Unified Application of Style Transfer for Face Swapping and Reenactment

Lê Minh Ngô[✉1,2,†][0000−0002−3810−1136], Christian aan de Wiel[1,2,†][0000−0003−0156−808X], Sezer Karaoğlu[1,2][0000−0001−9073−9420], and Theo Gevers[1,2][0000−0002−1190−5492]

[1] 3DUniversum, Amsterdam, The Netherlands
[2] Computer Vision Lab, University of Amsterdam, Amsterdam, The Netherlands
{l.m.ngo,th.gevers}@uva.nl, {m.c.aandewiel,s.karaoglu}@3duniversum.com

Abstract. Face reenactment and face swap have gained a lot of attention due to their broad range of applications in computer vision. Although both tasks share similar objectives (e.g. manipulating expression and pose), existing methods do not explore the benefits of combining these two tasks.

In this paper, we introduce a unified end-to-end pipeline for face swapping and reenactment. We propose a novel approach to isolated disentangled representation learning of specific visual attributes in an unsupervised manner. A combination of the proposed training losses allows us to synthesize results in a one-shot manner. The proposed method does not require subject-specific training.

We compare our method against state-of-the-art methods for multiple public datasets of different complexities. The proposed method outperforms other SOTA methods in terms of realistic-looking face images.

1 Introduction

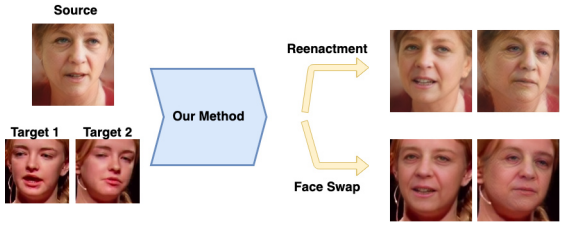

Fig. 1. Our algorithm takes source and target images and produces reenacted and swapped face results using a single unified pipeline.

† Equal contribution, alphabetic order.

Generating images or videos by manipulating facial attributes (i.e. face reenactment and swapping) has gained a lot of attention in recent years due to their broad range of computer vision and multimedia applications such as video dubbing [1], gaze correction [2], actor capturing [3, 4], and virtual avatar creation [5].

Face reenactment [3, 6] aims to manipulate facial attributes such as expression, pose or gaze of a video or a single image, whereas face swap [7, 8] tries to seamlessly replace a face from a source image with a target face while maintaining the realism of the facial appearance. To perform such transfer, face swap techniques manipulate face attributes such as expression, pose, and identity. Although the face attribute manipulation for both face reenactment and face swap is similar, they have never been considered in a unified pipeline. To this end, in this paper, we propose a single unified model for both face swapping and reenactment tasks allowing the model to produce a more robust face representation and exploiting the constraints from the two tasks to improve the realism of facial appearances.

Before the introduction of deep neural networks, face reenactment and swapping are typically solved by 3D modeling [3, 7, 9–11]. The 3D face image is transformed into a 3D representation, where latent parameters of the 3D representation are manipulated and projected back in a 2D space. Although those methods produce results with high realism, they are not able to generalize well on unseen data. Hence, for each target face the model parameters have to be tuned.

Current generative models make it feasible to synthesize realistic-looking images [12, 13]. Consequently, recent research is focused on improving the quality of the *face image generation* process [12–14] using generative models. Only a few methods explore the direction of using generative models for face reenactment or face swapping. Although these tasks share similarities, previous methods only focus on solving one of the two tasks independently and are supervised [15–17]. Recently methods show that face swap targeted methods can be used for face reenactment and vice versa. Unfortunately, the visual results on the second task are typically inferior to the first one [6, 8]. Since those methods are designed for one of the tasks separately, they are not optimal for both. In contrast to existing methods, we integrate both tasks into one combined model. To our knowledge, our method is the first unsupervised method designed to perform both tasks in a unified end-to-end manner.

In this paper, we propose a novel pipeline that unifies face swapping and reenactment (Fig. 1). A combined approach benefits from the similarities of the two tasks. Learning them together allows for robust face representation and enhances the realism of facial appearance. The proposed algorithm learns an isolated disentangled representation for face attributes without any supervision. Hence, our model can manipulate expression/pose, face identity, and style independently in latent space. We achieve this by directly mapping the disentangled latent representation to the latent space of a pre-trained generator. During inference time, the encoders condition the latent space by source and target face images together with their landmarks and generate the reenacted or swapped

face using the pre-trained decoder. Prediction is done in a one-shot manner (i.e. only a single image of a person is required). The model's training loss incorporates contextual and identity losses to preserve the face identity, regardless of the source face. As a result, our model obtains visually more appealing results in cross-gender face swapping compared to the baselines.

We evaluate our method on multiple datasets of various complexities: 300VW with videos of talking people [18], and UvA-NEMO with spontaneous and fake smiles in a controlled environment [19]. Experiments demonstrate that our method (on average) performs better on face reenactment and face swapping tasks than existing state-of-the-art methods focusing only on a single task.

To summarize, our contribution is four-fold:

- A novel method is proposed to perform face swapping and reenactment tasks in a joint manner. To our knowledge, our method is the first method to jointly perform the two tasks in a unified end-to-end architecture.
- The proposed method is *subject agnostic*: it does not require subject-specific training.
- A novel approach is proposed to learn an isolated disentangled representation for single visual attributes (i.e. the expression/pose, identity, and style) by using a pre-trained generator with a disentangled latent space. This allows for a full control over the face manipulation process in an unsupervised manner. Hence, our approach does not require ground truth data for expression/pose, identity, and style learning of reenactment outputs.
- A combination of training losses allows us to synthesize results in a one-shot manner and to outperform competitive methods in cross-gender face manipulation.

2 Related Works

2.1 Generative Models

Generative models based on Generative Adversarial Networks (GANs) are advantageous for the task of image synthesis [20, 21]. However, until recently, those models can be considered as black boxes with latent representations which are hard to interpret. In addition, the realism of the generated results, in particular for face image synthesis [12, 14], is limited (with artifacts in identity preservation).

Recently, StyleGAN [13] introduces a novel way to condition the latent code through an affine transformation, corresponding to a specific style [22], by using Adaptive Instance Normalization (AdaIN) [23]. AdaIN allows the model to generate images with more realistic face appearance compared to previous methods [24]. Furthermore, the aforementioned architecture modifications, combined with a revised training approach [13, 22], enable the separation of high-level and stochastic attributes making the latent representation easier to interpret. Hence, the face attributes of a generated image can be changed accordingly by manipulating the latent representation (i.e. disentanglement property). Recent

methods integrate StyleGAN into different applications as a pre-trained network for face enhancement and animation [25]. The state-of-the-art StyleGAN2 [22] enhances the architecture of StyleGAN by redesigning normalization flow and by applying the same network topology for low and high resolution. Image2StyleGAN [26] proposes a method to map an existing image to the latent representation of StyleGAN by iteratively optimizing a latent code to minimize the loss function. Mapping an image to latent space enables a user to change specific image attributes provided by the StyleGAN latent space. However, this method has a drawback in terms of efficiency and generalization: each new image is optimized separately until convergence to obtain a corresponding latent space limiting the applicability of the method for real-time applications. In contrast, our novel isolated disentangled representation learning method solves this problem by introducing encoders that learn to map the desired facial attributes to the corresponding changes in the latent representation. By constraining the mapping by encoders and by using a specific unsupervised training procedure, our approach manipulates the latent space in such a way that it is able to mix disentangled expression/pose, identity and style attributes in a robust manner.

2.2 Face Reenactment

Face reenactment focuses on changing attributes of the face image while keeping the face identity the same. Prior methods focus on different facial attributes like expression [14, 12], skin color [12], lighting [27] and pose, or a combination of those [3]. These methods are mostly used in applications such as virtual avatar or puppeteering, targeting high realistic-looking faces but ignoring background preservation [5]. Other approaches focus more on video dubbing and deepfake generation, preserving the realism for both the foreground and background of the scene [3, 4, 14]. Attribute conditioning is modeled by using different modalities like facial landmarks [28], action units [14] and 3D morphable models [4] for pose and/or expression, and spherical harmonics for lighting [27]. Some methods simplify attribute inference by conditioning directly on the face image. In contrast, our method uses a face image to condition identity and style together with facial landmarks for pose and expression.

Several methods perform face reenactment by manipulating the latent space [26, 29, 30]. [26] compute the relative difference between two images by calculating the differences in their latent spaces and applied to the source latent code afterwards. [29] propose an approach capable of reenacting faces by using encoders to compute the representations of pose, expression, style and identity in the latent space.

In combination with the use of a pre-trained generator, we aim to condition the generator in a one-shot manner during both training and testing time. InterFaceGAN [30] on the other hand, does use a pre-trained generator, but computes latent codes based on attribute scores (e.g. smile, glasses, gender etc.) making it a supervised method. Since our approach does not require ground-truth labels or attribute scores, we have full control over the face manipulation process using only a minimal amount of training data.

Methods that focus on the quality of images and identity/background preservation are typically target-specific [4]. Hence, the model is trained for a particular scene with a single face identity. Other non-target, one-shot methods [12, 14] produce decent results, but they fail in producing consistent face identities between images of the same person (video sequence) [6, 12]. Our method is also a one-shot method. However, in contrast to previous methods, the aim of our method is to produce realistic-looking faces with identity-preservation by exploiting the disentanglement property of the pre-trained model.

2.3 Face Swapping

Face swapping aims to change the facial identity but to keep other face attributes constant. Applications range from face identity obfuscation [31] to recreation [32] and entertainment [7]. Recent methods obtain realistic results by using GANs [8, 15, 17] conditioning identity attributes using either a face image or its facial landmarks. Besides, face segmentation is usually required to position a generated face on the original face [7, 8].

Most face reenactment and swapping approaches rely on the use of generative adversarial networks [12, 14, 26, 29, 33, 34]. A major drawback of the aforementioned methods is their training process, the interpolation quality and lack of disentanglement.

Despite the similarity between face reenactment and face swapping tasks, there are no methods, to the best of our knowledge, which successfully unify these tasks. [6] mainly focuses on the problem of face reenactment, but shows inferior results on the task of face swapping. [8] shows the opposite. This approach is mainly focused on face swapping, but the results on face reenactment lack realistic-looking appearance. Moreover, those methods are complex and multi-staged. Thus, [7] proposes four separate GANs for reenactment, segmentation, inpainting and blending. [6] uses a separate motion network to extract dense optical flow and requires an extra segmentation network for face swapping. In contrast, to our knowledge, we are the first method to unify face reenactment and face swap in one single unified pipeline.

3 Proposed Method

An overview of our method is shown in Fig. 2. Our goal is to produce a face image $\hat{\mathbf{x}}$ while predicting identity attributes w_i, style attributes w_s and pose/facial expression attributes w_{pe} from a given face image and its landmarks. We propose a novel isolated disentangled representation learning algorithm to separate w_i, w_s and w_{pe}. Using the proposed algorithm, attributes of the source and target images can be manipulated in the latent space via mixing using linear addition, since changing one attribute doesn't influence another due to their isolation. For the `face swapping` task, w_i is taken from the source image, while other attributes are taken from the target image. For the `face reenactment` task, w_{pe} and w_s are taken from the source image, keeping identity w_i from the target.

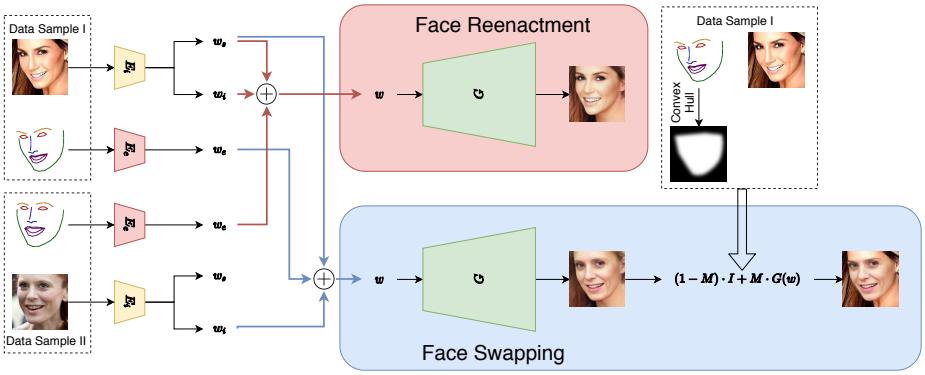

Fig. 2. Our architecture combines Face Swapping and Reenactment into a single unified pipeline with the help of our novel isolated disentangled representation learning algorithm.

3.1 Disentanglement Property and Vector Computations

Our encoders are trained to compute a latent code in the latent space $w \in \mathcal{W}^+$ of a pre-trained generator. Since the latent space is disentangled, face attributes can be manipulated by using vector arithmetics in $\mathcal{W}^+$ [26]. For example, given an image N_A and its latent code w_1 (person A with a neutral expression), N_B and its latent code w_2 (person B with a neutral expression) and another image S_B with a latent code w_3 (person B smiling), it's possible to generate an image of a person A smiling by conditioning the generator G on a latent code $G\big(w_1 + (w_3 - w_2)\big)$.

Our method uses that principle by predicting isolated latent codes for style w_s, identity w_i and pose/expression w_{pe} based on the input image and its corresponding landmarks, assuming those latent codes to be with a disentanglement property. Final latent code can be constructed via linear addition of the three isolated components $w = \mu_G + w_i + w_s + w_{pe}$, where μ_G is the mean of the generator's latent space $\mathcal{W}^+$ with disentanglement property [22, 26].

Since w is constructed from the latent codes w_s, w_i, and w_{pe}, full control is obtained for changing the style, identity, pose and expression of the resulting image I, by exploiting the high-quality images produced by the pre-trained generator. Note that our method allows for subject-agnostic face manipulation executed in a one-shot fashion during inference.

3.2 Architecture

The source face and the target face (together with its facial landmarks) are used as inputs to the two separate encoders E_i and E_{pe} respectively. These encoders approximate a latent code for face style w_s, identity w_i and pose/expression w_{pe}. The network latent space is manipulated using encoder outputs to obtain either

face swap by swapping the identity latent code or face reenactment by swapping the pose latent code. All latent codes are combined into the final latent code w. Then, w is fed to a decoder G to produce the final visual result. In the case of face swapping, a face mask M is generated by using the convex hull of the landmarks [35].

Encoders. Our architecture contains two different types of encoders: (1) the identity encoder E_i, and (2) the pose encoder E_{pe}. These encoders predict a latent code $w \in \mathcal{W}^+$ corresponding to either the identity, style, or pose of the input image.

For the design of the architecture, we base our encoders on the encoder of Pix2Pix [36]. To map the input images and landmarks to their corresponding latent codes, we add n separate fully connected blocks to the architecture, where n is the first dimension of the extended latent space. This fully connected blocks consist of 2 fully connected layers. E_i contains 2 of these fully connected block sets, for style (w_s) and identity (w_i) respectively.

Identity encoder $E_i(\mathbf{x})$ takes an input image $\mathbf{x}$ and estimates the identity latent code $w_i \in \mathcal{W}^+$ and style latent code $w_s \in \mathcal{W}^+$. Latent code w_i is trained to contain only pose- and expression-invariant identity features of the person.

Pose encoder $E_{pe}(\mathbf{x}_s)$ uses the facial landmarks of $\mathbf{x}$ denoted by $\mathbf{x}_s$ as an input. $E_{pe}(\mathbf{x}_s)$ predicts a latent code $w_{pe} \in \mathcal{W}^+$ containing both the pose and expression of $\mathbf{x}_s$. The landmarks are represented as RGB images of landmark boundaries [34].

$$w_i^{\mathbf{x}}, w_s^{\mathbf{x}} = E_i(\mathbf{x}), \quad w_{pe}^{\mathbf{x}} = E_{pe}(\mathbf{x}_s). \tag{1}$$

Decoder. Generator $G(w)$ is a pre-trained network with fixed weights. It takes a latent code $w \in \mathcal{W}^+$ as an input. Here $\mathcal{W}^+$ is the latent space of $G(w)$. $G(w)$ generates an image $\hat{\mathbf{x}}$ corresponding to latent code w. In this paper, the StyleGANv2 architecture is used. However, other models with similar disentanglement properties and continuous latent spaces can be used instead.

3.3 Face Reenactment and Swapping

The reconstructed original face is defined as a function $G(w)$ over its identity $w_i^{\mathbf{x}}$, style $w_s^{\mathbf{x}}$ and expression/pose $w_{pe}^{\mathbf{x}}$ parameters:

$$\hat{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{x}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{x}}). \tag{2}$$

Face Reenactment. Faces are reenacted by changing the expression and pose parameters $w_{pe}^{\mathbf{y}}$ to the pose/expression shown in the target image $\mathbf{y}$ and keeping other parameters identical $w_i^{\mathbf{x}}$ and $w_s^{\mathbf{x}}$. Since $w_i^{\mathbf{x}}$, $w_s^{\mathbf{x}}$ and $w_{pe}^{\mathbf{x}}$ parameters are separated, the resulting image $\dot{\mathbf{x}}$ is defined as a function of their sum:

$$\dot{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{x}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{y}}). \tag{3}$$

Face swapping is performed by keeping the $w_s^{\mathbf{x}}$ and $w_l^{\mathbf{x}}$ parameters unchanged and to modify the identity latent code to $w_i^{\mathbf{y}}$:

$$\tilde{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{y}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{x}}). \tag{4}$$

To swap faces, a facial mask $\mathbf{M}$ is obtained by computing a convex hull of the landmarks and to add a Gaussian blur [35]. The final swapped face is generated by interpolation $(1 - \mathbf{M}) \cdot \mathbf{x} + \mathbf{M} \cdot \tilde{\mathbf{x}}$.

3.4 Losses

The objective function, to train our unified face swapping/reenactment architecture, consists of 5 terms: reconstruction loss $\mathcal{L}_{MSE}$, perceptual loss $\mathcal{L}_{per}$, landmark loss $\mathcal{L}_L$, identity losses for the aligned reconstructed image $\mathcal{L}_{id}^a$ and for the unaligned identity-swapped/reenacted image $\mathcal{L}_{id}^u$. Those terms are weighted using hyperparameters $\lambda_i,\ \mathrm{i} \in \{1..5\}$.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{per} + \lambda_3 \mathcal{L}_L + \lambda_4 \mathcal{L}_{id}^a + \lambda_5 \mathcal{L}_{id}^u. \tag{5}$$

Reconstruction and Perceptual Losses We compute the mean squared error between input and predicted images as a reconstruction loss for efficient color embedding. $\mathcal{L}_{MSE}$ is calculated for the reconstructed image $\hat{\mathbf{x}}$ and the identity-swapped image $\tilde{\mathbf{x}}$. This loss function mainly helps to isolate w_s ensuring a proper color embedding. To capture finer features, the LPIPS distance is used [22, 37, 38]. L_{per} is taken as the reconstruction loss and is calculated only for the reconstructed image $\hat{\mathbf{x}}$.

$$\mathcal{L}_{MSE} = \| \hat{\mathbf{x}} - \mathbf{x} \|_2^2 + \| \tilde{\mathbf{x}} - \mathbf{x} \|_2^2, \qquad \mathcal{L}_{per} = \mathrm{LPIPS}(\hat{\mathbf{x}}, \mathbf{x}). \tag{6}$$

Landmark Loss The landmark loss term is used to isolate pose and expression from identity and style. A pre-trained facial landmark extraction network ψ [39] is taken to extract the landmark heatmaps from an image $\mathbf{x}$. The loss function attempts to minimize the L_2 distance between the extracted heatmaps of the facial landmarks of the source image $\mathbf{x}$ and the target image $\mathbf{y}$, while keeping the latent code for identity and style identical. Landmarks do contain identity (e.g. eye and mouth shape). This means that landmark loss adds an identity bias to the resulting image.

We separate the heatmap sets into two different sets, the expression landmarks ψ_E and the jaw landmarks ψ_J. Parameters γ_1, γ_2 adjust the importance of these landmark sets respectively.

$$\mathcal{L}_L = \gamma_1 \| \psi(\dot{\mathbf{x}})_E - \psi(\mathbf{y})_E \|_2^2 + \gamma_2 \| \psi(\dot{\mathbf{x}})_J - \psi(\mathbf{y})_J \|_2^2 . \tag{7}$$

Identity Loss The identity loss [29, 40] isolates identity in a separate latent code w_i. The layer activations are used of a pre-trained identity recognition network Φ [41]. For our purpose, we use activations $l \in L$ of two specific convolution layers and the last two fully connected layers.

The identity loss is applied to the convolution layers by calculating the contextual loss [42] $\mathcal{L}_{id}^{a}$ over these layers. Note that this will only work for images with the same pose (the reconstructed image), since the convolutions do not capture rotations properly.

$$\mathcal{L}_{id}^{a} = \sum_{l \in L} \parallel \mathrm{CX}(\Phi(\hat{\mathbf{x}}), \phi(\mathbf{x}) \parallel_2^2 . \tag{8}$$

To ensure correct identity in the reenacted frames, a loss function is required to detect the identity of a face independent of the pose. A mean squared error is calculated for the activations of the fully connected layers of Φ. During training, faces are reenacted with random landmarks from the dataset making our approach more robust to landmark biases.

$$\mathcal{L}_{id}^{u} = \sum_{l \in L} \left(\parallel \Phi(\dot{\mathbf{x}}, l) - \Phi(\mathbf{x}, l) \parallel_2^2 + \parallel \Phi(\tilde{\mathbf{x}}, l) - \Phi(\mathbf{y}, l) \parallel_2^2 \right) \tag{9}$$

3.5 Training Details

We trained both our method and the pre-trained generator on the subset of 183K images from the CelebA face dataset [43]. Faces are detected using the Dlib [44]. Face bounding boxes are computed based on an expanded by 10% bounding boxes over facial landmarks [39] and resized to 128×128. Parameters of the network were optimized using the Adam optimizer with a learning rate of 10^{-5} for 100 epochs, batch size = 4. In our experimental setup, we used $\lambda_1, \lambda_2 = 5$, $\lambda_3 = 1$, $\lambda_4, \lambda_5 = 0.05$, $\gamma_1 = 1$ and $\gamma_2 = 50$, since it yielded the best results.

We use StyleGANv2 in our experiments. For StyleGANv2 latent code manipulation, we use the extended latent space $w \in \mathcal{W}^+$, which predicts a different latent code for every level of a pre-trained generator. Using $\mathcal{W}^+$ allows for a better embedding of an image, but is also possible to cope with images that do not have a latent embedding.

4 Experiments

In this section, we evaluate the qualitative and quantitative performance of our proposed method and compare it to the state-of-the-art. We perform an ablation study to analyze the influence of the loss components in section 4.1. Results on latent space interpolation are discussed in section 4.2. Comparison to state-of-the-art in face swap and reenactment are provided in section 4.3. For all experiments, a cross-dataset evaluation is conducted for our method and baselines.

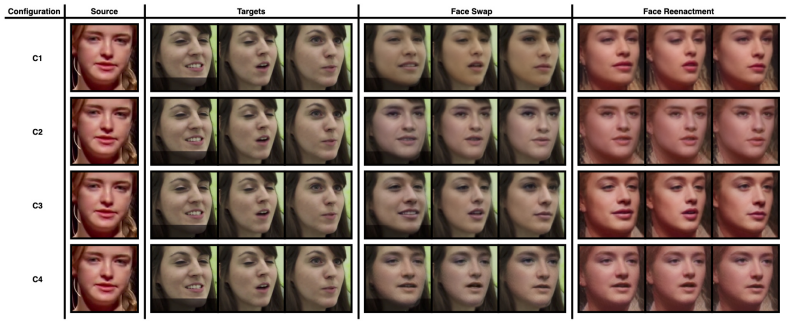

Fig. 3. Ablation Study. Face swap and reenactment results of our method trained with different loss configurations. Our full model results are shown in the last row.

Table 1. Quantitative ablation study evaluation on 300VW dataset. Reported metrics are (a) Inception Score, (b) FID source vs generated, (c) KID source vs generated, (d) FID target vs generated and (e) KID target vs generated.

	Face reenactment				Face swap			
Metric	C1	C2	C3	C4	C1	C2	C3	C4
(a)	2.4 ±0.08	2.02 ±0.08	1.96 ±0.1	**2.69 ±0.16**	**2.68 ±0.13**	2.59 ±0.12	2.63 ±0.12	2.56 ±0.14
(b)	1.57	1.57	1.48	**1.46**	1.54	1.50	1.51	**1.48**
(c)	7.12 ±0.22	7.47 ±0.21	6.84 ±0.2	**5.31 ±0.23**	5.44 ±0.21	5.01 ±0.21	5.27 ±0.2	**4.4 ±0.21**
(d)	N/A	N/A	N/A	N/A	**0.42**	0.49	0.52	0.51
(e)	N/A	N/A	N/A	N/A	1.84 ±0.18	1.78 ±0.16	2.08 ±0.16	**1.45 ±0.15**

4.1 Ablation Study

An ablation study is conducted for the loss components to assess their influence on the face swapping and reenactment tasks on the 300VW dataset [18]. This dataset contains 114 high-quality videos of talking people. The dataset is preprocessed by cropping faces based on the given (ground truth) landmark bounding boxes with 10% extension to each direction.

The qualitative results of our method trained with 4 different loss configurations are shown in Fig. 3: C1 - $\mathcal{L}$ without contextual loss $\mathcal{L}_{id}^{a}$ and identity loss $\mathcal{L}_{id}^{u}$; C2 - $\mathcal{L}$ without $\mathcal{L}_{id}^{u}$; C3 - $\mathcal{L}$ without $\mathcal{L}_{id}^{a}$; C4 - our final model with $\mathcal{L}$. Configurations with other losses being disabled produce significantly degenerated visual results. Consequently, they are crucial for our method.

Contextual loss $\mathcal{L}_{id}^{a}$ supports identity preservation of the source image both in reenactment and face swapping tasks (C2 vs C1). However, it has difficulty with the pose and expression preservation of the target image. Thus, expression and pose are influenced by the content of the source face.

Identity loss $\mathcal{L}_{id}^{u}$ is beneficial for expression/pose isolation and visual sharpness. However, it has difficulties in identity preservation (C3 vs C1). Besides, for the face reenactment task, the reenacted shape of the source person is morphed by target images. It can be seen that the source rounded face becomes oval (C3:

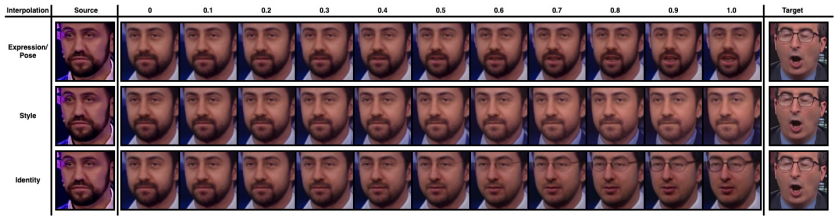

Fig. 4. Interpolation of the latent space. Row 1: expression and pose interpolation. Row 2: Style interpolation. Row 3: identity interpolation. The last column represents a target expression/pose, style, or identity respectively. The results show that our novel disentangled representation learning algorithm can robustly isolates face attributes so that we can manipulate each attribute independently.

Face Reenactment, columns 1, 2). A trade-off result is obtained by combining $\mathcal{L}_{id}^{a}$ and $\mathcal{L}_{id}^{u}$ together (C4 vs C1).

For quantitative evaluation, different metrics are computed which are commonly used in image synthesis evaluation and shown in Table 1. Inception Score [45] uses pre-trained on ImageNet Inception Network to compute the KL divergence between conditional and marginal label distributions over generated data (higher - better). Frechet-Inception distance [46] computes Wasserstein-2 distance between distributions of real and generated samples in the Inception Net feature space (lower - better). Kernel-Inception distance [47] measures dissimilarity between distributions of real and generated samples (lower - better).

Since the generated results of our method are unaligned in term of face attributes, FID and KID metrics are used only as an indicator of how our face identity is similar to the real data distribution. In case of face reenactment, the identity should be as close as possible to source face image. In case of face swap, we want a generated face to capture both properties of source and target image. Consequently, for face swap generated images, the FID and KID metrics are reported both in comparison with the source and target image data distributions. Source and target subjects are randomly selected from the 300VW dataset. Evaluation is performed on a sample of 10K generated images.

In the task of face reenactment, the evaluation metrics support our qualitative experimental results: our method with combined contextual and identity losses generates visual results with identity closer to the source face image distribution (C1 vs C2, 1.57 vs 1.46 FID). In the case of face swap, it can be observed that the distribution of generated images is closer to the distribution of target face images (C4 1.48 vs 0.51 FID). With the introduction of the additional regularization into our model, visual results start to capture more and more properties from the source image (C1 vs C4, 1.54 vs 1.48 FID).

4.2 Latent Space Interpolation

In this section, our method is analyzed to interpolate over different face attribute dimensions. Given a source image, its face attributes are gradually changed where

expression/pose, style or identity are modified to become closer to the target face image. Qualitative results on the 300VW dataset are shown in Fig. 4. The 300VW dataset is preprocessed in the same way as described in the section 4.1. The first column shows the source image. Our algorithm changes gradually an attribute dimension to become closer to the target image of the last column.

Given a source w_1 and target attribute w_2, our model generates meaningful face images conditioned on the interpolated latent code $\alpha w_1 + (1 - \alpha)w_2$. Note that in the case of style, a costume of John Oliver gradually starts to appear, while in the case of identity, we can observe the disappearance of beard, an emergence of his glasses and eyebrows. Despite the challenges given by cross-dataset evaluation, our model preserves attributes dimensions on challenging cases with face accessory and occlusion. Image2StyleGAN [26] show the capability to map face attributes into the latent space of StyleGAN. However, the latent space of expression/pose, identity and style are not fully disentangled. For example, it's not possible to manipulate expression/pose property separately without influencing identity or style. In contrast, our mapping to the latent space provides more flexibility.

4.3 Face Swap and Reenactment State-of-the-Art Comparison

Qualitative Evaluation We evaluate qualitatively our method on the face swapping and face reenactment tasks. We perform cross-dataset evaluation of our method with results produced by FSGAN [8] and First Order Motion Model [6] on the 300VW [18] and UvA-NEMO [19] datasets. These methods are selected because they are state-of-the-art which can do both face swap and reenactment. For our purpose, the available pre-trained model is used provided by authors of FSGAN and First Order Motion. For fairness of comparison, we use models trained on a different dataset from UvA-NEMO and 300VW. The datasets are preprocessed by cropping faces based on landmark bounding boxes with 10% extension to each direction. For 300VW, the provided ground truth landmarks are used. For UvA-NEMO, the landmarks are extracted by using FAN [39].

In the first experiment, we qualitatively compare our method with the state-of-the-art for the face reenactment task. The visual comparison is shown in the Fig. 5. For the First Order Motion model, its pre-trained model is used with *absolute motion* for both face reenactment and face swap experiments, since only the absolute motion mode is capable of computing face swaps. Our method shows comparable quality of reenactment results to First Order Motion and outperforms FSGAN in terms of identity preservation. Besides, since our latent space is constrained by the pre-trained generator, it's less prone to produce artifacts not inherent to a human face (First Order Motion, second row, middle image, eyes). However, this constraint has also a drawback in terms of facial accessories it's capable of modeling (the disappearance of a microphone in the second row). Note that, since First Order Motion is focused on the face reenactment task, it produces better results than the FSGAN model.

In the second experiment, we qualitatively compare our method with state-of-the-art in the context of face swapping. The visual comparison is shown in

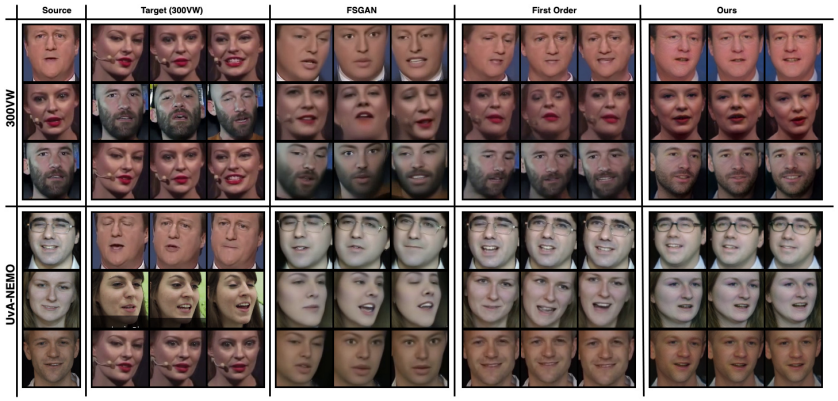

Fig. 5. Qualitative comparison of face reenactment results on 300VW and UvA-NEMO datasets. Pose and expression from target images (second column) are applied on the source image (first column). Faces are produced by the baseline methods, FSGAN and First Order Motion Model, and predictions provided by our novel unified pipeline.

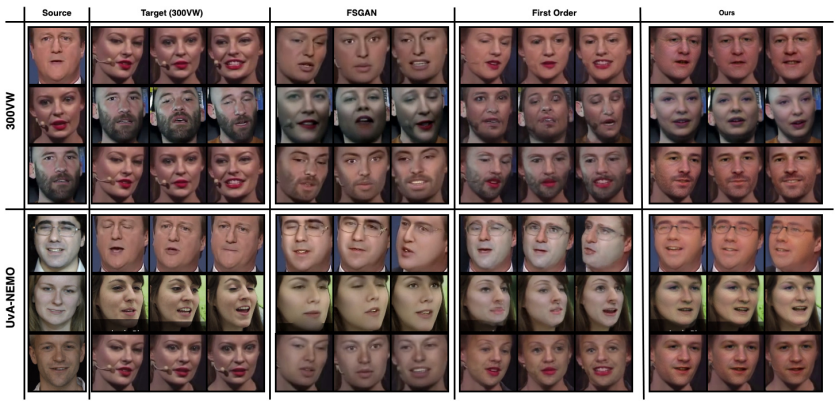

Fig. 6. Qualitative comparison of face swapping results on the 300VW and UvA-NEMO datasets. First column: source image from which identity properties are taken. Second column: target images, on which those properties are applied. Faces produced by the baseline methods, FSGAN [8] and First Order Motion Model [6], and predictions provided by our novel unified pipeline.

the Fig. 6. For the face swapping task, GAN based methods may fail in cross-gender face swapping due to the difference between gender appearance and shape. We show that our method produces realistic-looking results both for male-to-female (rows 1,3,6) and female-to-male swapping (row 2) compared to competitive methods: First Order Motion keeps the lipstick color of a target face (row 3), FSGAN loses the identity of the source image (rows 1,3,6). Note that, since FSGAN is focused on the face swapping task, it produces better results than the First Order Motion model.

Quantitative Evaluation We provide additional quantitative evaluations on 300VW to verify preservation of identity/expression/pose w.r.t. SOTA and to motivate the benefit of joint learning (Table 2). We compare identity preservation using cosine similarity between latent space of VGGFace2 features [48]. Headpose correctness is compared using absolute distance in degrees of yaw-pitch-roll predicted from a pretrained Hopenet [49]. Expression correctness is compared using mean absolute distance of facial landmarks (in pixels, image resized to 256) using a pretrained FAN detector [39]. Our method outperforms SOTA in the swapping task on 3 benchmarks. On the reenactment task Firt Order Motion performs better on identity and headpose preservation however, on average, our method outperforms SOTA.

Table 2. Quantitative evaluation on 300VW.

	Identity ↑			Headpose ↓			Expression ↓		
	First Order	FSGAN	Ours	First Order	FSGAN	Ours	First Order	FSGAN	Ours
Reenactment	**0.578**	0.461	0.517	**2.811**	4.268	3.364	4.883	51.56	**3.983**
Swap	0.308	0.317	**0.412**	2.628	2.823	**2.113**	3.902	2.554	**3.072**
Avg	0.443	0.389	**0.464**	**2.719**	3.546	2.739	4.393	27.057	**3.528**

5 Limitations

Despite promising results presented in this paper, our method has several limitations. First, the expressiveness of the generated facial expressions is dependent on its presence in the training dataset and the quality of face landmarks provided by the landmark detector. Second, our model does not explicitly model occlusion and consequently relies on a pre-trained generator to have a capacity of modeling occlusions, such as accessories or makeup. Finally, both landmark plots and source images contain a bias in terms of identity, pose and expression.

6 Conclusion

In this work, we proposed a novel approach to isolated disentangled representation learning combined with an end-to-end method capable of performing both face reenactment and swapping. To our knowledge, our method is the first approach which is designed to solve both objectives in a unified pipeline.

We showed that our method is trained in an unsupervised way to achieve equally good visual results on both tasks. In addition, it's capable of producing results in a one-shot manner during inference time. The qualitative results on multiple public datasets show that the proposed method is outperforming SOTA methods which can perform both face reenactment and swap.

References

1. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: Learning lip sync from audio. ACM Trans. Graph. **36** (2017)
2. Kuster, C., Popa, T., Bazin, J.C., Gotsman, C., Gross, M.: Gaze correction for home video conferencing. ACM Trans. Graph. **31** (2012)
3. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2387–2395
4. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37** (2018) 163
5. Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., Li, H.: Pagan: Real-time avatars using dynamic textures. ACM Trans. Graph. **37** (2018)
6. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Conference on Neural Information Processing Systems (NeurIPS). (2019)
7. Nirkin, Y., Masi, I., Tran, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: IEEE Conference on Automatic Face and Gesture Recognition. (2018)
8. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7184–7193
9. Garrido, P., Valgaerts, L., Rehmsen, O., Thormaehlen, T., Perez, P., Theobalt, C.: Automatic face reenactment. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '14, USA, IEEE Computer Society (2014) 4217–4224
10. Dale, K., Sunkavalli, K., Johnson, M.K., Vlasic, D., Matusik, W., Pfister, H.: Video face replacement. ACM Trans. Graph. **30** (2011) 1–130
11. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Transactions on Graphics (TOG) **34** (2015)
12. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8789–8797
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4401–4410
14. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 818–833
15. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. CoRR **abs/1611.09577** (2016)
16. Wu, W., Zhang, Y., Li, C., Qian, C., Loy, C.C.: Reenactgan: Learning to reenact faces via boundary transfer. In: ECCV. (2018)
17. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)

18. Chrysos, G., Antonakos, E., Zafeiriou, S., Snape, P.: Offline deformable face tracking in arbitrary videos. In: Proceedings of IEEE International Conference on Computer Vision Workshops, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop, Santiago, Chile, IEEE (2015) 1–9
19. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Are you really smiling at me? spontaneous versus posed enjoyment smiles. In: European Conference on Computer Vision, Springer (2012) 525–538
20. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015)
21. Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., Clune, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. CoRR **abs/1612.00005** (2016)
22. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. CoRR **abs/1912.04958** (2019)
23. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. CoRR **abs/1703.06868** (2017)
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
25. Gabbay, A., Hoshen, Y.: Style generator inversion for image enhancement and animation. CoRR **abs/1906.11880** (2019)
26. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4432–4441
27. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
28. Sanchez, E., Valstar, M.F.: Triple consistency loss for pairing distributions in gan-based face synthesis. CoRR **abs/1811.03492** (2018)
29. Fu, C., Hu, Y., Wu, X., Wang, G., Zhang, Q., He, R.: High fidelity face manipulation with extreme pose and expression. arXiv preprint arXiv:1903.12003 (2019)
30. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR. (2020)
31. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: Automatically replacing faces in photographs. ACM Trans. Graph. **27** (2008) 1–8
32. Kemelmacher-Shlizerman, I.: Transfiguring portraits. ACM Trans. Graph. **35** (2016)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232
34. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 9459–9468
35. Yang, C., Lim, S.N.: Unconstrained facial expression transfer using style-based generator. arXiv preprint arXiv:1912.06253 (2019)
36. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
37. Banerjee, S., Scheirer, W.J., Bowyer, K.W., Flynn, P.J.: On hallucinating context and background pixels from a face mask using multi-scale gans. CoRR **abs/1811.07104** (2018)

38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. (2018)
39. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision. (2017)
40. Hu, Y., Wu, X., Yu, B., He, R., Sun, Z.: Pose-guided photorealistic face rotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8398–8406
41. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13** (2018) 2884–2896
42. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 768–783
43. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (2015)
44. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10** (2009) 1755–1758
45. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. CoRR **abs/1606.03498** (2016)
46. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR **abs/1706.08500** (2017)
47. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans (2018)
48. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition. (2018)
49. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2018)