).

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics, 40*(4), 1–14.

Tripathy, S., Kannala, J., & Rahtu, E. (2020). Icface: Interpretable and controllable face reenactment using GANs. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3385–3394).

Tripathy, S., Kannala, J., & Rahtu, E. (2021). Facegan: Facial attribute controllable reenactment GAN. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1329–1338).

Tzelepis, C., Oldfield, J., Tzimiropoulos, G., & Patras, I. (2022). ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences. Preprint arXiv:2206.02104

Tzelepis, C., Tzimiropoulos, G., & Patras, I. (2021). WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6393–6402).

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. Preprint arXiv:1812.01717.

Voynov, A., & Babenko, A. (2020). Unsupervised discovery of interpretable directions in the GAN latent space. In *International conference on machine learning* (pp. 9786–9796). PMLR.

Wang, C., Chai, M., He, M., Chen, D., & Liao, J. (2021a). Cross-domain and disentangled face manipulation with 3d guidance. Preprint arXiv:2104.11228.

Wang, T.C., Mallya, A., & Liu, M.Y. (2021b). One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10039–10049).

Wang, T., Zhang, Y., Fan, Y., Wang, J., & Chen, Q. (2022a). High-fidelity GAN inversion for image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11379–11388).

Wang, Y., Yang, D., Bremond, F., & Dantcheva, A. (2022b). Latent image animator: Learning to animate images via latent space navigation. In *International conference on learning representations*.

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

Yang, H., Chai, L., Wen, Q., Zhao, S., Sun, Z., & He, S. (2021). Discovering interpretable latent space directions of GANs beyond binary attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12177–12185).

Yang, K., Chen, K., Guo, D., Zhang, S. H., Guo, Y. C., & Zhang, W. (2022). Face2face ρ: Real-time high-resolution one-shot face reenactment. In *European conference on computer vision* (pp. 55–71). Springer.

Yao, X., Newson, A., Gousseau, Y., & Hellier, P. (2021). A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13789–13798).

Yao, X., Newson, A., Gousseau, Y., & Hellier, P. (2022a). A style-based GAN encoder for high fidelity reconstruction of images and videos. In *European conference on computer vision*.

Yao, X., Newson, A., Gousseau, Y., & Hellier, P. (2022b). A style-based GAN encoder for high fidelity reconstruction of images and videos. In X. V. Part (Ed.), *Computer Vision-ECCV 2022: 17th European conference*, (pp. 581–597). Springer.

Yao, G., Yuan, Y., Shao, T., & Zhou, K. (2020). Mesh guided one-shot face reenactment using graph convolutional networks. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1773–1781).

Zakharov, E., Ivakhnenko, A., Shysheya, A., & Lempitsky, V. (2020). Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*.

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9459–9468).

Zeng, X., Pan, Y., Wang, M., Zhang, J., & Liu, Y. (2020). Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12757–12764).

Zhang, J., Zeng, X., Wang, M., Pan, Y., Liu, L., Liu, Y., Ding, Y., & Fan, C. (2020). Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5326–5335).

Zhang, R., Isola, P., Efros, A.A., & Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., & Wen, F. (2022). General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18697–18709).

Zhou, H., Liu, J., Liu, Z., Liu, Y., & Wang, X. (2020). Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5911–5920).

Zhu, J., Shen, Y., Zhao, D., & Zhou, B. (2020). In-domain GAN inversion for real image editing. In *European conference on computer vision* (pp. 592–608). Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.