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

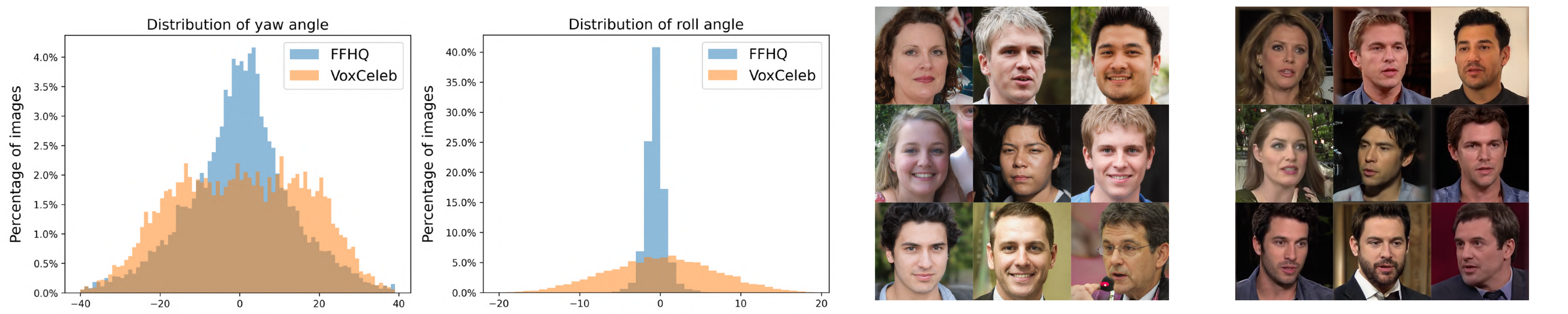
Stella Bounareli[1], Vasileios Argyriou[1], and Georgios Tzimiropoulos[2]
[1]Kingston University, London, UK
[2]Queen Mary University of London, UK

A. Introduction

- SOTA methods for neural face reenactment train generative models to learn disentangled embeddings for identity and facial pose using paired data.
- The main challenges are: a) realistic image generation, b) identity preservation and c) faithful facial pose transfer.
- We present a novel method for face reenactment leveraging the high quality generation of a **pretrained StyleGAN2** and the disentangled properties of a **3D shape model**.
- Our method is able to **create realistic facial images**, and also **faithfully transfer the target head pose and expression**.

B. Preliminaries

1. We finetune StyleGAN2 (trained on FFHQ) on VoxCeleb dataset, which is more diverse in terms of head poses and expressions compared to FFHQ dataset.



FFHQ generated images VoxCeleb generated images

2. We use a 3D shape model [2] to extract the 3D facial model $\mathbf{s}$ and the facial pose parameter $\mathbf{p}$ defined as:

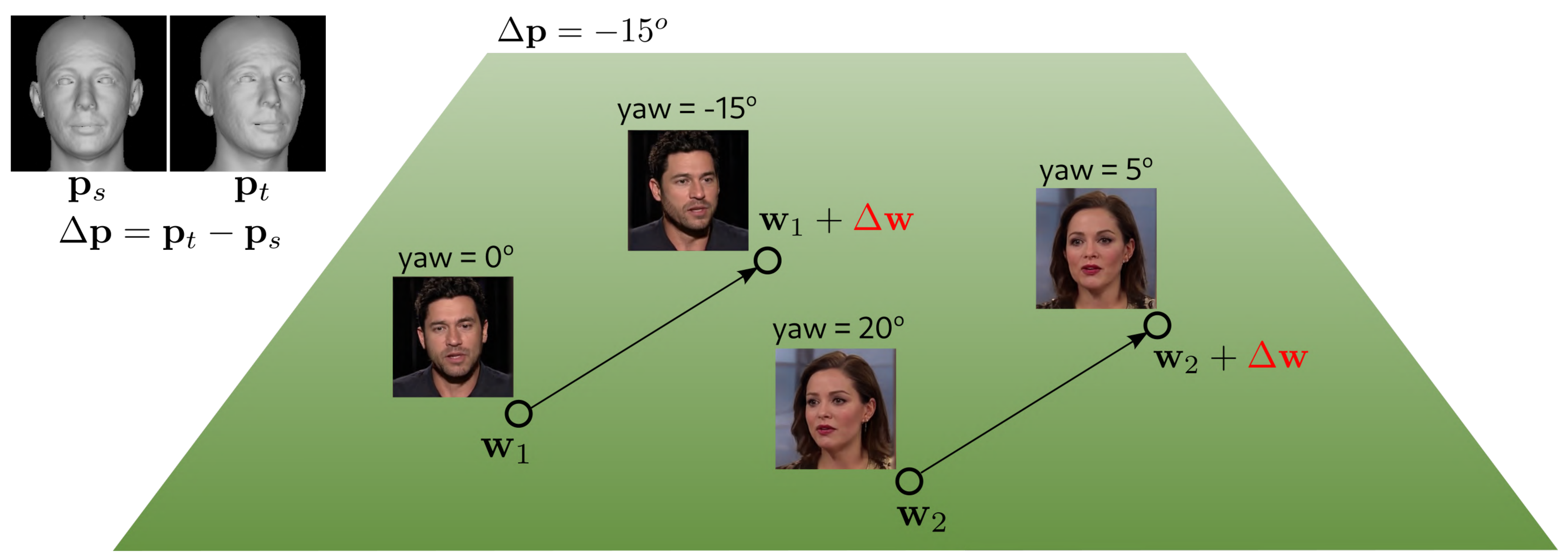
$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}_i\mathbf{p}_i + \mathbf{S}_e\mathbf{p}_e\,, \quad \mathbf{p} = [\mathbf{p}_\theta, \mathbf{p}_e]$$

where $\mathbf{p}_i, \mathbf{p}_e$ are the identity and expression coefficients, and $\mathbf{p}_\theta$ the head orientation.

C. Goal

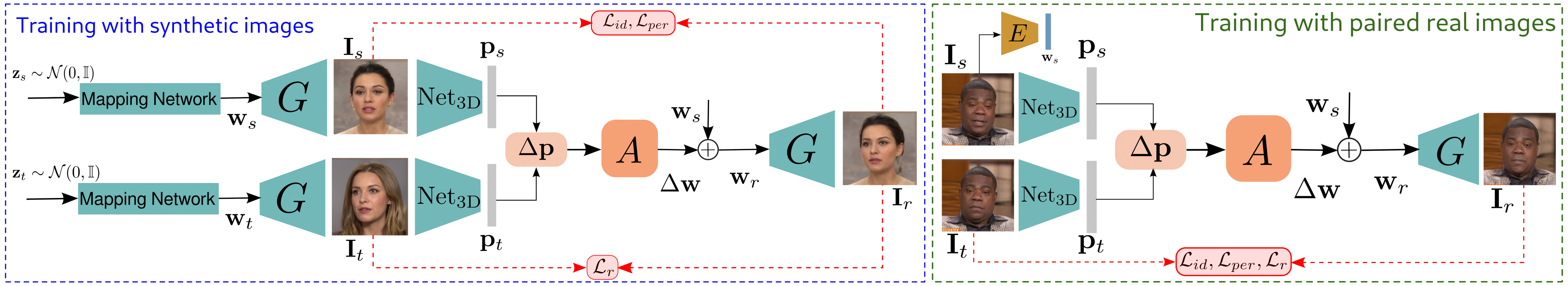
Learn the directions in the latent space of StyleGAN2 that control different facial attributes without altering the identity of the generated face.

We propose to associate a change $\Delta\mathbf{p}$ in the parameter space, with a change $\Delta\mathbf{w}$ in the intermediate latent space $\mathcal{W}+$.



D. Our method

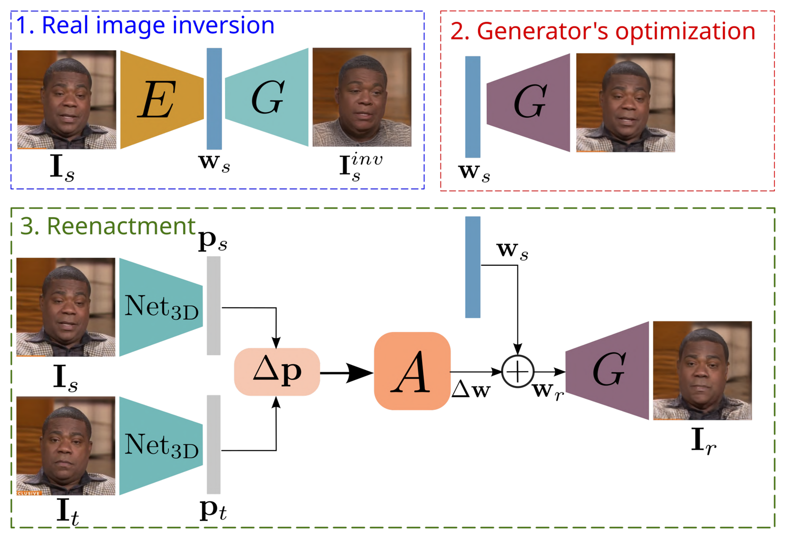
- We train the matrix of directions $\mathbf{A}$, which takes as input the difference of facial pose parameters $\Delta\mathbf{p}$ and outpus a shift vector $\Delta\mathbf{w}$.
-The reenacted image is generated by shifting the source latent code using the predicted shift $\Delta\mathbf{w}$.



Training with synthetic images:
The reenacted image should have:
- the identity of the source image (identity and perceptual losses)
- the target facial pose (reenactment loss)

Training with paired real images:
The reenacted image should have:
- the identity of the target image (identity and perceptual losses)
- the target facial pose (reenactment loss)

E. Inference

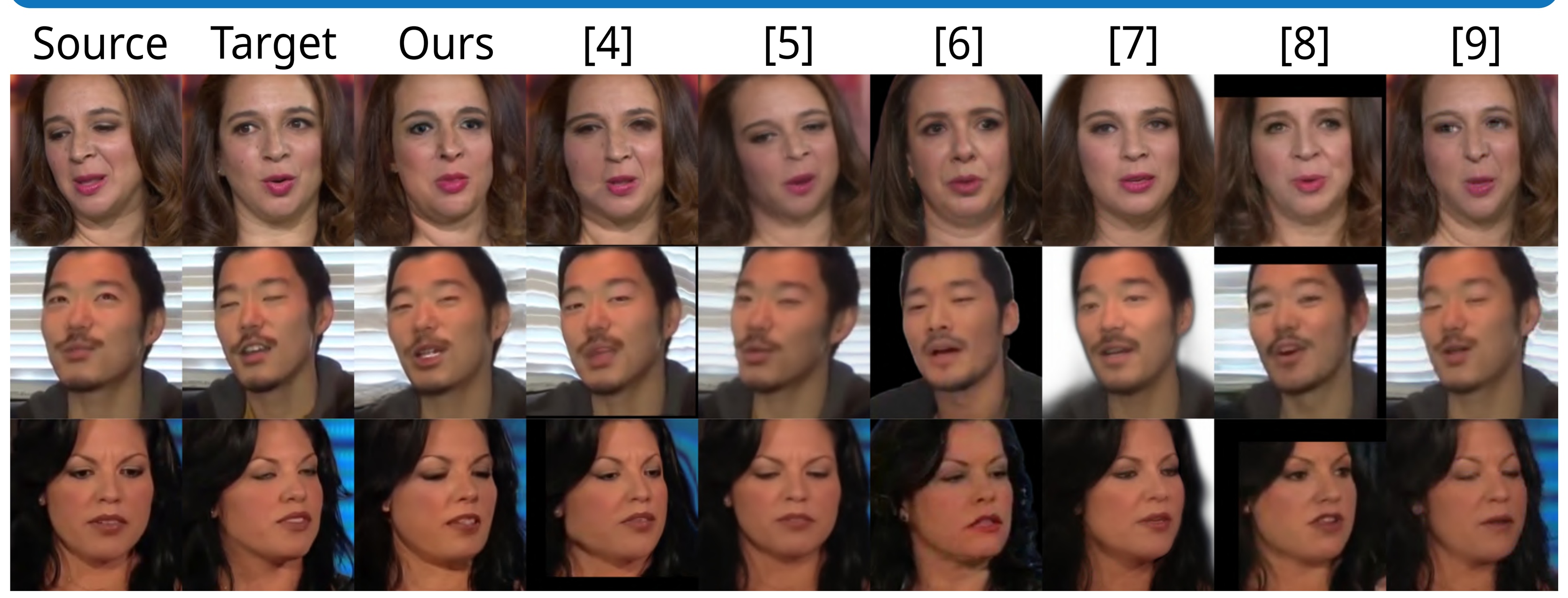


Given a source face and a target video:
1. We invert the source image to get the source latent code $\mathbf{w}_s$.
2. We finetune the generator to get a better reconstruction result [3].
3. We reenact the source face given a target pose.

F. Quantitative Results

	Self Reenactment							Cross Reenactment		
Method	CSIM	LPIPS	FID	FVD	NME	Pose	Exp.	CSIM	Pose	Exp.
X2Face [4]	<u>0.70</u>	0.13	<u>35.5</u>	409	17.8	1.5	0.90	0.57	2.2	1.5
FOMM [5]	0.65	0.14	35.6	<u>402</u>	34.1	5.0	1.3	**0.73**	7.7	2.0
Fast Bi-layer [6]	0.64	0.23	52.8	634	**13.2**	<u>1.1</u>	0.80	0.48	1.5	1.3
Neural-Head [7]	0.40	0.22	98.4	587	15.5	1.3	0.90	0.36	1.7	1.6
LSR [8]	0.59	0.13	45.7	464	17.8	**1.0**	<u>0.75</u>	0.50	<u>1.4</u>	<u>1.2</u>
PIR [9]	**0.71**	<u>0.12</u>	57.2	414	18.2	1.86	0.94	0.62	2.2	1.4
Ours	0.66	**0.11**	**35.0**	**345**	<u>14.1</u>	<u>1.1</u>	**0.68**	<u>0.63</u>	**1.2**	**1.0**

G. Qualitative Results (I)



Self reenactment: Source and target images have the same identity.

G. Qualitative Results (II)

Cross-subject reenactment: Source and target images have different identities.

Facial image editing: Only one facial attribute (yaw, pitch, smile etc.) is edited, without altering the identity and any other attribute of the source face (shown inside the red box).

[1] Tov et al., Designing an encoder for stylegan image manipulation. ACM TOG, 2021
[2] Feng et al., Learning an animatable detailed 3D face model from in-the-wild images. ACM TOG, 2021
[3] Roich et al., Pivotal tuning for latent-based editing of real images. ACM TOG, 2021
[4] Wiles et al., X2face: A network for controlling face generation using images, audio, and pose codes. ECCV, 2018
[5] Siarohin et al., First order motion model for image animation. NeurIPS, 2019
[6] Zakharov et al., Fast bi-layer neural synthesis of one-shot realistic head avatars. ECCV, 2020
[7] Burkov et al., Neural head reenactment with latent pose descriptors. CVPR, 2020
[8] Meshry et al., Learned spatial representations for few-shot talking-head synthesis. ICCV, 2021
[9] Ren et al., Pirenderer: Controllable portrait image generation via semantic neural rendering. ICCV, 2021

Project Page