

# Face Identity Disentanglement via Latent Space Mapping

YOTAM NITZAN, Tel-Aviv University

AMIT BERMANO, Tel-Aviv University

YANGYAN LI, Alibaba Cloud Intelligence Business Group

DANIEL COHEN-OR, Tel-Aviv University

Learning disentangled representations of data is a fundamental problem in artificial intelligence. Specifically, disentangled latent representations allow generative models to control and compose the disentangled factors in the synthesis process. Current methods, however, require extensive supervision and training, or instead, noticeably compromise quality.

In this paper, we present a method that learns how to represent data in a disentangled way, with minimal supervision, manifested solely using available pre-trained networks. Our key insight is to decouple the processes of disentanglement and synthesis, by employing a leading pre-trained unconditional image generator, such as StyleGAN. By learning to map into its latent space, we leverage both its state-of-the-art quality, and its rich and expressive latent space, without the burden of training it.

We demonstrate our approach on the complex and high dimensional domain of human heads. We evaluate our method qualitatively and quantitatively, and exhibit its success with de-identification operations and with temporal identity coherency in image sequences. Through extensive experimentation, we show that our method successfully disentangles identity from other facial attributes, surpassing existing methods, even though they require more training and supervision.

**CCS Concepts:** • Computing methodologies → Learning latent representations; Computer graphics; Neural networks.

**Additional Key Words and Phrases:** Disentanglement, Deep Learning, Generative Adversarial Networks

**ACM Reference Format:**

Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. 2020. Face Identity Disentanglement via Latent Space Mapping. *ACM Trans. Graph.* 39, 6, Article 225 (December 2020), 23 pages. <https://doi.org/10.1145/3414685.3417826>

## 1 INTRODUCTION

Since the dawn of machine learning, learning a disentangled representation has been one of its core problems. Disentanglement can be defined as the ability to control a single factor, or feature, without affecting other ones [Locatello et al. 2018]. A properly disentangled representation can benefit semantic data mixing [Johnson et al. 2016; Xiao et al. 2019], transfer learning for downstream tasks [Bengio et al. 2013; Tschannen et al. 2018], or even interpretability [Mathieu

---

Authors' addresses: Yotam Nitzan, Tel-Aviv University, [yotamnitzan@gmail.com](mailto:yotamnitzan@gmail.com); Amit Bermano, Tel-Aviv University; Yangyan Li, Alibaba Cloud Intelligence Business Group; Daniel Cohen-Or, Tel-Aviv University.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0730-0301/2020/12-ART225 \$15.00  
<https://doi.org/10.1145/3414685.3417826>

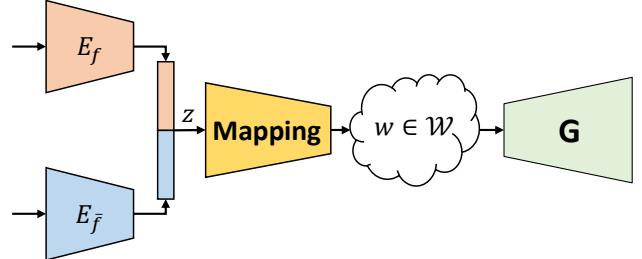


Fig. 1. Our disentanglement framework uses two encoders (left) to generate the latent code  $z$ , consisting of a description of the property of interest ( $f$ ), and all the rest ( $\bar{f}$ ). The code is then mapped to the latent space  $\mathcal{W}$  of the employed pre-trained generator  $G$ . This decouples the tasks of learning quality image generation and disentanglement.

et al. 2018]. Achieving disentanglement, however, is a notoriously difficult task, which has been addressed by many approaches.

A key challenge to learning a disentangled representation is reducing supervision. Similarly to many other learning-related objectives, fully supervised solutions are most effective [Aberman et al. 2019; Reed et al. 2015], but impose often infeasible data collection requirements. It is not tractable, for example, to find a data set of paintings depicting the same scenes in different styles. The opposite approach, of a completely unsupervised disentanglement, is equally impractical at the moment, as it typically struggles with producing satisfying results [Locatello et al. 2018]. Therefore, middle ground forms of supervision have been proposed. A prominent example is *class-supervision*, where only the feature of interest is labeled throughout the dataset, partitioning it into classes. The class-supervised setting assumes the existence of multiple samples in each class, and that the intra-class variation of this feature is significantly lower than the inter-class ones [Gabbay and Hoshen 2019]. While being more feasible, this approach still requires meticulous gathering and labeling of data. Avoiding the labeling requirement would enable using virtually endless amounts of data.

In this paper, we present a novel method to disentangle face identity from all other facial attributes, using no data specific supervision. In our case, supervision is solely realized through the use of relevant pre-trained networks — a plausible prerequisite, as shall soon be demonstrated, especially since these networks need not be trained on the same dataset or for the same task. Our key idea is to directly map the disentangled latent representation to the latent space of a pre-trained generator, as depicted in Figure 1. Given a feature of interest  $f$ , identity in our case, we propose training two encoders, as is commonly done in disentanglement setting [Bao et al. 2018; Gabbay and Hoshen 2019],  $E_f$  and  $E_{\bar{f}}$ , seeking to encode only

$f$ , and everything but  $f$ , respectively. Unlike traditional methods, however, we then propose to map the resulting latent code  $z$  to the latent space  $\mathcal{W}$  of a powerful, pre-trained generator  $G$ , and assess the quality of the disentanglement only on the latter’s output. This mapping is the heart of our approach. It allows us to use a state-of-the-art pre-trained generator, inheriting its high-quality and fidelity, and to control its output in a disentangled manner with minimal training. Furthermore, our approach relaxes the requirement for a distinct disentanglement, where the representation is split into two parts which are mutually exclusive, and carry completely separate information. In our approach, the mapping is trained to disentangle and extract the relevant information from each part to be combined into a complete representation of the target image. In practice, this approach decouples the disentanglement task from the synthesis one, allowing the native employment of the most expressive and high-quality image generation techniques, and a dedicated training process for content control without compromising generation quality. We demonstrate our approach using arguably the most powerful unconditional image generator available nowadays — StyleGAN [Karras et al. 2019b], in one of the most challenging image generation domains — the human face and head. Generating and manipulating faces is highly applicable on one hand, but is also known to be particularly hard, on the other. Besides the challenges of dealing with human faces that arise from the keen human perception of them [Mori et al. 1970], and their high photometric, geometric and kinematic complexities, the human face has many independent, high dimensional attributes. From these, we choose to demonstrate image synthesis with disentangled control over the person identity attribute, as illustrated in Figure 2. This type of control is useful in applications such as de-identification, reenactment, and many others. Unlike other methods [Bouchacourt et al. 2018; Denton et al. 2017; Gabay and Hoshen 2019; Hadad et al. 2018], our training data does not contain examples of the same person twice, nor does it have any indication or labeling regarding the person’s identity. Through the use of available networks for evaluating identity and facial landmarks, our approach effectively transforms StyleGAN into a conditional image generator, conditioned on either the identity of the person or all other facial attributes, such as expression, pose and illumination.

As we shall see, the performance of our disentangled image generation heavily depends on the capabilities of the selected generator. In the case of StyleGAN, this means phenomenal image quality, outperforming all previous disentangled control attempts we compare to, but at the cost of expressiveness, as some of the faces do not reside within the attainable domain of the generator, due to the data used during its training. Nevertheless, in addition to superior quality, our method also successfully handles the generation of the entire head, including the hair — a region that is known to have a strong impact on identity [Abudarham et al. 2019]. This is in contrast to state-of-the-art identity manipulation methods, which manipulate facial features only [Bao et al. 2018; Gafni et al. 2019; Li et al. 2019; Nirkin et al. 2019].

As validation, we offer several experiments, evaluating qualitatively and quantitatively face identity and attributes manipulations, and compare them to previous methods. The experiments assess

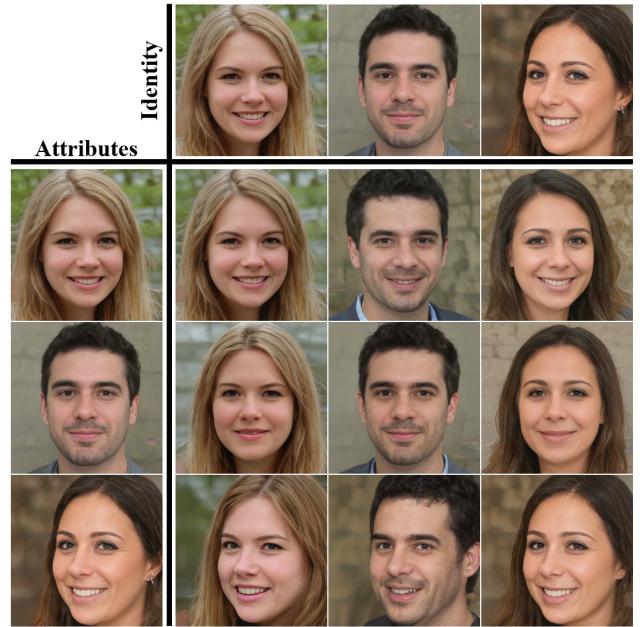


Fig. 2. Sample results generated by our method, demonstrating the ability to disentangle identity from other facial attributes: pose, expression and illumination and preserving one while manipulating the other. Three images are used as input, forming a 3 by 3 table combinations generated by our method. As can be seen, identity is preserved along the columns, and attributes are preserved along the rows.

feature combination, de-identification operations, and temporal coherency of identity over sequences. The methods we compare to include state-of-the-art class-supervised disentanglement methods, which rely on a more structured, better labeled, but harder to curate, data. We evaluate all methods in terms of the quality of disentanglement and preservation of composing factors on unseen faces, as well as image quality and diversity. Our method is shown to outperform previous art, in addition to offering unique advantages, such as the said generation of the entire head and hair, and minimal supervision, which does not necessitate multiple images of the same person at any point.

## 2 RELATED WORK

### 2.1 Disentanglement

Many works learn disentangled representations and use various levels of supervision. Fully supervised methods [Aberman et al. 2019; Reed et al. 2015] learn from a dataset in which all relevant underlying factors are labeled. In this case, a sample from the dataset takes the form of  $(\text{input}, \text{transformation}, \text{result of transformation})$ . Critically, a ground truth is available for any transformation performed on the data. This data requirement is infeasible in many domains and tasks. On the other end of the spectrum, fully unsupervised methods [Chen et al. 2016; Higgins et al. 2017; Kim and Mnih 2018] learn from a dataset with no associated information. These methods employ information-theoretic regularization losses to encourage

disentanglement. These methods trade-off quality for disentanglement, often explicitly [Higgins et al. 2017; Kim and Mnih 2018], and thus produce low visual quality results. Class-supervised methods [Bao et al. 2018; Bouchacourt et al. 2018; Denton et al. 2017; Gabbay and Hoshen 2019; Hadad et al. 2018; Li et al. 2019; Xiao et al. 2019] use additional labels that partition the images to classes, defined by a set of mutually exclusive attribute values. For example, each class contains a set of images of a single identity only, while other attributes vary. In contrast, we examine a more challenging but easily attainable dataset, in which no person appears twice, and no labeling is offered.

All aforementioned fully-supervised and class-supervised methods follow a similar approach, where separate encoders infer latent representations of their inputs. Next, those representations are fed into a generator that produces the output image. As described in Section 3, our choice of architecture follows the same approach. Among these works, perhaps the most similar to ours are the ones presented by Bao et al. [2018] and its successor, FaceShifter [Li et al. 2019]. Both focus specifically on disentangling identity from other attributes and use a pre-trained face recognition network as an encoder to infer the identity representation. There are several notable differences between these methods and ours. Both methods train a generator themselves, while we use a pre-trained generator. Additionally, both methods deal with inner-facial features only, not including the head and hair. Most importantly, both are class-supervised, explicitly requiring identity labels. We qualitatively compare our method with FaceShifter in Section 4.1.

## 2.2 Latent Space of GANs

With the rapid evolution of GANs, many works have tried to understand and control their latent space. Several methods apply *GAN Inversion*, where the latent vector that most accurately reconstructs a given image is sought. Some methods train an encoder to map images to the latent space in conjunction to training the generator (e.g., [Huang et al. 2018; Pidhorskyi et al. 2020]). Training the generator together with the encoder requires long training and often compromises the generated images' quality. Therefore, other methods consider inversion as a post-hoc task where the generator is pre-trained and then inversion is solved separately. Such methods either directly optimize the latent vector to minimize reconstruction error for every image [Abdal et al. 2019; Creswell and Bharath 2018; Lipton and Tripathi 2017], train an encoder to map images to the latent space [Luo et al. 2017; Perarnau et al. 2016; Richardson et al. 2020], or use a hybrid approach combining both [Baylies 2019; Zhu et al. 2020, 2016]. A separate line of research deals with learning to traverse the latent space in a semantically meaningful manner. A popular approach is to find linear directions that correspond to changes in a given binary labeled attribute, such as young  $\leftrightarrow$  old, or no-smile  $\leftrightarrow$  smile [Denton et al. 2019; Goetschalckx et al. 2019; Shen et al. 2019]. Tewari et al. [2020] utilize a pre-trained 3DMM to learn semantic face edits in latent space. Jahanian et al. [2019] find latent space paths that correspond to a specific image transformation, such as zoom or rotation, in a self-supervised manner. Härkönen et al. [2020] find useful paths in a completely unsupervised manner. These paths are set to be the principal component axes (PCA) on

an intermediate activation space. Similar to the other methods, the computed transitions control one-dimensional attributes such as age or gender, as well as image transformations like zoom or rotation.

Our work is different from the two lines of work discussed above. First, inversion methods start from a single image and aim to infer a latent vector that reconstructs it. This means that the input image serves as a particularly strong, pixel-level, supervision for the expected output. In contrast, our method infers a latent vector that represents an unseen image, and therefore cannot rely on pixel-level supervision, but on a significantly weaker form of supervision. The second difference concerns image manipulation. We perform manipulation by directly mapping properties from two given images to a latent vector that represents the unseen output comprising both properties. This is in contrast to previous methods that propose traversing the latent space to perform image editing.

It may be possible to devise a way to apply inversion followed by a traversal of the latent space to perform a task similar to the one we describe, i.e., to produce a novel image that keeps properties from one image, and alters some other ones. However, all properties we generate seek to match those of the two input images, i.e., identity from one image and facial attributes from the other. On the other hand, state-of-the-art methods based on image inversion cannot achieve the same goal. For example, they cannot generate an image featuring the identity of the input image, but also the smile of another specific one. Such an approach would require traversing the latent space, which renders mimicking a specific expression extremely challenging. Furthermore, latent traversal methods depend more on the local behavior of the latent space, assuming it is well-behaved everywhere. Our method, on the other hand, bypasses the need to explicitly invert an image but rather produces the output image through a direct mapping.

## 2.3 Controlling Facial Attributes

There is an abundance of works on face manipulation with various means of controlling the facial attributes. These can be categorized by whether they preserve identity and manipulate other attributes or vice versa.

As for the former, some works are based on conditional GANs that translates between different domains such as young  $\rightarrow$  old, or happy  $\rightarrow$  angry [Choi et al. 2018, 2019; Liu et al. 2017; Perarnau et al. 2016; Pumarola et al. 2018]. These methods are limited to a discrete number of domains and require datasets with their associated labels, but work on unseen image. By comparison, face reenactment methods transfer expression and head pose (sometimes also illumination and eye gaze) from a source video to a target identity [Averbuch-Elor et al. 2017; Kim et al. 2018; Thies et al. 2016, 2018; Zakharov et al. 2019]. But, there methods usually require training a network for each given identity, and assumes the availability of videos of it. Another line of identity-preserving works [Liu et al. 2018; Shen et al. 2018; Tian et al. 2018] train GAN-based Image-to-Image translation networks that preserve the identity of the input image, while a subset of other attributes are controlled by a different image. Unlike our method, these methods are all class-based, i.e. relying on having an identity-labeled image dataset with multiple images for each identity.

Face de-identification methods are of the latter category, editing the identity of a target image, while preserving other attributes, like expression, pose, illumination, etc. Sun et al. [2018] remove the face from the target image and complete it using a GAN, while other methods [Gafni et al. 2019; Wu et al. 2018] shift the identity using a pre-trained face recognition network, while keeping the general image relatively similar to the source. A popular approach is performing face swap [Bao et al. 2018; deepfakes 2019; Li et al. 2019; Nirkin et al. 2019], in which the face in the target image is replaced by the one from the source image. Unlike previously mentioned methods, face swap allows controlling the generated identity which is copied from the source image. However, face swapping methods are generally limited to using a target face relatively similar to the source face.

Our method, allows the editing of both identity and facial attributes, while also controlling the generated identity according to an input image. In contrast to other methods, our method is successful even when the input images are completely different, demonstrating our disentanglement capability. Furthermore, all aforementioned methods only alter the internal features of the face, usually by cropping a tight face region or by using a segmentation mask of the face, leaving the head and hair intact. This approach conflicts with the fact that numerous works have identified that the appearance of the head as a whole, and specifically the hair, are crucial for identification [Abudarham et al. 2019; Sendik et al. 2019; Sinha and Poggio 1996; Toseeb et al. 2012]. Our method, on the other hand, generates an entire human head, including the hair, thus better controlling and preserving the generated identity.

### 3 METHOD

Our method takes two images as input:  $I_{id}$ ,  $I_{attr}$ . The goal is to generate an image with the identity from  $I_{id}$  and all other attributes, specifically pose, expression and illumination from  $I_{attr}$ . The disentanglement task is therefore to disentangle identity from all other attributes. Once achieved, We can extract the identity from  $I_{id}$  and the attributes from  $I_{attr}$ , and reassemble them into a combined representation of a new human head. This representation is then fed to the generator, to generate an image that respects both the identity of  $I_{id}$  and attributes of  $I_{attr}$ . Encouraging this disentanglement, while generating state-of-the-art quality images, is notoriously difficult. The key idea of our work is to separate the two objectives, by learning to map the combined representation into the latent space of a pre-trained generator, and its existing semantics. For this reason, we use a pre-trained generator with semantics-rich, and expressive latent space.

As depicted in Figure 3, the network consists of two encoders  $E_{id}$  and  $E_{attr}$ , a mapping network  $M$ , and a generator network  $G$ . An additional encoder,  $E_{lnd}$  and discriminator  $D_{\mathcal{W}}$  are used for loss calculation only and shall be discussed in Section 3.2. The task of generating an image with the same identity as in  $I_{id}$ , and the attributes portrayed in  $I_{attr}$ , consists of two parts: extracting identity and attributes from the corresponding input images, and then reassembling them to create a new head representation and generating an image accordingly. For the former, we use the two encoders

$E_{id}$ ,  $E_{attr}$ . For the latter, we combine the codes by concatenation:

$$z = [E_{id}(I_{id}), E_{attr}(I_{attr})] \quad (1)$$

and map  $z$ , using  $M$ , into the latent space of a pre-trained state-of-the-art generator  $G$ , which then generates the output image,  $I_{out}$ . We use state-of-the-art StyleGAN [Karras et al. 2019b] as the pre-trained generator for all our experiments. Differently from other GANs, StyleGAN has two latent spaces:  $\mathcal{Z}$ , which is induced by a fixed distribution, and  $\mathcal{W}$  induced by a learned mapping from  $\mathcal{Z}$ . We choose to map the combined face code into  $\mathcal{W}$ , as it is a more disentangled latent space than  $\mathcal{Z}$ , thus more suitable to facilitate and accommodate image editing [Karras et al. 2019b; Shen et al. 2019]. The design choice of using an existing latent space is crucial for a few reasons. Usually disentanglement is performed with the premise that given enough samples with constant factors, while other factors are varying, one could learn to identify the constant factors and disentangle it from the others e.g., LORD [Gabbay and Hoshen 2019]. In our setting, the dataset does not contain more than one image of the same person. Therefore, it is unclear how the network could learn to disentangle. Our approach resolves this problem by leveraging a latent space that already exhibits some degree of disentanglement, achieved in a completely unsupervised manner. Moreover, by using a state-of-the-art generator, we alleviate the difficulty of learning to generate high-quality and high-fidelity images. However, training the mapping between the latent space of the encoder and  $\mathcal{W}$ , is not trivial. Thus, we add a discriminator  $D_{\mathcal{W}}$  to help  $M$  predict features that lie within  $\mathcal{W}$ .  $D_{\mathcal{W}}$  is trained in an adversarial manner to discriminate between real samples from StyleGAN’s  $\mathcal{W}$  space and  $M$ ’s predictions. Note that, thanks to our use of a pre-trained generator, there is no discriminator employed on  $I_{out}$ . Thus, side-stepping much of the difficulty of training adversarial methods.

#### 3.1 Network Architecture

The  $E_{id}$  encoder is a pre-trained ResNet-50 [He et al. 2016] face recognition model, trained on VGGFace2 [Cao et al. 2018], with loose crops including the hair. The  $E_{attr}$  encoder is implemented as InceptionV3 [Szegedy et al. 2016]. For both encoders, their output is taken from the last feature vector before the FC classifier. The mapping network,  $M$ , is a 4-layers MLP with LReLU [He et al. 2015] activation layers. The generator,  $G$  is a pre-trained StyleGAN synthesis network, trained on FFHQ [Karras et al. 2019b].  $G$  takes our predicted  $w$  vector as input and employs it normally through the AdaIN [Huang and Belongie 2017] layers. Both  $E_{id}$  and  $G$  are kept frozen during training, while all other networks are trainable.

#### 3.2 Training and Losses

We create a dataset using StyleGAN in the following manner. We sample 70,000 random Gaussian vectors and forward them through a pre-trained StyleGAN. In the forward process, the Gaussian noise is mapped into a latent vector  $w$ , from which an image is generated, and we record both the image and the  $w$  vector. The StyleGAN generated images are used as our training dataset, and the latent  $w$  vectors are used as "real" samples for training  $D_{\mathcal{W}}$ .

StyleGAN cannot create the entire human head space, specifically all human identities, from its latent space  $\mathcal{W}$ . Some works [Abdal

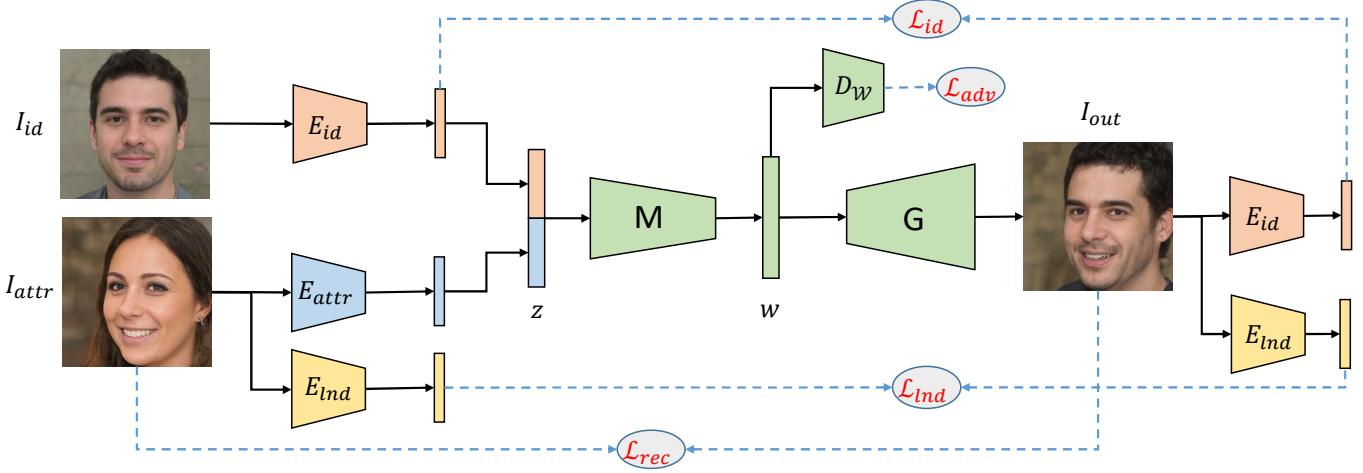


Fig. 3. Our disentanglement scheme, as utilized in the human head domain. Data flow is marked by solid lines, and losses by dashed ones. The identity and attributes codes are first extracted from two input images using encoders  $E_{id}$  and  $E_{attr}$  respectively. Through our mapping network  $M$ , the concatenated codes are mapped to  $\mathcal{W}$ , the latent space of the pre-trained generator  $G$ , which in turn generates the resulting image. An adversarial loss  $\mathcal{L}_{adv}$  ensures proper mapping to the  $\mathcal{W}$  space. Identity preservation is encouraged using  $\mathcal{L}_{id}$ , that penalizes differences in identity between  $I_{id}, I_{out}$ . Attributes preservation is encouraged using  $\mathcal{L}_{rec}$ ,  $\mathcal{L}_{lnd}$ , that penalizes pixel-level and facial landmarks differences respectively, between  $I_{attr}, I_{out}$ .

et al. 2019; Zhu et al. 2020] used an artificially enlarged latent space, named  $\mathcal{W}+$ , from which the generator may also create non-human images, including cats and bedrooms. We therefore choose to use a StyleGAN generated dataset to prevent the conflict between identity preserving and mapping into StyleGAN's rich latent space  $\mathcal{W}$ .

For adversarial loss, we use the non-saturating loss [Goodfellow et al. 2014] with  $R_1$  regularization [Mescheder et al. 2018]:

$$\begin{aligned} \mathcal{L}_{adv}^D = & -\mathbb{E}_{w \sim \mathcal{W}} [\log D_{\mathcal{W}}(w)] - \mathbb{E}_z [\log(1 - D_{\mathcal{W}}(M(z)))] + \\ & \frac{\gamma}{2} \mathbb{E}_{w \sim \mathcal{W}} [\|\nabla_w D_{\mathcal{W}}(w)\|_2^2] \end{aligned} \quad (2)$$

$$\mathcal{L}_{adv}^G = -\mathbb{E}_z [\log D_{\mathcal{W}}(M(z))] \quad (3)$$

An  $L_1$  cycle consistency loss between  $I_{id}$  and  $I_{out}$  is used to enforce identity preservation:

$$\mathcal{L}_{id} = \|E_{id}(I_{id}) - E_{id}(I_{out})\|_1 \quad (4)$$

As discussed, human perception is highly sensitive to minor artifacts in facial appearance, this is especially true when generating sequences of frames, where not only does every individual frame must look realistic, but the motion across frames must also be realistic. Facial landmarks model the possible motion of the human face, Therefore, we incorporate a sparse  $L_2$  cycle consistency landmarks loss. Landmarks are extracted using a pre-trained network noted as  $E_{lnd}$ :

$$\mathcal{L}_{lnd} = \|E_{lnd}(I_{attr}) - E_{lnd}(I_{out})\|_2 \quad (5)$$

Additional loss is exercised to encourage pixel-level reconstruction of  $I_{attr}$ . This loss is clearly motivated by our desire for  $I_{out}$  to be generally similar to  $I_{attr}$ . Intuitively, if  $I_{id}, I_{attr}$  are the same image, we would expect our method to reconstruct this image. Furthermore, we would like to capture and preserve pixel-level information such as colors and illumination, not modeled by any other loss. For this

end, we adopt the "mix" loss suggested in Zhao et al. [2016], and use a weighted sum of  $L_1$  loss and MS-SSIM loss:

$$\mathcal{L}_{mix} = \alpha(1 - \text{MS-SSIM}(I_{attr}, I_{out})) + (1 - \alpha) \|I_{attr} - I_{out}\|_1 \quad (6)$$

However, a pixel reconstruction loss might also affect the identity of  $I_{out}$  by reconstructing facial features from  $I_{attr}$ . In order to prevent this, we employ the reconstruction loss only when  $I_{id} = I_{attr}$ , i.e. :

$$\mathcal{L}_{rec} = \begin{cases} \mathcal{L}_{mix}, & I_{id} = I_{attr} \\ 0, & \text{Otherwise} \end{cases} \quad (7)$$

The overall generator non-adversarial loss is a weighted sum of the above losses:

$$\mathcal{L}_{non-adv}^G = \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{lnd} + \lambda_3 \mathcal{L}_{rec} \quad (8)$$

Our training procedure is simple. We uniformly randomly sample images and latent vectors from our generated dataset. The images are used for  $I_{id}, I_{attr}$  and the latent vectors are used as "real" samples for  $D_{\mathcal{W}}$ . When  $I_{id} \neq I_{attr}$ , the network learns to disentangle identity from attributes. Whereas when  $I_{id} = I_{attr}$  it learns to encode all the information needed for proper reconstruction.

### 3.3 Implementation Details

We use StyleGAN pre-trained at 256x256 resolution in all our experiments, to easily compare with other methods. Please see the supplementary material for 1024x1024 results as well. The ratio of training samples with  $I_{id} \neq I_{attr}$  and  $I_{id} = I_{attr}$  is an hyper-parameter that controls the weight for disentanglement and reconstruction. We empirically take  $I_{id} \neq I_{attr}$  every third iteration, and  $I_{id} = I_{attr}$  otherwise.  $E_{lnd}$  is implemented using a pre-trained landmarks regression network [Feng et al. 2018], trained to regress 68 facial keypoints.

We optimize the adversarial loss  $\mathcal{L}_{adv}^G$  and non-adversarial losses  $\mathcal{L}_{non-adv}^G$  separately, which proved to be more stable during training. Training is performed using the Adam [Kingma and Ba 2015] optimizer, with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . We use learning rate of  $5e^{-5}$  when optimizing  $\mathcal{L}_{non-adv}^G$ . For adversarial learning rates we follow Heusel et al. [2017] and use  $5e^{-6}$  for G’s adversarial loss  $\mathcal{L}_{adv}^G$  and  $2e^{-5}$  for  $D_{\mathcal{W}}$ . Loss weights are set to  $\lambda_1 = 1$ ,  $\lambda_2 = 1$ ,  $\lambda_3 = 0.001$ ,  $\lambda_4 = 0.02$ ,  $\alpha = 0.84$  and  $\gamma = 10$ . When calculating  $\mathcal{L}_{Ind}$  we use only the 52 inner-face landmarks, removing the jawline landmarks which were found to strongly affect the head shape, harming identity preservation. The network is trained end-to-end with batch size 6 on a single NVIDIA Titan XP GPU and requires roughly a day to converge. Note that this is incredibly efficient, as training StyleGAN itself would require more than 30 days on the same GPU.

## 4 EXPERIMENTS

We perform extensive experimentation to evaluate our method, mainly through two aspects: the quality of the disentanglement, or how well we control identity and facial attributes without them affecting each other, and the quality of the synthesized images. We further perform ablation studies, which show the importance of individual components in our pipeline, and compare our performance to state-of-the-art methods.

First, a qualitative inspection of the results are shown in Figures 4 and 5 for StyleGAN and FFHQ input images, respectively. The array of images illustrates the degree of identity preservation of our method (along the columns) and preservation of the rest of the attributes (along the rows). In addition, our method successfully preserves the overall head shape and the hair – a pivotal yet elusive part of true identity preservation. Additionally, we observe consistency in details such as the existence and appearance of glasses. This is a crucial element when considering consistency, and is especially relevant when, for example, generating consecutive frames for a sequence. Note that these disentanglement and preservation capabilities cannot be achieved by the style mixing approach proposed in the original StyleGAN paper, where styles entangle identity and other semantic attributes. A performance gap can be observed between our results on “real” images (FFHQ) and on StyleGAN generated images. As previously discussed, we inherit the performance of the employed pre-trained generator and its latent space. As discussed in the literature [Abdal et al. 2019; Karras et al. 2019a; Zhu et al. 2020], StyleGAN is unable to generate the entire space of human heads from  $\mathcal{W}$ . This is most evident for the head pose, where faces generated by StyleGAN feature no roll angle, since the faces in the training data were vertically aligned. Similarly, not all human identities can be generated by StyleGAN. By inheriting StyleGAN’s performance, our method generates the closest possible identity, which qualitatively and quantitatively is very similar as evident in Table 1.

### 4.1 Comparison with Previous Methods

We qualitatively compare our results against those of LORD [Gabbay and Hoshen 2019] on images from CelebA [Liu et al. 2015] in Figure 6. Note that the comparison is performed on images from the dataset on which LORD was trained. On the other hand, our method was

Table 1. Quantitative Comparison of our method with LORD [Gabbay and Hoshen 2019] and FSGAN [Nirkin et al. 2019]

Method	FID ↓	Identity ↑	Expression ↓	Pose ↓
LORD	23.08	$0.20 \pm 0.11$	$0.085 \pm 0.018$	$13.34 \pm 15.00$
FSGAN	8.90	$0.35 \pm 0.19$	<b><math>0.013 \pm 0.013</math></b>	<b><math>6.73 \pm 13.62</math></b>
Ours	<b>4.28</b>	<b><math>0.60 \pm 0.09</math></b>	$0.017 \pm 0.019$	$9.74 \pm 13.16$

trained on a significantly different dataset composed from StyleGAN generated images. Therefore, making this comparison significantly more challenging for our method. LORD uses low resolution, 64x64 tight face crops, while our method handles higher resolution, of 256x256 loose crops. For a fair comparison, each method is run using its native input configuration, and in post-process we crop and resize the output images to make them visually comparable. The red frames indicate the region of the image that is input to LORD.

As can be observed, our method better preserves the facial expression of the attributes image, regardless of the expression of the identity image, indicating a strong disentanglement between identity and expression. This is most noticeable when observing the mouth, where our method is able to generate various mouth shapes. On the other hand, LORD struggles in preserving expression when the identity image has a non-neutral expression, and preserving the challenging open-mouth expression. Furthermore, our results have a much higher visual quality than LORD’s, which are of low resolution and contain artifacts, most noticeable is the checkerboard effect, which should not be attributed to the resizing of the image as it also exists in the original resolution.

We also compare our results with the latest face swapping methods FaceShifter [Li et al. 2019] and FSGAN [Nirkin et al. 2019]. As discussed earlier, face swapping is a different yet related task that focuses on replacing inner face features only. We qualitatively demonstrate the differences in our application from face swapping in Figure 7. Our method preserves the inner face features from the identity image, with similar quality to previous methods. However, we also preserve the head shape and hair from the identity image, regardless of that from the attributes image, overall preserving the identity better. Furthermore, we produce results of the same quality even when input images are completely different, as opposed to face swapping methods that are limited to operate on relatively similar images. When input images are different, noticeable artifacts like phantom hair (rows 1,2) and two jaw lines (row 2) may appear. Even when no artifacts are created, the output may be not recognized as either of the input identities and often create an unrealistic face simply because it depicts a very unusual appearance (rows 3, 4).

We further quantify the aforementioned differences in performance by conducting a quantitative comparison to LORD and FSGAN. We evaluate the methods’ ability to disentangle and preserve underlying factors composing the human head from different sources, as well as evaluating image quality quantitatively. The evaluation is performed by randomly sampling 10K pairs of images from FFHQ, that are used as identity and attribute inputs. We then apply all methods to infer output images. Next, we assess the preservation of identity, pose and expression from the respective source image,

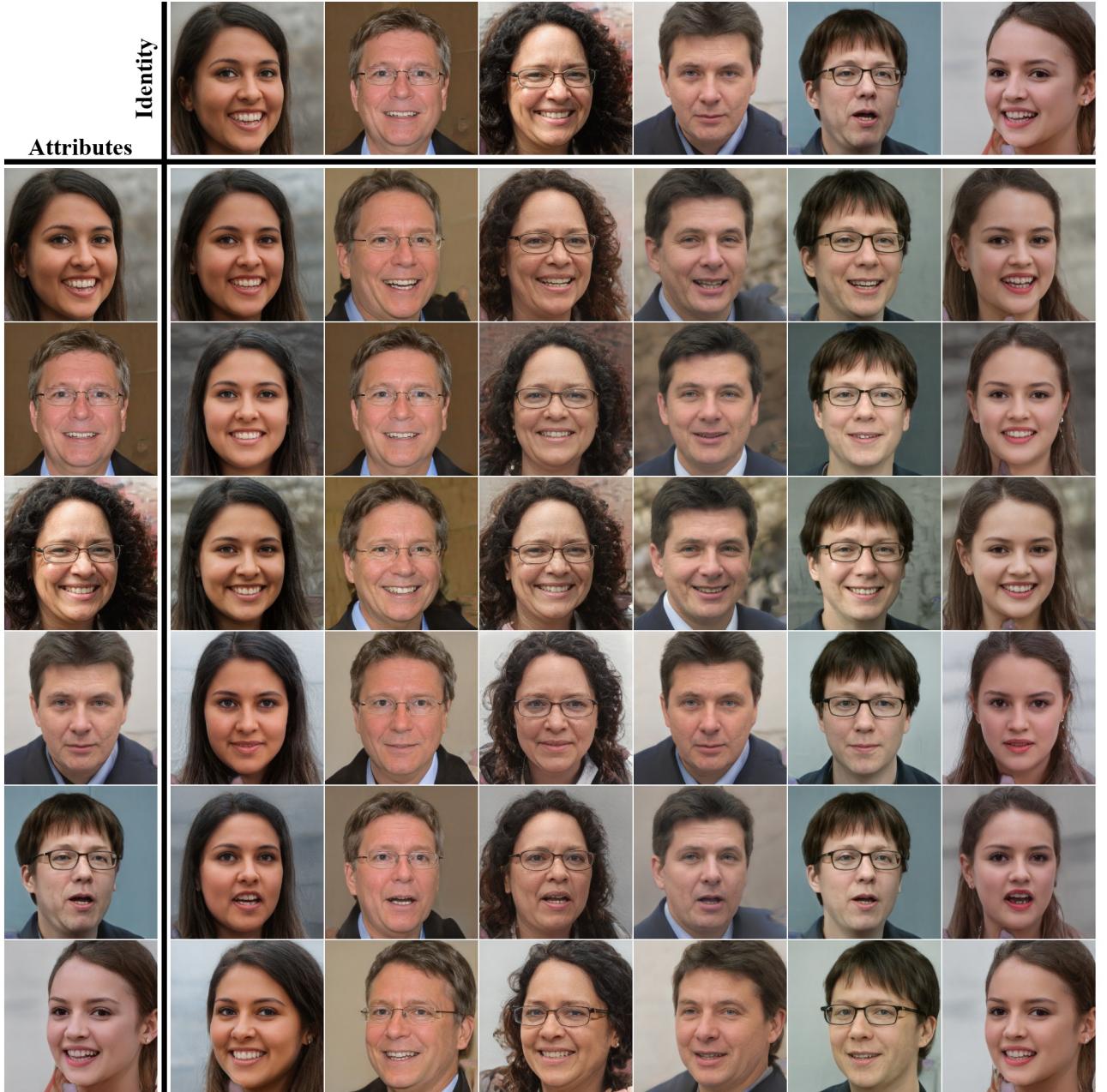


Fig. 4. Feature combination results. For every image in the table, identity is taken from the top, and the rest of the attributes (including expressions, orientation, lighting conditions, etc.) from the left. All images (both inputs and output) were generated using StyleGAN.

and the quality of the output image. The results are displayed in Table 1.

To test identity preservation, we employ state-of-the-art face recognition network, namely ArcFace [Deng et al. 2019], and adopt the cosine similarity metric to compare the identity of  $I_{id}$  and  $I_{out}$ . Note that, ArcFace is completely different than the face recognition network used during training, in terms of both the training set and

losses. The accuracy of expression preservation is calculated by Euclidean distance between 2D landmarks of  $I_{attr}$  and  $I_{out}$ , inferred using dlib [King 2009]. We normalize the landmarks values by dividing them by the output image resolution, to make the methods comparable. Similarly, pose preservation is measured by the Euclidean distance between Euler angles of  $I_{attr}$  and  $I_{out}$ . For each of the above, we calculate the mean and standard deviation across the



Fig. 5. Feature combination results on FFHQ images. The setting is identical to Fig. 4.

test set. Last, we evaluate the quality of the image, specifically on the face region, which is the interest of this work. We sample new 10K images to serve as real images. We then detect [Zhang et al. 2016] and crop faces from the output images and the real images, creating the test sets. Afterward, we calculate the FID [Heusel et al. 2017] score for all methods. As can be seen in Table 1, our approach is superior to both methods in terms of image quality and identity

preservation, while being comparable to them in expression and pose.

#### 4.2 Comparison with Reconstruction Methods

As previously discussed, our task and setting is significantly different than that of *GAN Inversion*. However, if the identity and attribute images fed to our network are the same, our network is implicitly tasked with an inversion problem. For StyleGAN generated faces our

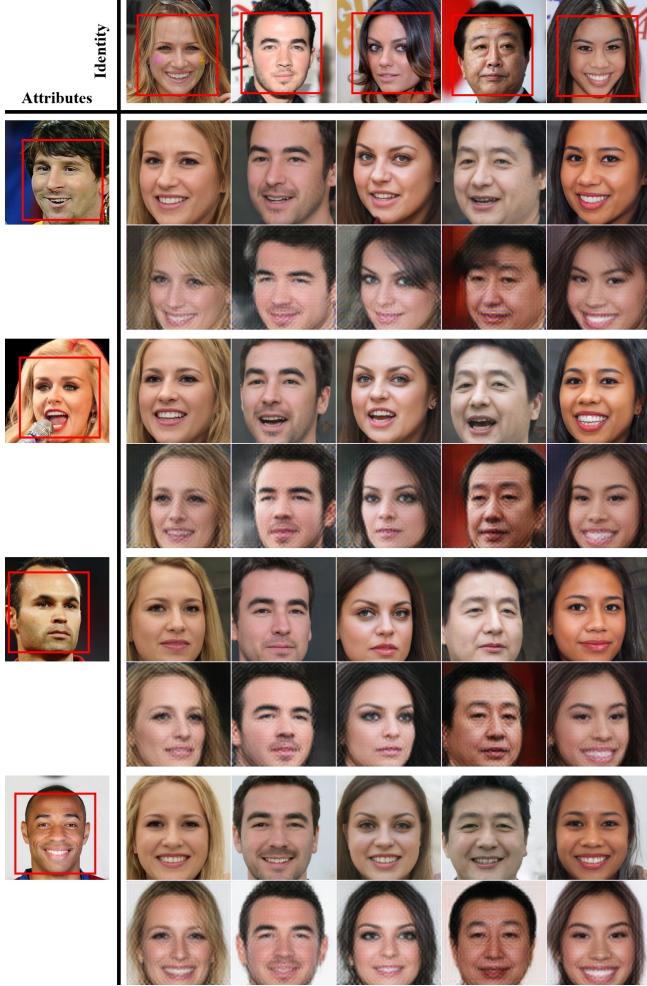


Fig. 6. Qualitative Comparison of our method (odd rows) to LORD [Gabbay and Hoshen 2019] (even rows) on samples from CelebA. Our results have a much better visual quality and preservation of identity and facial attributes (see Table 1). The red frame represents the region LORD gets as input. The reader is recommended to zoom in to better observe details.

inversion is accurate as demonstrated in the diagonal in Figures 2 and 4. We also study the quality of our reconstruction on real faces by comparing it to two recent state-of-the-art encoder-based inversion methods: ALAE [Pidhorskyi et al. 2020] and pSp [Richardson et al. 2020]. All methods were trained on FFHQ and evaluated on CelebA-HQ [Karras et al. 2017]. We perform a quantitative evaluation of the three methods, to assess their reconstruction performance. We randomly sample 5K images from CelebA-HQ for evaluation and evaluate the reconstruction performance by measuring pixel-wise reconstruction with RMSE and PSNR as well as the preservation of semantic features that is identity, expression and pose, in the same manner as performed in Table 1. Results are displayed in Table 2 and a sample of visual results are displayed in Figure 8. As can be observed, pSp is superior, while our method is comparable to ALAE. It is important to stress again, that unlike ALAE and pSp, our method does not aim to reconstruct pixel-level information, as it was devised

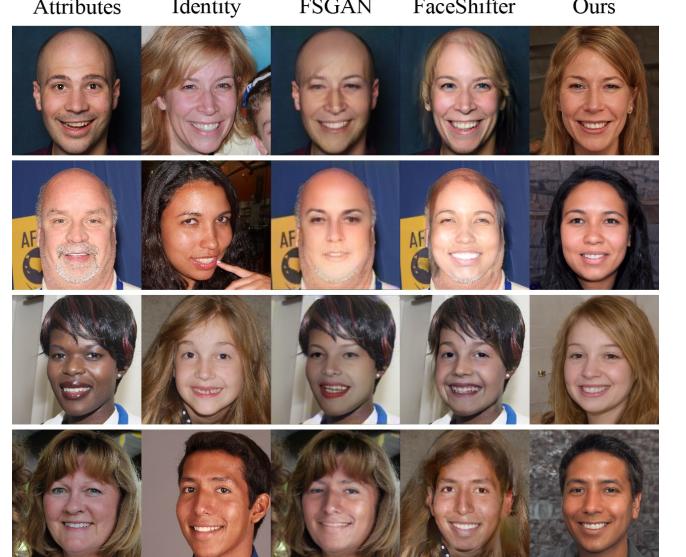


Fig. 7. Qualitative comparison to state-of-the-art face swapping methods FaceShifter [Li et al. 2019] and FSGAN [Nirkin et al. 2019] on samples from FFHQ [Karras et al. 2019b]. As can be seen, our method better preserves the identity as it does not only preserve inner facial features, but the entire head and hair, which are known to be crucial for identity recognition by humans. It can also be observed that face swapping methods struggle with faces with significantly different appearances.

Table 2. Reconstruction quantitative comparison with ALAE [Pidhorskyi et al. 2020] and pSp [Richardson et al. 2020] on CelebA-HQ [Karras et al. 2017]

Method	RMSE ↓	PSNR ↑	ID ↑	Expression ↓	Pose ↓
ALAE	0.192	14.903	0.300	0.014	6.919
pSp	<b>0.095</b>	<b>21.001</b>	<b>0.821</b>	<b>0.008</b>	<b>4.531</b>
Ours	0.202	14.155	0.600	0.013	6.943

for the task of disentanglement rather than reconstruction. However, it is evident that direct inversion approaches that map directly to the enlarged latent space  $\mathcal{W}^+$  [Abdal et al. 2019; Richardson et al. 2020] perform better. This suggests that there is a potential room for improvement in our algorithm by mapping the images into  $\mathcal{W}^+$ .

### 4.3 Ablation Study

In the following we present an ablation study that analyzes the contribution of individual components and losses in our proposed method. One of the most intriguing components of our method is  $D_{\mathcal{W}}$ , which encourages the mapping network to predict latent representations that lie within StyleGAN’s  $\mathcal{W}$  space. To validate its effect, we train a baseline model, that consists of our architecture without the discriminator  $D_{\mathcal{W}}$  and its associated adversarial losses. We explore and compare three different  $\mathcal{W}$  spaces: StyleGAN’s original, ours, and the baseline predicted spaces. We sample 10K vectors from each space and compute PCA, visualized in Figure 9. As can be clearly seen, our predicted  $\mathcal{W}$  coincides with StyleGAN’s  $\mathcal{W}$ ,

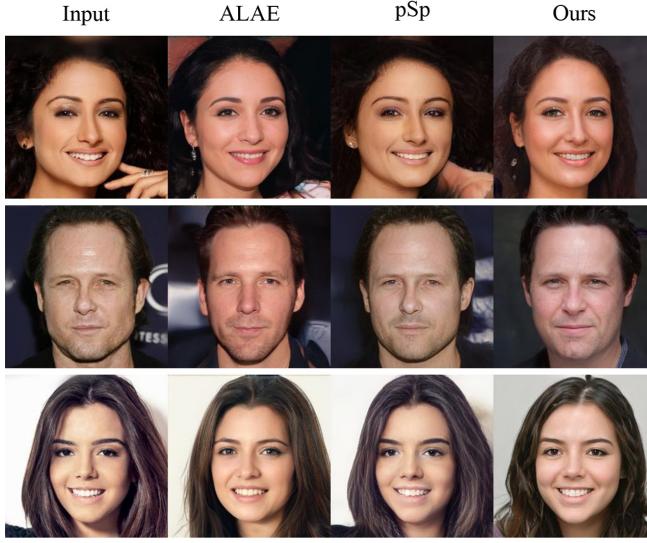


Fig. 8. Qualitative comparison to state-of-the-art learning-based reconstruction methods ALAE [Pidhorskyi et al. 2020] and pSp [Richardson et al. 2020].

Table 3. Quantitative analysis of  $\mathcal{L}_{lnd}$

Method	Identity $\uparrow$	Expression $\downarrow$	Pose $\downarrow$
LORD	$0.20 \pm 0.11$	$0.085 \pm 0.018$	$13.34 \pm 15.00$
Ours w/o $\mathcal{L}_{lnd}$	<b><math>0.63 \pm 0.08</math></b>	$0.023 \pm 0.017$	$13.90 \pm 13.36$
Ours	$0.60 \pm 0.09$	<b><math>0.017 \pm 0.019</math></b>	<b><math>9.74 \pm 13.16</math></b>

while the baseline’s  $\mathcal{W}$  is significantly different, indicating that  $D_{\mathcal{W}}$  is crucial for mapping  $z$  into the actual  $\mathcal{W}$  space of the pre-trained generator. Furthermore, as stated by Karras et al. [2019b], peripheral regions in the latent space often correspond to generated images with inferior quality and resemblance to the real data distribution. Without  $D_{\mathcal{W}}$ , the predictions may lie in such peripheral regions of  $\mathcal{W}$ , causing the generation of erroneous images, such as those in Figure 10.

Next, we evaluate the contribution of our landmarks consistency loss. This loss is calculated using an accessible off-the-shelf network, pre-trained on an independent dataset and task. In the following, we train our network without  $E_{lnd}$ , forming a baseline, and show that the supervision of the landmarks network serves merely to improve the attributes preservation, but is not required for disentanglement. Quantitative evaluation is presented in Table 3. As can be seen, removing  $E_{lnd}$  decreases only the expression and pose preservation. Yet, even with this decrease the baseline is overall superior to LORD in those aspects. Note that, since this landmarks supervision is not necessary to perform the disentanglement, our overall supervision is significantly weaker than the common class-supervision used by other methods.

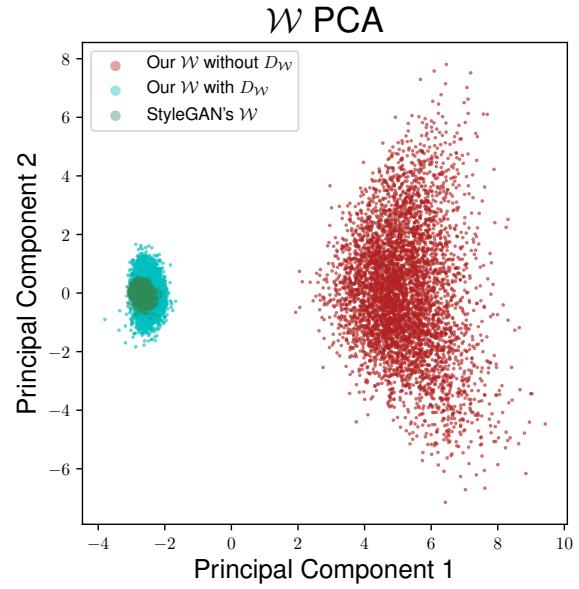


Fig. 9. Dimension reduction of different configurations of  $\mathcal{W}$ , using PCA ( $n=2$ ). As can be clearly seen, our predicted  $\mathcal{W}$  space coincides with StyleGAN’s, expanding it only slightly. However, without  $D_{\mathcal{W}}$  the predicted  $\mathcal{W}$  space is significantly different, stressing the necessity of  $D_{\mathcal{W}}$ .



Fig. 10. The effect of  $D_{\mathcal{W}}$ . Images in the top row were generated by our method. Images in the bottom row were generated by our method, but without  $D_{\mathcal{W}}$ , on the same inputs. As can be seen,  $D_{\mathcal{W}}$  significantly improves image quality, which is also expressed by the FID score that is improved from 6.96 to 4.28.

#### 4.4 Applications

As previously mentioned, our method inherits the properties of the chosen generator  $G$  and its latent space. For the running example of this paper, this generator is StyleGAN. Shen et al. [2019] have demonstrated that StyleGAN’s latent space  $\mathcal{W}$  is well behaved, permitting the smooth editing of features via interpolation of the latent code. However, this and other previous art [Zhu et al. 2020] have demonstrated this property for one-dimensional properties, such

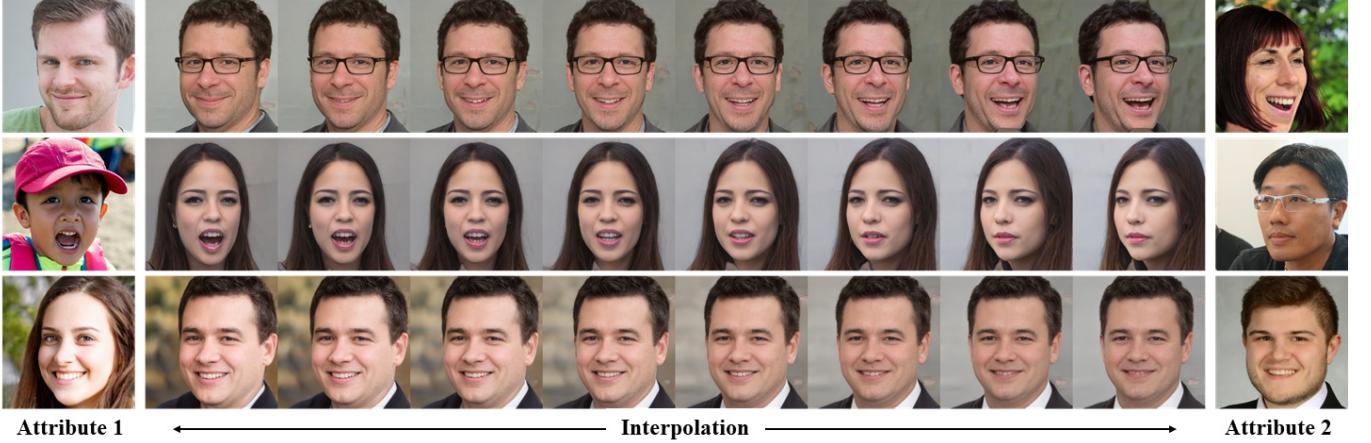


Fig. 11. Disentangled interpolation of attributes while preserving identity. In each line, attributes are extracted from two images (both ends of the spectrum), and the identity is extracted from a third image (not shown). For each of these we infer a  $w$  vector, and interpolate between the resulting two (middle).

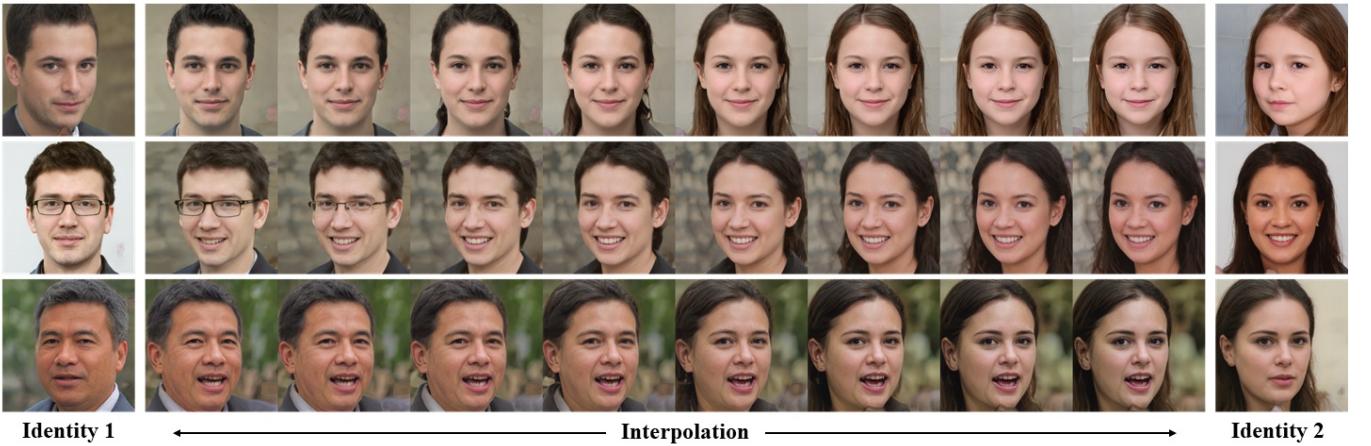


Fig. 12. Disentangled interpolation of identity while preserving the other attributes. The setting is identical to the one in Fig. 11, only here the attributes are extracted from the same image (not shown), and the identity is extracted from two images (both ends of the spectrum), and is interpolated in the space of  $\mathcal{W}$  (middle).

as age or the extent of a smile. In contrast, our proposed mapping identifies latent codes which represent more involved differences, namely the much discussed high-dimensional identity property, or a combination of expression, pose, and lighting.

In Figures 11 and 12 we demonstrate the smooth editing of these elaborate features by interpolation of the latent codes, thus showing that the StyleGAN's latent space  $\mathcal{W}$  is well behaved even with respect to such features, and that we indeed inherit these advantages.

In each figure, the interpolated feature is extracted from the images on the far ends, while the constant feature is extracted from a third image that is not shown. From these inputs we infer two  $w$  vectors and interpolate between them. The images generated by the interpolated values appear in the middle of each of these figures, and portray a pleasant and natural transition between the various explored properties. In Figure 11, we demonstrate that we accurately

and consistently preserve the identity while smoothly and properly interpolate expression, pose and illumination. Specifically, note the successful interpolation of illumination presented on the bottom line of the figure. In Figure 12 we demonstrate that we accurately and consistently preserve the attributes while smoothly interpolating the identity. Note that all images generated during interpolation are of high visual quality and realism.

Next, we turn to examining the coherency of the network in terms of identity preservation. In other words, we examine the stability of the generated identity while perturbing the other attributes. We do this through the generation of sequences. To generate the sequence, we use a single, unseen image to define the target identity, and a sequence of a facial performance to define the rest of the attributes. Generating such a sequence can be considered as a step in the practical direction of the case where the de-identification of the

person in the driving input sequence is desired, or when one would like to reenact a single, unseen, given image. For the case of de-identification, our method is unique because it completely hides the original identity, both from state-of-the-art face recognition networks and from human eyes. This is in contrast to previous methods [Gafni et al. 2019], that perform minimal facial modification to fool face recognition methods. In this case, if the input and output images were put side by side, they would still be recognized as the same person by a human. Our method generates a different person, having a completely different appearance. A sample of our results is displayed in Figure 13. Consecutive frames from the driving sequence are displayed in the first row, and the rest of the rows are our results. Note the different overall appearance of our results compared to the driving sequence. For example, the bottom three rows are generations of different women, all with long hair, while the input is a bald man with a beard. More results, including an animated sequence, can be found in the supplementary material. As can be seen, this simple approach achieves smooth temporal control over pose and expression, and a very stable and coherent identity for the entire sequence, even though every image was generated completely independently. To better observe details, we crop the mouth region from these frames and display them in Figure 14. Note the subtle changes in mouth shapes along the sequence.

## 5 DISCUSSION AND CONCLUSIONS

This paper presented a novel disentanglement method, applied to the highly challenging domain of human heads. The key idea of mapping the disentangled representation to the latent space of a pre-trained GAN is both novel and crucial. It enables state-of-the-art quality synthesis, while requiring modest supervision. Through extensive experimentation, we have further demonstrated the effectiveness and versatility of the method. We have proposed a novel concept of disentanglement, achieved by mapping to the semantically rich latent space of a pre-trained GAN. This concept is generic and could possibly be applied to other data domains and GAN architectures, assuming its latent space is well behaved. Thus, our work is orthogonal to ongoing research on unconditional image generation, from which the proposed framework can only gain.

Furthermore, this concept concentrates on the preservation of image properties, rather than explicitly reducing "leakage" of information between the two parts of the disentangled representation  $\mathcal{Z}$ . Preventing leakage is a sufficient condition for disentanglement, but not a necessary one. We separate the latent space  $\mathcal{Z}$  in which a disentangled representation is formed and the latent space of the generator  $\mathcal{W}$ , allowing greater degree of freedom for both tasks. The mapping process connects these two latent spaces, by taking only the relevant information from each of the parts, and disregards any impurities in the separation of the representation.

As many works [Härkönen et al. 2020; Jahanian et al. 2019; Shen et al. 2019; Zhu et al. 2020] have shown, the latent space of GANs is well-behaved and allows great controlled editing opportunities. All of them, however, have found generic linear directions, along which linear properties can be increased or decreased, in a disentangled fashion. Identity, on the other hand, is a complex and high-dimensional factor that cannot be edited by one-dimensional

value changes. By exhibiting control over the identity, our method continues the direction of demonstrating the incredible strength and possibilities hidden in the latent space of GANs, and StyleGAN in particular.

Our approach further relies on the existence of a network, or any other derivable method, to classify, or evaluate, the feature of interest  $f$ . In the case of human faces, we have also leaned on a similar network to help with identifying facial landmark positions. This was needed due to the extremely sensitive human perception of faces. Other than these networks, our dataset does not contain any labeling, nor several images of the same identity. This setting poses a rather weak form of supervision, especially when compared to the common setting of class-supervised disentanglement. The supervision in our method is manifested solely through the used pre-trained networks. These, however, can be trained on fundamentally different datasets and tasks, so they do not impose an inhibiting requirement.

As discussed, we inherit the generative capabilities of the used pre-trained generator. Of course, alongside these, we also inherit any limitations the generator might have, including those imposed by its training dataset. For example, the preprocessing method used by StyleGAN aligns heads such that there is no roll angle, and renders yaw rotations to be highly correlated with translation. Thus, StyleGAN-based generations, including ours, inherit these properties. Furthermore, it was recently shown that StyleGAN does not cover the entire manifold of human faces and heads, forcing many approaches [Abdal et al. 2019; Baylies 2019; Zhu et al. 2020] to work with an artificially enlarged latent space, named  $\mathcal{W}^+$ . Introducing manipulations on  $\mathcal{W}^+$  to our methods may significantly increase the expressiveness of our model, but may come at the cost of both generation and disentanglement qualities. We leave this investigation as an exciting avenue for future work. Regardless, this work introduces a powerful concept, where generative networks are employed as "backbones" for disentanglement tasks, which can most probably be explored much further in future research.

## ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments. We also thank Tal Hassner, Yuval Nirkin, Aviv Gabbay and Mingchao Sun for their help and useful suggestions. This work was supported by the Israel Science Foundation (grant no. 2366/16 and 2472/17) and in part by the National Science Foundation of China General Program grant No. 61772317.

## REFERENCES

- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *Proceedings of the IEEE International Conference on Computer Vision*. 4432–4441.
- Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Learning character-agnostic motion for motion retargeting in 2D. *arXiv preprint arXiv:1905.01680* (2019).
- Naphtali Abudarham, Lior Shkoller, and Galit Yovel. 2019. Critical features for face recognition. *Cognition* 182 (2019), 73–83.
- Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)* 36, 6 (2017), 196.
- Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2018. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6713–6722.



Fig. 13. Talking head sequence. The first row consists of consecutive frames from a driving attributes sequence, while the rest of the rows are frames generated by our method. We demonstrate smooth control over facial expression and pose while maintaining constant identity.



Fig. 14. Cropped mouth regions from Fig. 13. Our method is able to preserve subtle lips movement, critical for talking head sequence realism.

- Baylies. 2019. stylegan-encoder. <https://github.com/pbaylies/stylegan-encoder>. Accessed: April 2020.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
- Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 67–74.
- Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8789–8797.
- Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2019. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *arXiv preprint arXiv:1912.01865* (2019).
- Antonia Creswell and Amil Anthony Bharath. 2018. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems* 30, 7 (2018), 1967–1974.
- deepfakes. 2019. faceswap. <https://github.com/deepfakes/faceswap>. Accessed: April 2020.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.
- Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439* (2019).
- Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*. 4414–4423.
- Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2235–2245.
- Aviv Gabbay and Yedid Hoshen. 2019. Demystifying Inter-Class Disentanglement. *arXiv preprint arXiv:1906.11796* (2019).
- Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live Face De-Identification in Video. In *Proceedings of the IEEE International Conference on Computer Vision*. 9378–9387.
- Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. GANalyze: Toward Visual Definitions of Cognitive Image Properties. *arXiv:cs.CV/1906.10112*
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
- Naama Hadad, Lior Wolf, and Moni Shahar. 2018. A two-step disentanglement method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 772–780.

- Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. *arXiv preprint arXiv:2004.02546* (2020).
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr* 2, 5 (2017), 6.
- Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. 2018. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in neural information processing systems*. 52–63.
- Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the “steerability” of generative adversarial networks. *arXiv preprint arXiv:1907.07171* (2019).
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
- Tero Karras, Samuli Laine, and Timo Aila. 2019b. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4401–4410.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019a. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958* (2019).
- Heeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.
- Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. *arXiv preprint arXiv:1802.05983* (2018).
- Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1412.6980>
- Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *arXiv preprint arXiv:1912.13457* (2019).
- Zachary C Lipton and Subarna Tripathi. 2017. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782* (2017).
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*. 700–708.
- Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. 2018. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2080–2089.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2018. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359* (2018).
- Junyu Luo, Yong Xu, Chenwei Tang, and Jiancheng Lv. 2017. Learning inverse mapping by autoencoder based generative adversarial nets. In *International Conference on Neural Information Processing*. Springer, 207–216.
- Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. 2018. Disentangling Disentanglement in Variational Autoencoders. [arXiv:stat.ML/1812.02833](https://arxiv.org/abs/stat.ML/1812.02833)
- Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge? *arXiv preprint arXiv:1801.04406* (2018).
- Masahiro Mori et al. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.
- Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*. 7184–7193.
- Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. 2016. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355* (2016).
- Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. [to appear].
- Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 818–833.
- Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. 2015. Deep visual analogy-making. In *Advances in neural information processing systems*. 1252–1260.
- Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *arXiv preprint arXiv:2008.00951* (2020).
- Omry Sendik, Dani Lischinski, and Daniel Cohen-Or. 2019. What’s in a Face? Metric Learning for Face Characterization. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 405–416.
- Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2019. Interpreting the latent space of gans for semantic face editing. *arXiv preprint arXiv:1907.10786* (2019).
- Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. 2018. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 821–830.
- Pawan Sinha and Tomaso Poggio. 1996. I think I know that face... *Nature* 384, 6608 (1996), 404–404.
- Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. 2018. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 553–569.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
- Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. *arXiv preprint arXiv:2004.00121* (2020).
- Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.
- Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. 2018. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.
- Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. 2018. CR-GAN: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191* (2018).
- Umar Toseeb, David RT Keeble, and Eleanor J Bryant. 2012. The significance of hair for face recognition. *PloS one* 7, 3 (2012).
- Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069* (2018).
- Yifan Wu, Fan Yang, and Haibin Ling. 2018. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1806.08906* (2018).
- Fanyi Xiong, Haotian Liu, and Yong Jae Lee. 2019. Identity from here, Pose from there: Self-supervised Disentanglement and Generation of Objects using Unlabeled Videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 7013–7022.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*. 9459–9468.
- Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.
- Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging* 3, 1 (2016), 47–57.
- Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-Domain GAN Inversion for Real Image Editing. *arXiv preprint arXiv:2004.00049* (2020).
- Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*. Springer, 597–613.

## Appendix

### A TEMPORAL COHERENT SEQUENCES

As presented in Section 4.4 in the main paper, our method can generate temporally coherent sequences. To test the stability of the generated identity, we perturb facial attributes through a driving sequence of a facial performance. Figure 15 depicts how the identity is preserved almost to perfection in the presence of varying driving attributes, even though each image is generated independently. The sequences along each row are generated using the same driving sequence, and hence portray corresponding facial expression and pose. Note how the identity remains constant along the entire sequence, and how the expression, illumination and pose attributes match well along each row. The successful temporal coherence in the above experiment holds a promising avenue for future research in one-shot face reenactment.

### B INTERPOLATION IN $\mathcal{Z}$

In Figures 16 and 17 we present the results of an experiment regarding the disentangled interpolation in  $\mathcal{Z}$ . This is different than the disentangled interpolation in  $\mathcal{W}$ , which we present in Section 4.4 in the paper. Equivalently to the latter figure, in each figure, the interpolated feature is extracted from the images on the two far ends, while the constant feature is extracted from a third image, which is not shown. These extracted features are our disentangled representations. We keep the representation of the constant feature fixed, while interpolating between the two representations of the other feature. The fixed and interpolated values are concatenated to form our latent code  $z$ . Our mapping networks converts this code to  $w$  - a code in StyleGAN’s latent space, which in turn is used to generate the image. The generated in-between images of each of these figures portray a pleasant and natural transition between the various explored properties, like they do in the parallel experiment in the paper. In Figure 16, we demonstrate that we accurately and consistently preserve the identity, while smoothly interpolating the expressions, poses and illumination, which are also well preserved from their respective inputs. In Figure 17, we demonstrate that we accurately and consistently preserve the attributes, while smoothly interpolating the identity. Specifically, note the successful interpolation of illumination presented on the bottom row of Figure 16 and the smooth disappearance of sunglasses, while maintaining high realism in the middle row of Figure 17.

It is important to distinguish between interpolation in  $\mathcal{W}$  and  $\mathcal{Z}$ . The well-behaved interpolation in  $\mathcal{W}$  is inherited by StyleGAN, which is not the case for  $\mathcal{Z}$ , that is learned by our method. The fact that interpolating one factor representation in  $\mathcal{Z}$ , while keeping the other fixed, causes only a single factor to change in the generating image indicates that our representation is in fact disentangled.

### C HIGH RESOLUTION RESULTS

All results in the main paper were obtained using a pre-trained StyleGAN with 256x256 resolution, to make results better comparable to previous methods. However, our method is able to fully harness the power of StyleGAN by generating 1024x1024 resolution. We

Fig. 15. Sample results for temporally coherent sequence generation experiment. For each identity, a single never-before-seen image was given, note the temporal coherency of identity along all sequences, and the coherency of the rest of the attributes along each row. NOTE: this is an animated figure. Viewing it using Adobe Acrobat, or another animation supporting viewer, is strongly recommended.

simply replace the pretrained 256x256 network with a pre-trained 1024x1024 one, and train in the same manner. In order to adjust to the higher memory requirements, we change the batch size to 2, and divide all learning rates by 10. The network is trained end-to-end on a single NVIDIA Titan XP GPU and requires roughly 3 days to converge. Note that this is incredibly efficient, compared to training StyleGAN itself on the same GPU, which requires roughly 60 days. Results are displayed in Figures 18 to 21.

### D RESULTS

We provide more 256x256 results, generated by our method in Figures 23 and 24.

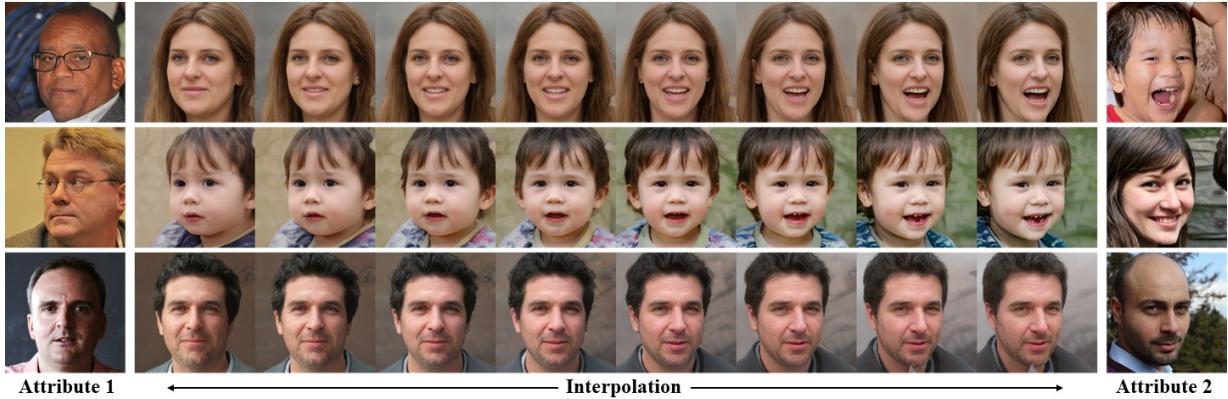


Fig. 16. Disentangled interpolation of attributes while preserving identity. In each line, attributes representations are extracted from two images (both ends of the spectrum), and the identity representation is extracted from a third image (not shown). We interpolate the attributes representation, between the two extracted value. We then concatenate the interpolated attributes value with the fixed identity representation, forming  $z$ , a new whole latent face representation which is then used to generate an image (middle).

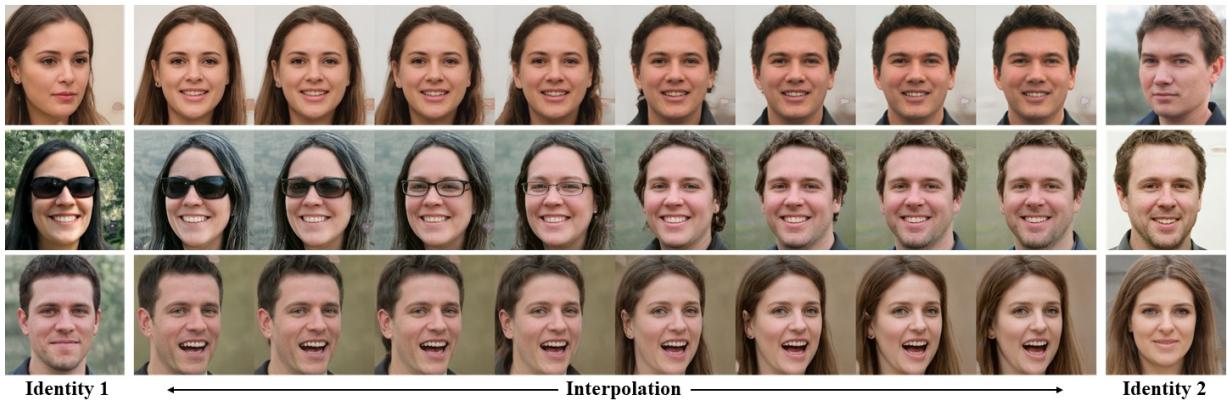


Fig. 17. Disentangled interpolation of identity while preserving the other attributes. The setting is identical to the one in Fig. 16, only here the attributes are extracted from the same image (not shown), and the identity is extracted from two images (both ends of the spectrum).

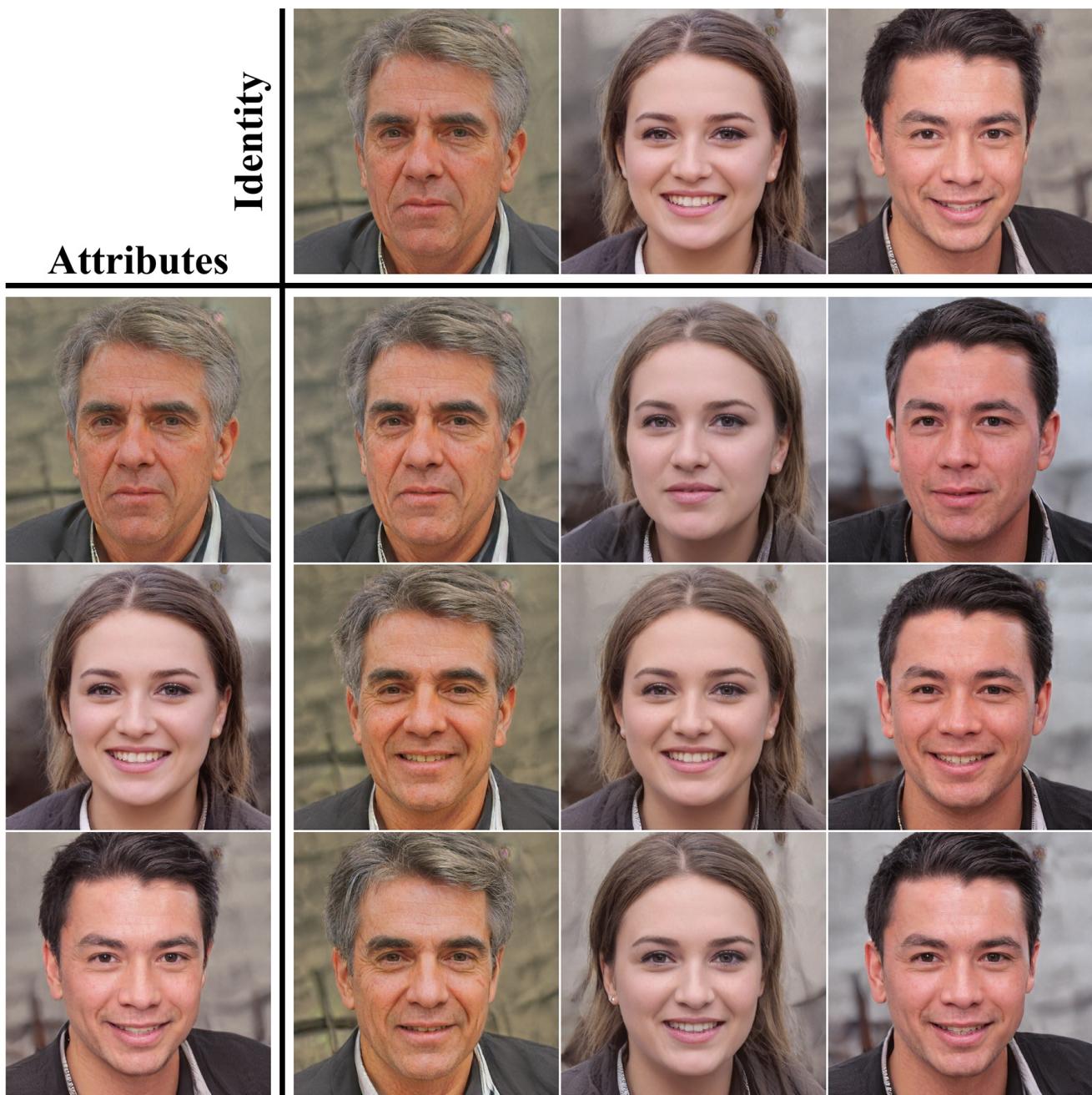


Fig. 18. Feature combination results, in 1024x1024 resolution. For every image in the table, identity is taken from the top, and the rest of the attributes (including expressions, pose, lighting conditions, etc.) from the left. All images were generated using StyleGAN generator.

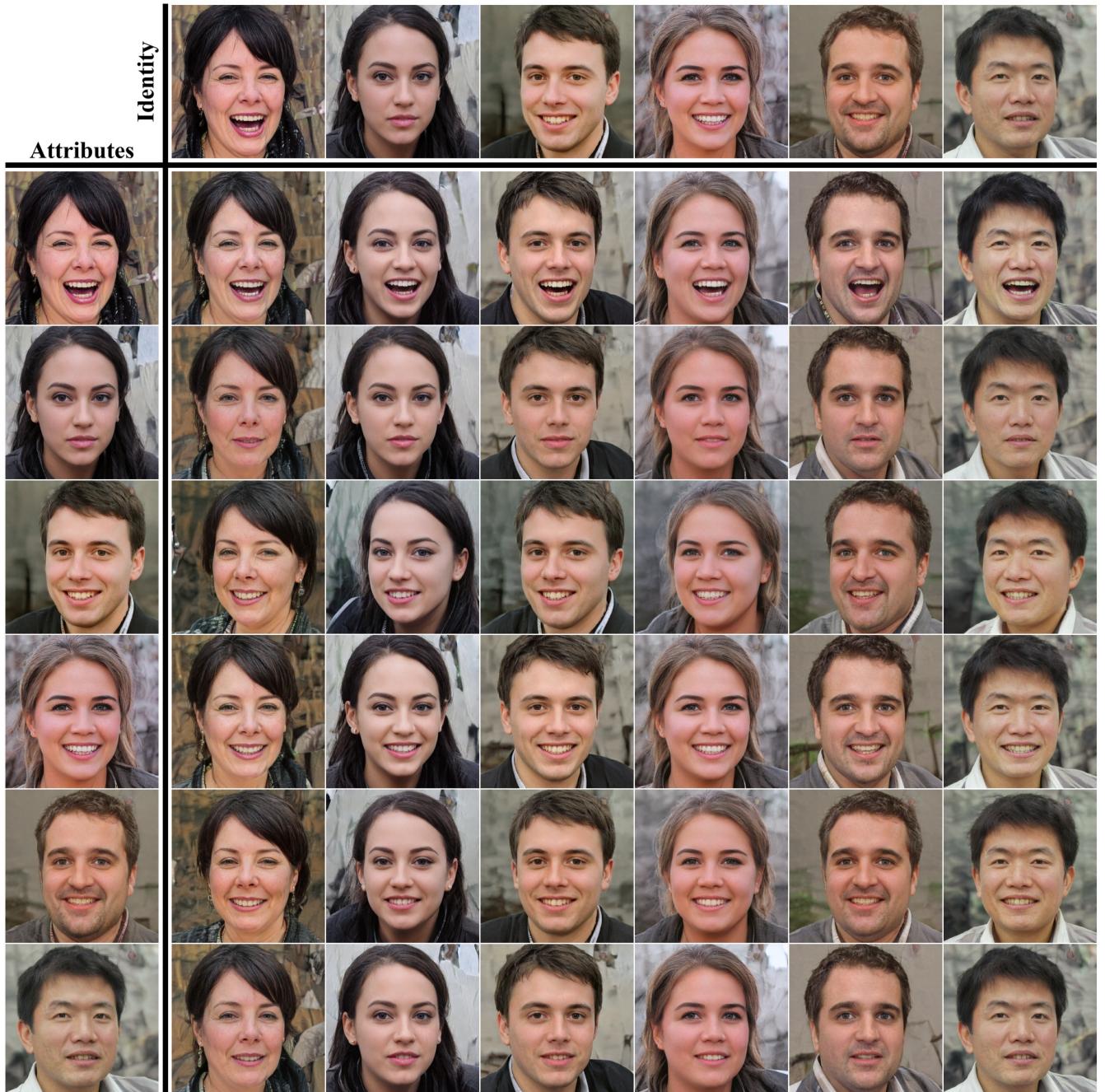


Fig. 19. Feature combination results, in 1024x1024 resolution, as in Fig. 18.



Fig. 20. Feature combination results, in 1024x1024 resolution, as in Fig. 18.



Fig. 21. Feature combination results, in 1024x1024 resolution, as in Fig. 18.



Fig. 22. Feature combination results, in 1024x1024 resolution. The setting is identical to 18, only here input images are sampled from FFHQ



Fig. 23. Further feature combination results, in 256x256 resolution. For every image in the table, identity is taken from the top, and the rest of the attributes (including expressions, pose, lighting conditions, etc.) from the left. All images were generated using StyleGAN.



Fig. 24. Further feature combination results, in 256x256 resolution, as in Fig. 23