Live Face De-Identification in Video

Oran Gafni, Lior Wolf
Facebook AI Research and Tel-Aviv University
{oran,wolf}@fb.com

Yaniv Taigman
Facebook AI Research
yaniv@fb.com

Abstract

We propose a method for face de-identification that enables fully automatic video modification at high frame rates. The goal is to maximally decorrelate the identity, while having the perception (pose, illumination and expression) fixed. We achieve this by a novel feed-forward encoder-decoder network architecture that is conditioned on the high-level representation of a person's facial image. The network is global, in the sense that it does not need to be retrained for a given video or for a given identity, and it creates natural looking image sequences with little distortion in time.

1. Introduction

In consumer image and video applications, the face has a unique importance that stands out from all other objects. For example, face recognition (detection followed by identification) is perhaps much more widely applicable than any other object recognition (categorization, detection, or instance identification) in consumer images. Similarly, putting aside image processing operators that are applied to the entire frame, face filters remain the most popular filters for consumer video. Since face technology is both useful and impactful, it also raises many ethical concerns. Face recognition can lead to loss of privacy and face replacement technology may be misused to create misleading videos.

In this work, we focus on video de-identification, which is a video filtering application that both requires a technological leap over the current state-of-the-art, and is benign in nature. This application requires the creation of a video of a similar looking person, such that the perceived identity is changed. This allows, for example, the user to leave a natural-looking video message in a public forum in an anonymous way, that would presumably prevent face recognition technology from recognizing them.

Video de-identification is a challenging task. The video needs to be modified in a seamless way, without causing flickering or other visual artifacts and distortions, such that the identity is changed, while all other factors remain identical, see Fig. 1. These factors include pose, expression, lip positioning (for unaltered speech), occlusion, illumination and shadow, and their dynamics.

Figure 1. De-identification video results demonstrated on a variety of poses, expressions, illumination conditions and occlusions. Pairs of the source frame (first row) and the output frame (second row) are shown. The high-level features (e.g. nose, eyes, eyebrows and mouth) are altered, while the pose, expression, lip articulation, illumination, and skin tone are preserved.

In contrast to the literature methods, which are limited to still images and often swap a given face with a dataset face, our method handles video and generates de novo faces. Our experiments show convincing performance for unconstrained videos, producing natural looking videos. The person in the rendered video has a similar appearance to the person in the original video. However, a state-of-the-art face-recognition network fails to identify the person. A similar experiment shows that humans cannot identify the generated face, even without time constraints.

Our results would not have been possible, without a host of novelties. We introduce a novel encoder-decoder architecture, in which we concatenate to the latent space the activations of the representation layer of a network trained to perform face recognition. As far as we know, this is the first time that a representation from an existing classifier network is used to augment an autoencoder, which enables the feed-forward treatment of new persons, unseen during training. In addition, this is the first work to introduce a new

	Newton, '05 [32]	Gross, '08 [10]	Samarzija, '14 [41]	Jourabloo, '15 [16]	Meden, '17 [31]	Wu, '18 [49]	Sun'18 [43, 44]	Our
Preserves expression	-	-	-	-	-	-	-	+
Preserves pose	-	+	+	-	+	-	+	+
Generates new faces	-	†	-	†	+	+	+	+
Demonstrated on video	-	-	-	-	-	-	-	+
Demonstrated on a diverse dataset (gender, ethnicity, age, etc.)	-	+	-	+	-	-	-	+
Reference to a comparison with ours			Fig. 7		Fig. 4	Fig. 8	Fig. 5, 14	

Table 1. A comparison to the literature methods. The final row references comparison figures in this work. We compare to all methods that provide reasonable quality images in their manuscript, under conditions that are favorable to previous work (we crop the input images from the pdf files, except for the images received from the authors of [43, 44]). †The face is swapped with an average of a few dataset faces.

type of attractor-repeller perceptual loss term. This term distinguishes between low- and mid-level perceptual terms, and high-level ones. The former are used to tie the output frame to the input video frame, while the latter is used to distance the identity. In this novel architecture, the injection of the representation to the latent space enables the network to create an output that adheres to this complex criterion. Another unique feature is that the network outputs both an image and a mask, which are used, in tandem, to reconstruct the output frame. The method is trained with a specific data augmentation technique that encourages the mapping to be semantic. Additional terms include reconstruction losses, edge losses, and an adversarial loss.

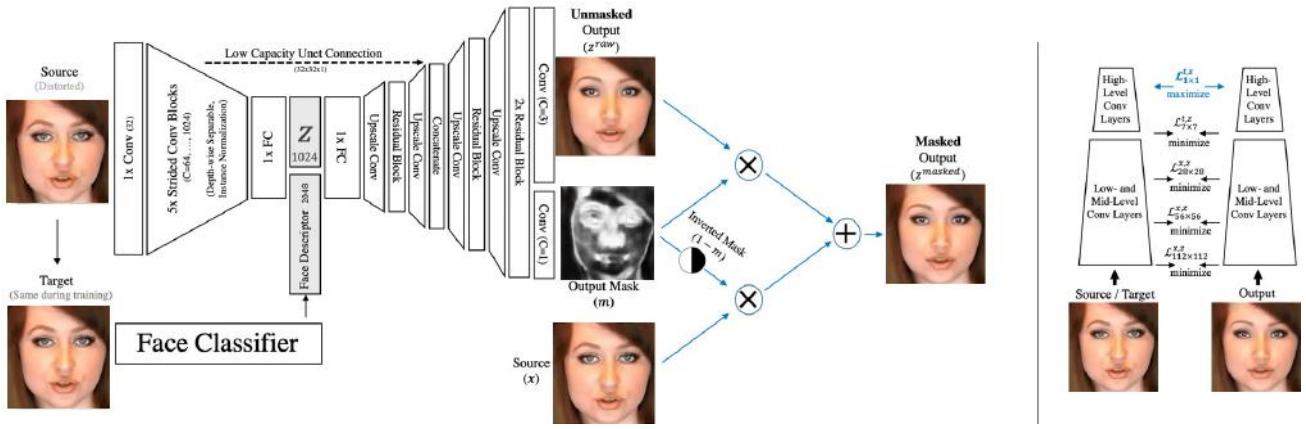
2. Previous Work

Faces have been modeled by computer graphics systems for a long time. In machine learning, faces have been one of the key benchmarks for GAN-based generative models [9, 37, 40] since their inception. High resolution natural looking faces were recently generated by training both the generator and the discriminator of the GAN progressively, starting with shallower networks and lower resolutions, and enlarging them gradually [17].

Conditional generation of faces has been a key task in various unsupervised domain translation contributions, where the task is to learn to map, e.g., a person without eyewear to a person with eyeglasses, without seeing matching samples from the two domains [20, 51, 1, 27]. For more distant domain mapping, such as mapping between a face image and the matching computer graphics avatar, additional supervision in the form of a face descriptor network was used [45]. Our work uses these face descriptors, in order to distance the identity of the output from that of the input.

As far as we know, our work is the first de-identification work to present results on videos. In still images, several methods have been previously suggested. Earlier work implemented different types of image distortions for face de-identification [33, 10], while more recent works rely on techniques for selecting distant faces [41] or averaging/fusing faces from pre-existing datasets [32, 16, 31]. The experiments conducted by the aforementioned techniques are restricted, in most cases, to low-resolution, black and white results. Although it is possible to create eye-pleasing results, they are not robust to different poses, illuminations and facial structures, making them inadequate for video generation. The use of GANs for face de-identification has been suggested [49]. However, the experiments were restricted to a homogeneous dataset, with no apparent expression preservation within the results. In the GAN-based methods of [43, 44], face de-identification is employed for the related task of person obfuscation. The work of [43] conditions the output image based on both a blurred version of the input and the extracted facial pose information. The follow-up work [44] combines the GAN-based reconstruction with a parametric face generation network. As both methods are applied over full upper-body images, they result in low facial resolution outputs of 64×64. These methods do not preserve expressions, are unsuitable for video, and occasionally provide unnatural outputs.

Tab. 1 provides a comparative view of the literature. The current literature on de-identification often involves face swapping (our method does not). Face swapping, i.e., the replacement of a person's face in an image with another person's face, has been an active research topic for some time, starting with the influential work of [3, 2]. Recent contributions have shown a great deal of robustness to the source image, as well as for the properties of the image, from which the target face is taken [19, 34]. While these classical face swapping methods work in the pixel space and copy the expression of the target image, recent deep-learning based work swaps the identity, while maintaining the other aspects of the source image [23]. In comparison to our work, [23] requires training a new network for every target person, the transferred expression does not show subtleties (which would be critical, e.g., for a speaking person), and the results are not as natural as ours. These limitations are probably a result of capturing the appearance of the target, by restricting the output to be similar, patch by patch, to a collection of



(a) (b)

Figure 2. (a) The architecture of our network. For conditioning, a pre-trained face recognition network is used. (b) An illustration of the multi-image perceptual loss used, which employs two replicas of the same face recognition network.

patches from the target person. Moreover, [23] is limited to stills and was not demonstrated on video.

The face swapping (FS) project [8] is an unpublished work that replaces faces in video in a way that can be very convincing, given suitable inputs. Unlike our network, the FS is retrained for every pair of source-video and target-video persons. The inputs to the FS system, during training, are two large sets of images, one from each identity. In order to obtain good results, thousands of images from each individual with a significant variability in pose, expression, and illumination are typically used. In many cases, a large subset of the images of the source person are taken from the video that is going to be converted. In addition, FS often fails, and in order to obtain a convincing output, the person in the source video and the target person need to have a similar facial structure. These limitations make it unsuitable for de-identification purposes.

Like ours, the FS method is based on an encoder-decoder architecture, where both an image and output mask are produced. A few technical novelties of FS are shared with our work. Most notable is the way in which augmentation is performed in order to train a more semantic encoder-decoder network. During the training of FS, the input image is modified by rotating or scaling it, before it is fed to the encoder. The image that the decoder outputs is compared to the undistorted image. Another common property is that the GAN variant used employs virtual examples created using the mixup technique [52]. In addition, in order to maintain the pose and expression, which are considered low- or mid-level features in face descriptors (orthogonal to the identity) FS employs a perceptual loss [15, 47] that is based on the layers of a face-recognition network.

Another line of work that manipulates faces in video is face reanimation, e.g., [46]. This line of work reanimates the face in the target video, as controlled by the face in a source video. This does not provide a de-identification solution in the sense that we discuss – the output video is reanimated in a different scene, and not in the scene of the source video. In addition, it always provides the same output identity.

We do not enforce disentanglement [14, 26, 5] between the latent representation vector Z and the identity, since the network receives the full information regarding the identity using the face descriptor. Therefore, washing out the identity information in Z may not be beneficial. Similarly, the U-Net connection means that identity information can bypass Z. In our method, the removal of identity is not done through disentanglement but via the perceptual loss. As Fig. 9 demonstrates, this loss provides a direct and quantifiable means for controlling the amount of identity information. With disentanglement, this effect would be brittle and sensitive to hyperparameters, as is evident in work where the encoding is set to be orthogonal, even to simple multiclass label information, e.g, [25].

3. Method

Our architecture is based on an adversarial autoencoder [29], coupled with a trained face-classifier. By concatenating the autoencoder's latent space with the face-classifier representation layer, we achieve a rich latent space, embedding both identity and expression information. The network is trained in a counter-factual way, i.e., the output differs from the input in key aspects, as dictated by the conditioning. The generation task is, therefore, highly semantic, and the loss required to capture its success cannot be a conventional reconstruction loss.

For the task of de-identification, we employ a target image, which is any image of the person in the video. The method then distances the face descriptors of the output video from those of the target image. The target image does not need to be based on a frame from the input video. This

contributes to the applicability of the method, allowing it to be applied to live videos. In our experiments, we do not use an input frame in order to show the generality of the approach. To encode the target image, we use a pre-trained face classifier ResNet-50 network [12], trained over the VGGFace2 dataset [4].

The process during test time is similar to the steps taken in the face swapping literature and involves the following steps: (a) A square bounding box is extracted using the 'dlib' [21] face detector. (b) 68 facial points are detected using [18]. (c) A transformation matrix is extracted, using an estimated similarity transformation (scale, rotation and translation) to an averaged face. (d) The estimated transformation is applied to the input face. (e) The transformed face is passed to our network, together with the representation of the target image, obtaining both an output image and a mask. (f) The output image and mask are projected back, using the inverse of the similarity transformation. (g) We generate an output frame by linearly mixing, per pixel, the input and the network's transformed output image, according to the weights of the transformed mask. (h) The outcome is merged into the original frame, in the region defined by the convex hull of the facial points.

At training time, we perform the following steps: (a) The face image is distorted and augmented. This is done by applying random scaling, rotation and elastic deformation. (b) The distorted image is fed into the network, together with the representation of a target image. During training, we select the same image, undistorted. (c) A linear combination of the masked output (computed as in step (g) above) and the undistorted input is fed to the discriminator. This is the mixup technique [52] discussed below. (d) Losses are applied on the network's mask and image output, as well as to the masked output, as detailed below.

Note that there is a discrepancy between how the network is trained and how it is applied. Not only do we not make any explicit effort to train on videos, the target images are selected in a different way. During training, we extract the identity from the training image itself and not from an independent target image. The method is still able to generalize to perform the real task on unconstrained videos.

3.1. Network architecture

The architecture is illustrated in Fig. 2(a). The encoder is composed of a convolutional layer, followed by five strided, depth-wise separable [6] convolutions with instance normalization [48]. Subsequently, a single fully connected layer is employed, and the target face representation is concatenated. The decoder is composed of a fully connected layer, followed by a lattice of upscale and residual [12] blocks, terminated with a $tanh$ activated convolution for the output image, and a sigmoid activated convolution for the mask output. Each upscale block is comprised of a 2D convolution, with twice the number of filters as the input channel size. Following an instance normalization and a LReLU [11] activation, the activations are re-ordered, so that the width and height are doubled, while the channel size is halved. Each residual block input is summed with the output of a Conv2D-LReLU-Conv2D chain.

A low-capacity U-net connection [38] is employed (32x32x1), thus relieving the autoencoder's bottleneck, allowing a stronger focus on the encoding of transfer-related information. The connection size does not exceed the bottleneck size (1024) and due to the distortion of the input image, a collapse into a simple reconstructing autoencoder in early training stages is averted.

The discriminator consists of four strided convolutions with LReLU activations, with instance normalization applied on all but the first one. A sigmoid activated convolution yields a single output.

The network has two versions: a lower resolution version generating 128x128 images, and a higher resolution version, generating 256x256 images. The higher resolution decoder is simplified and enlarged and consists of a lattice of 6x(Upscale block –> Residual block). Unless otherwise specified, the results presented in the experiments are done with the high-res model.

3.2. Training and the Losses Used

For training all networks, except for the discriminator D, we use a compound loss $\mathcal{L}$, which is a weighted sum of multiple parts:

$$\begin{aligned}\mathcal{L} =&\alpha_0\mathcal{L}_G + \alpha_1\mathcal{L}_R^{raw} + \alpha_1\mathcal{L}_R^{masked} + \alpha_2\mathcal{L}_x^{raw} \\ &+ \alpha_2\mathcal{L}_y^{raw} + \alpha_2\mathcal{L}_x^{masked} + \alpha_2\mathcal{L}_y^{masked} \\ &+ \alpha_3\mathcal{L}_p^{raw} + \alpha_3\mathcal{L}_p^{masked} + \alpha_4\mathcal{L}^m + \alpha_5\mathcal{L}_x^m + \alpha_5\mathcal{L}_y^m,\end{aligned}$$

where $\mathcal{L}_G$ is the generator's loss, $\mathcal{L}_R^{raw}$ and $\mathcal{L}_R^{masked}$ are reconstruction losses for the output image of the decoder z^{raw} and the version after applying the masking z^{masked}, $\mathcal{L}_x^*$ and $\mathcal{L}_y^*$ are reconstruction losses applied to the spatial images derivatives, $\mathcal{L}_p^*$ are the perceptual losses, and $\mathcal{L}_*^m$ are regularization losses on the mask. The discriminator network is trained using its own loss $\mathcal{L}_D$. Throughout our experiments, we employ $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 0.5, \alpha_4 = 3 \cdot 10^{-3}, \alpha_5 = 10^{-2}$.

To maintain realistic looking generator outputs, an adversarial loss is used with a convex combination of example pairs (known as mixup) [52] over a Least Square GAN [30]:

$$\begin{aligned}\mathcal{L}_D &= \|D(\delta_{mx}) - \lambda_\beta \mathbb{1}\|_2^2 \\ \mathcal{L}_G &= \alpha_0\|D(\delta_{mx}) - (1-\lambda_\beta)\mathbb{1}\|_2^2\end{aligned}$$

While, $\delta_{mx} = \lambda_\beta \cdot x + (1-\lambda_\beta)z^{masked}$ and λ_β is sampled out of a Beta distribution $\lambda_\beta \sim Beta(\alpha, \alpha)$, x is the undistorted input "real" sample and z^{masked} is the post masking

generated sample. A value of $\alpha = 0.2$ is used throughout the experiments.

Additional losses are exercised to both retain source-to-output similarity, yet drive a perceptible transformation. Several losses are distributed equally between the raw and masked outputs, imposing constraints on both. An L1 reconstruction loss is used to enforce pixel-level similarity:

$$\mathcal{L}_R^{raw} = \alpha_1 \|z^{raw} - x\|_1 \quad \mathcal{L}_R^{masked} = \alpha_1 \|z^{masked} - x\|_1$$

where z^{raw} is the output image itself. This results in a non-trivial constraint, as the encoder input image is distorted. An edge-preserving loss is used to constrain pixel-level derivative differences in both the x and y image axes. Calculated as the absolute difference between the source and output derivatives in each axis direction for both the raw and masked outputs:

$$\mathcal{L}_x^{raw} = \alpha_2 \|z_x^{raw} - x_x\|_1 \quad \mathcal{L}_x^{masked} = \alpha_2 \|z_x^{masked} - x_x\|_1$$
$$\mathcal{L}_y^{raw} = \alpha_2 \|z_y^{raw} - x_y\|_1 \quad \mathcal{L}_y^{masked} = \alpha_2 \|z_y^{masked} - x_y\|_1$$

where x_x is the derivative of the undistorted input image x along the x axis, and similarly for outputs z and the y axis.

Additional losses are applied to the blending mask m, where 0 indicates that the value of this pixel would be taken from the input image x, 1 indicates taking the value from z^{raw}, and intermediate values indicate linear mixing. We would like the mask to be both minimal and smooth and, therefore, employ the following losses:

$$\mathcal{L}^m = \|m\|_1 \qquad \mathcal{L}_x^m = \|m_x\|_1 \qquad \mathcal{L}_y^m = \|m_y\|_1$$

where m_x and m_y are the spatial derivatives of the mask.

3.2.1 A Multi-Image Perceptual Loss

A new variant of the perceptual loss [15] is employed to maintain source expression, pose and lighting conditions, while capturing the target identity essence. This is achieved by employing a perceptual loss between the undistorted source and generated output on several low-to-medium abstraction layers, while distancing the high abstraction layer perceptual loss between the target and generated output.

Let $a_{n\times n}^r$ be the activations of an $n \times n$ spatial block within the face classifier network for image r, where in our case, r can be either the input image x, the target image t, the raw output z^{raw}, or the masked output z^{masked}.

We consider the spatial activations maps of size $112 \times 112, 56 \times 56, 28 \times 28$ and 7×7, as well as the representation layer of size 1×1. The lower layers (larger maps) are used to enforce similarity to the input image x, while the 7×7 layer is used to enforce similarity to t, and the 1×1 feature vector is used to enforce dissimilarity to the target image.

Let us define $\ell_{n\times n}^{r_1,r_2} = c_n \|a_{r_1,n\times n} - a_{r_2,n\times n}\|_1$, where c_n is a normalizing constant, corresponding to the size of the spatial activation map.

Figure 3. Sample results for video de-identification (zoom). Triplets of source frame, converted frame and target are shown. The modified frame looks similar but the identity is completely different.

The perceptual loss is given by:

$$\mathcal{L}_p^c = \ell_{112\times112}^{x,z^c} + \ell_{56\times56}^{x,z^c} + \ell_{28\times28}^{x,z^c} + \ell_{7\times7}^{t,z^c} - \lambda \ell_{1\times1}^{t,z}$$

for c that is either raw or $masked$, and where $\lambda > 0$ is a hyperparameter, which determines the generated face's high level features distance from those of the target image.

The application of the multi-image perceptual loss during training is depicted in Fig. 2(b). During training, the target is the source, and there is only one input image. The resulting image has the texture, pose and expression of the source, but the face is modified to distance the identity. Note that we refer to it as a multi-image perceptual loss, as its aim is to minimize the analog error term during inference (generalization error). However, as a training loss, it is only applied during train, where it receives a pair of images, similar to other perceptual losses.

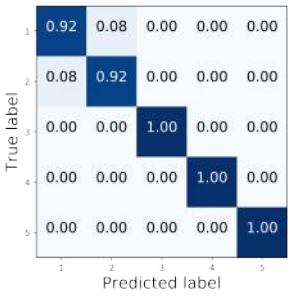
Note that the perceptual loss parameters c_n are normalizing constants obtained by counting the number of elements. In addition, $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3$ are simply set to one, and α_4, α_5 were chosen arbitrarily. Therefore, there is effectively, only a single important hyperparameter: λ, which provides a direct control of the strength of the identity distance which requires tuning (see Fig. 9).

At inference time, the network is fed an input frame and a target image. The target image is transmitted through the face classifier, resulting in a target feature vector, which, in turn, is concatenated to the latent embedding space. Due to the way the network is trained, the decoder will drive the output image away from the target feature vector.

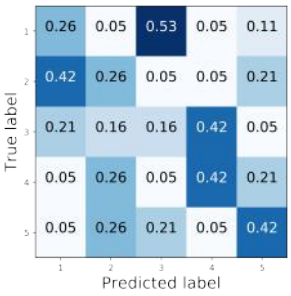
Video	lower	higher
1	28.7%	34.2%
2	66.7%	45.8%
3	61.9%	64.3%
4	52.4%	62.1%
5	42.9%	43.8%
6	47.6%	27.0%
7	57.1%	56.8%
8	71.4%	73.5%

(a)

(b)



(c)



(d)

Table 2. (a) Video user study - Success rate in user identification of a real video from a modified one for both lower-resolution and higher-resolution models. Closer to 50% is better. (b) Each column is a different individual from the still image user study. [Row 1] The gallery images, i.e, the album images the users were asked to select the identity from. [Row 2] The input images. [Row 3] The de-identified version of [Row 2]. (c) The confusion matrix in identifying the five persons for the real images (control). (d) The confusion matrix for identifying, based on the de-identified images.

Person in	Method	RGB values	Face desc.
Row 1	[41]	5.46	1.21
	Our high	2.72	1.50
Row 2	[41]	4.91	1.35
	Our high	2.35	1.53
Row 3	[41]	4.51	1.20
	Our high	3.92	1.32

Table 3. The distance between the original and de-identified image, for the images in Fig. 7. Our method results in lower pixel differences but with face descriptor distances that are higher.

4. Experiments

Training is performed using the Adam [22] optimizer, with the learning rate set to 10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.99$. At each training iteration, a batch of 32 images for the lower resolution model, and 64 for the higher resolution model, are randomly selected and augmented. We initialize all convolutional weights using a random normal distribution, with a mean of 0 and a standard deviation of 0.02. Bias weights are not used. The decoder includes LReLU activations with $\alpha = 0.2$ for residual blocks and $\alpha = 0.1$ otherwise. The low-resolution network was trained on a union of LFW [13], CelebA [28] and PubFig [24], totaling 260,000 images, the vast majority from CelebA. The identity information is not used during training. The high-resolution network was train on a union of CelebA-HQ [17], and faces extracted out of the 1,000 source videos used by [39], resulting in 500,000 images. Training was more involved for the lower resolution model, and it was trained for 230k iterations with a gradual increasing strength of the hyperparameter λ, ranging from $\lambda = 1 \cdot 10^{-7}$ to $\lambda = 2 \cdot 10^{-6}$, in four steps. Without this gradual increase, the naturalness of the generated face is diminished. For the higher resolution model, 80k iterations with a fixed $\lambda = 2 \cdot 10^{-6}$ were sufficient.

Sample results are shown in Fig. 3. In each column, we show the original frame, the modified (output) frame, and the target image from which the identity was extracted. As can be seen, our method produces natural looking images that match the input frame. Identity is indeed modified, while the other aspects of the frame are maintained.

The supplementary media (https://youtu.be/cCYnBtni7Wg) contains sample videos, with significant motion, pose, expression and illumination changes, to which our method was applied. Our method can deal with videos, without causing motion- or instability-based distortions. This is despite being strictly based on per-frame analysis.

It is also evident that the lower resolution model seems blurry at times. This is a consequence of the fixed resolution and not of the generated image, which is in fact sharp. The higher resolution model clearly provides more pleasing results, when the required resolution is high.

To test the naturalness of the approach, we tested the ability of humans to discriminate between videos that were modified to those that were not. Although the human observers ($n = 20$) were fully aware of the type of manipulation that the videos had undergone, the human performance was close to random, with an average success rate of 53.6% (SD=13.0%), see Tab. 2(a). In order to avoid a decision based on a familiar face, this was evaluated on a non-celebrity dataset created specifically for this purpose, which contained 8 videos.

Familiar identities, can often be recognized by non-facial cues. To establish that given a similar context around a facial identity (e.g. hair, gender, ethnicity), the perceived identity is shifted in a way that is almost impossible to place, we considered images of five persons of the same ethnicity and similar hair styles from a TV show, and collected two sets of images: reference (gallery) and source. The source

Person	Original frames		Lower-res de-ID model		Higher-res de-ID	
	Median	Mean ±SD	Median	Mean ±SD	Median	Mean ±SD
Simone Biles	1	3 ±50	1730	2400.6±2142	1725	2223±1814
Billy Corgan	1	95.6±313	3156	3456.3±2601	901	1334±1518
Selena Gomez	1	1 ±0	2256	2704±1873	8058	8110±2186
Scarlett Johansson	1	3.8±38.6	9012	7753.5±3112	4493	4830±2544
Steven Yeun	1	1.02±0.6	5806	4976.2±3167	1069	1814±2544
Sarah J. Parker	1	1±0	679	1069.3±1096	408	620±665
Average	1	17	3773	3726	2776	3155

Table 4. Ranking of the true identity out of a dataset of 54,000 persons (SD=Standard Deviation). Evaluation is performed on the pre-trained LResNet50E-IR ArcFace network. Results are given for both the lower- and higher-resolution models.

images were modified by our method, using them as targets as well, see Tab. 2(b). As can be seen in the confusion matrix of Tab. 2(c), the users could easily identify the correct gallery images, based on the source images. However, as Tab. 2(d) indicates, post de-identification, the answers had little correlation with the true identity, as desired.

In order to automatically quantify the performance of our de-identification method, we applied a state-of-the-art face-recognition network, namely, the ArcFace [7] LResNet50E-IR network. This network was selected both for its performance, and for the dissimilarity between this network and the VGGFace2 network, used as part of our network, in both the training set and loss.

The results of the automatic identification are presented in Tab. 4 for both the lower resolution and the higher resolution models. Identification is performed out of the 54,000 persons in the ArcFace verification set. The table reports the rank of the true person out of all persons, when sorting the softmax probabilities that the face recognition network produces. The ranking of the true identity in the original video shows an excellent recognition capability, with most of the frames identifying the correct person as the top-1 result. For the de-identified frames, despite the large similarity between the original and the modified frames (Fig. 3), the rank is typically in the thousands.

Another automatic face recognition experiment is conducted on the LFW benchmark [13]. Tab. 5 presents the results on de-identified LFW image pairs for a given person (de-identification was applied to the second image of each pair), for two FaceNet [42] models. The true positive rate for the LFW benchmark drops from almost 0.99, to less than 0.04 after applying de-identification.

An additional experiment, evaluating our method on the LFW benchmark, can be found in the appendix.

A comparison of our method with the recent work of [31] is given in Fig. 4. This method relies on the generation of a new identity, given the k-closest identities, as selected by a trained CNN feature-extractor. As can be seen, this can result in the same rendered identities for multiple inputs, and does not maintain the expression, illumination and skin tone.

FaceNet Model	Original	De-ID
VGGFace2	0.986 ± 0.010	$\mathbf{0.038} \pm 0.015$
CASIA	0.965 ± 0.016	$\mathbf{0.035} \pm 0.011$

Table 5. Results on the LFW benchmark, employing the FaceNet network trained on VGGFace2 or CASIA-WebFace. Shown is the True Positive Rate for a False Acceptance Rate of 0.001.

To emphasize the ability of identity-distancing, while maintaining pixel-space similarity, we compare our method to [41]. While the method of [41] relies on finding a dissimilar identity within a given dataset, ours is single-image dependent, in the sense that it does not rely on other images within a dataset. It is, therefore, resilient to different poses, expressions, lighting conditions and face structures. Given the figures provided in the work of [41], we compare our generated outputs by high-level perceptual distance from the source face, taking into account pixel-level similarity (Fig. 7). A comparison of the distance between the original and the de-identified image for the two methods (Tab. 3) reveals that our method results in lower pixel differences, yet with face descriptor distances that are higher.

A comparison with the work of [49] is given in Fig. 8. Our results are at least as good as the original ones, despite having to run on the cropped faces extracted from the paper PDF. Although [49] presents visually pleasing results, they do not maintain low-level and medium-level features, including mouth expression and facial hair. In addition, the work of [49] presents results on low-resolution black and white images only, with no pose or gender variation.

Figure 5 compares with the recent work of [43, 44]. Our method is able to distance the identity in a more subtle way, while introducing less artifact. Our generated image contains only the face, which is enabled by the use of the mask. Their

(a) (b) (c)

Figure 4. (a) Input images from [31], (b) our results, (c) those of [31]. Our method maintains the expression, pose, and illumination. Furthermore, our work does not assign the same new identity to different persons.

(a) (b) (c)

Figure 5. (a) Input images from [43, 44], (b) our results, (c) those of [43] (row 1) and [44] (rows 2-3).

method generates both the face and the upper body using the same 256×256 generation resolution, which makes our results of a much higher effective resolution. A full set of results is given in the appendix, Fig. 14.

To further demonstrate the robustness of our method, we applied our technique to images copied directly from the very difficult inputs of [36]. As can be seen in Fig. 6, our method is robust to very challenging illuminations.

To demonstrate the control of the hyperparameter λ over the identity distance, we provide a sequence of generated images, where each trained model is identical, apart from the strength of λ. The incremental shift in identity can be seen in Fig. 9. Ablation analyses are given in the appendix. The analyses compare various variants of our method, and depict the artifacts introduced by removing parts of it.

5. Conclusions

Recent world events concerning the advances in, and abuse of face recognition technology invoke the need to understand methods that successfully deal with de-identification. Our contribution is the only one suitable for video, including live video, and presents quality that far surpasses the literature methods. The approach is both elegant and markedly novel, employing an existing face descriptor concatenated to the embedding space, a learned mask for blending, a new type of perceptual loss for getting the desired effect, among a few other contributions.

Minimally changing the image is important for the method to be video-capable, and is also an important factor in the creation of adversarial examples [35]. Unlike adversarial examples, in our work, this change is measured using low- and mid-level features and not using norms on the pixels themselves. It was recently shown that image perturbations caused by adversarial examples distort mid-level features [50], which we constrain to remain unchanged.

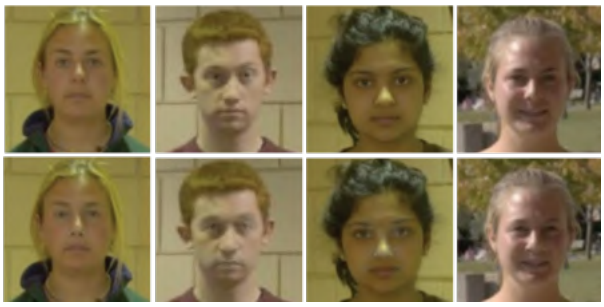
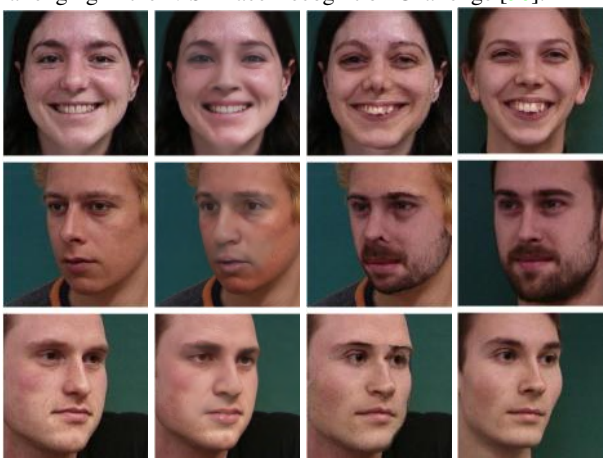


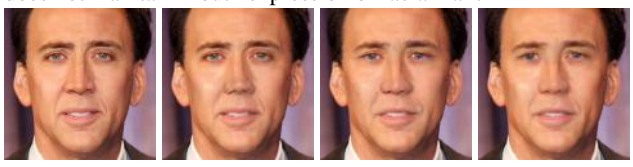
Figure 6. De-Identification applied to the examples labeled as very challenging in the NIST Face Recognition Challenge [36].



(a) (b) (c) (d)

Figure 7. Comparison with [41] (from the paper sample image). (a) Original image (also used for the target of our method). (b) Our generated output. (c) Result of [41]. (d) Target used by [41].

Figure 8. Comparison with [49]. Row 1 - Original images. Row 2 - results of [49]. Row 3 - Our generated outputs. The previous work does not maintain mouth expression or facial hair.



(a) (b) (c) (d)

Figure 9. Incrementally growing λ in the lower resolution model. A gradual identity shift can be observed. (a) Source. (b) $\lambda = -5 \cdot 10^{-7}$. (c) $\lambda = -1 \cdot 10^{-6}$. (d) $\lambda = -2 \cdot 10^{-6}$.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *NIPS*, 2017. 2

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face swapping: Automatically replacing faces in photographs. In *SIGGRAPH*, 2008. 2

[3] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004. 2

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. 4

[5] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*. 2016. 3

[6] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 4

[7] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 7

[8] Faceswap. Github project, https://github.com/deepfakes/faceswap. 2017. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014. 2

[10] Ralph Gross, Latanya Sweeney, Fernando De La Torre, and Simon Baker. Semi-supervised learning of multi-factor models for face de-identification. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[13] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report. 6, 7

[14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3, 5

[16] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face deidentification. In *In ICB*, 2015. 2

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2, 6

[18] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 4

[19] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Trans. Graph.*, 35(4), 2016. 2

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 2

[21] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. 4

[22] Kingma, Diederik P., and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2016. 6

[23] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *The IEEE International Conference on Computer Vision*, 2017. 2, 3

[24] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *CVPR*, 2009. 6

[25] Guillaume Lample et al. Fader networks: Manipulating images by sliding attributes. In *NIPS*, 2017. 3

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*. 2017. 2

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[29] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 3

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 4

[31] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 11(9):1046–1054, 2017. 2, 7, 8

[32] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. 2

[33] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. 2

[34] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. *arXiv preprint arXiv:1704.06729*, 2017. 2

[35] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017. 8

[36] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O'Toole, David S Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Automatic Face & Gesture Recognition*, 2011. 8

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[39] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018. 6

[40] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 2

[41] Branko Samarzija and Slobodan Ribaric. An approach to the de-identification of faces in different poses. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1246–1251. IEEE, 2014. 2, 6, 7, 8

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 7

[43] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. 2, 7, 8

[44] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 553–569, 2018. 2, 7, 8, 12, 13

[45] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[46] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 3

[47] Dmitry Ulyanov, Vadim Lebedev, Victor Lempitsky, et al. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016. 3

[48] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[49] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1806.08906*, 2018. 2, 7, 8

[50] Cihang Xie et al. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018. 8

[51] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017. 2

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3, 4

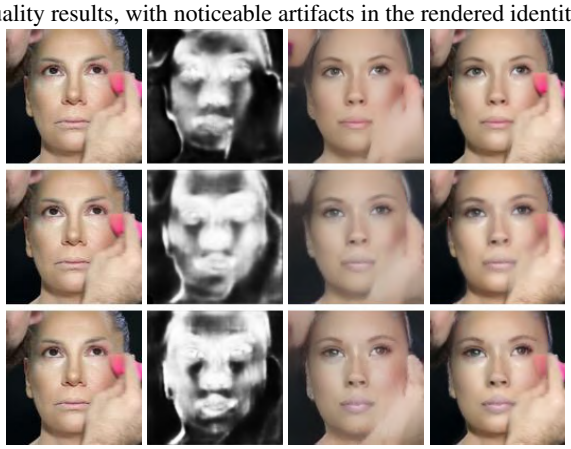
A. Ablation Analyses

A.1. General Ablation Analysis

An ablation analysis is shown in Fig. 12. In order to efficiently run multiple-models, it is done with the low-res architecture. The various options include: a no-mask option, a partial adversarial loss that applies only to the masked output and not to the raw output, training without the gradual increase of λ, and an attempt to incorporate an additional output with a lower resolution to be taken into account, as part of the compound loss. All of these ablation experiments were conducted on the lower-resolution model.

The following description of methods and the associated artifacts correspond to the columns of Fig. 12: (c) No mask. *Bad face edge, glasses occlusion handled poorly.* (d) Adversarial loss on masked output only. *Various artifacts, e.g., around the right eye, one can also observe green stripes near the mouth.* (e) No gradual increment of λ. *Collapse into unnatural blurred face*. (f) Lower resolution output added for the compound loss. *Weak de-id, checkerboard pattern near the center of the face, and when handling occlusions.* (g) Weak λ, adversarial loss on masked output only. *Weak de-identification, artifacts near the eyes and eyebrows*.

Figure 10. No face-descriptor ablation study. Source (row 1), our model (row 2), and no face-descriptor (row 3), resulting in lower quality results, with noticeable artifacts in the rendered identity.



(a) (b) (c) (d)

Figure 11. Mask regularization ablation study and mask outputs. (a) Source, (b) mask, (c) raw output, (d) masked output. Compared to our model (row 1), the effects of not minimizing the mask norm ($\alpha_4 = 0$, row 2) can be observed as occlusions (hand, pink element, upper-left face) not handled well, and excessive face regions taken from the rendered image, resulting in distortions. No mask derivative regularization ($\alpha_5 = 0$, row 3) effects can be seen as high-frequency patterns generated by the mask and output frame.

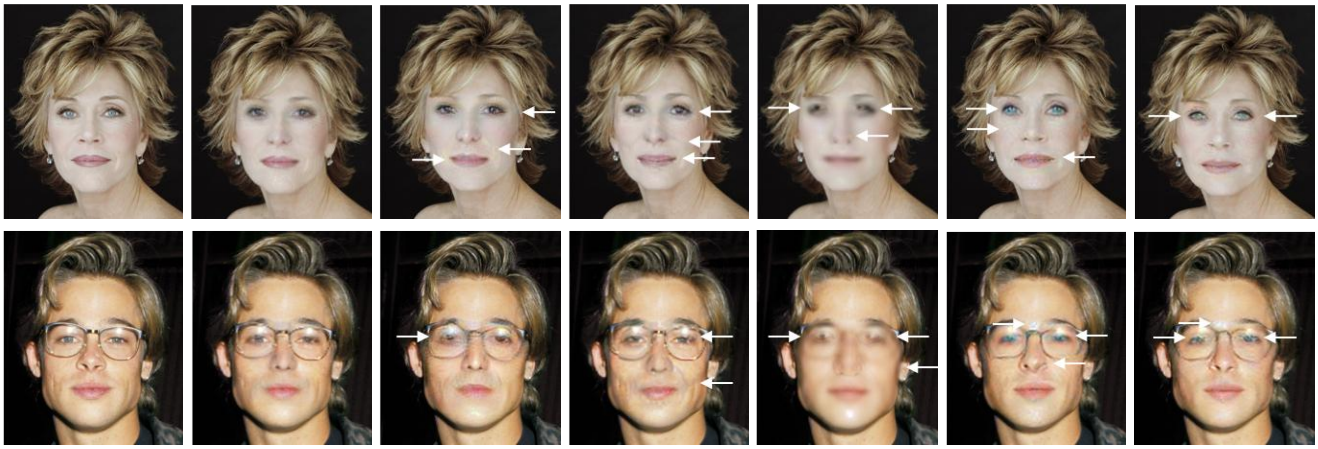
A numerical analysis plot of the ablation study with the same options is provided in Fig. 13. Each method is evaluated along two axes of comparison between the input image and the output image: on the x-axis we show the difference in appearance as measured by the L1 norm between the images; the y-axis shows the difference in ID, as computed by the L1 norm between the VGGFace2 representation of the two images. The plot shows mean results obtained for our method (marked (b) to match the columns of Fig. 12) and the various ablation methods (marked (c)–(g)). As can be seen, our method maintains image similarity and also has a difference in ID that is similar or larger than any other method, with the exception of the method marked as (c). This is expected, since this variant is the mask-less one, which does not blend-in the original image. Variant (f) is considerably more similar to the original image on both axes, since the de-ID performed is very weak with this variant.

A.2. Face-Descriptor Ablation Analysis

A face-descriptor-specific ablation analysis is provided to emphasize its necessity in Fig. 10. The face-descriptor is highly motivating for the decoder to use, otherwise, minimizing the high-level perceptual loss ($l_{1\times1}$) would be more challenging, as can be seen in Fig. 10. For each source image (row 1), our model result (row 2) can be seen to produce higher-quality results with less artifacts, compared to the model that lacks a face-descriptor concatenated to the latent space (row 3). In the results of the third row, the face descriptor is not concatenated to the z embedding, but still used in the perceptual loss.

A.3. Mask Regularization Ablation Analysis

The mask regularization parameters $\alpha_{4,5}$ importance can be observed in Fig. 11. They assist in dealing with occlusions, and handling irrelevant regions, that can be taken from the source image, rather than generated (e.g. regions that are not related to the generated face, teeth, etc.). α_4 keeps the mask minimal, i.e. blending maximal regions from the source image, rather than the generated one. By avoiding excessive blending of generated regions, less artifacts are apparent on the final output (as observed in row 2). α_4 keeps the mask smooth, by penalizing mask derivatives. This can be seen to reduce high-frequency patterns, (as observed in row 3).



(a) (b) (c) (d) (e) (f) (g)

Figure 12. An ablation study. (a) Source image. (b) Our result. (c)–(g) variants, see text for details.

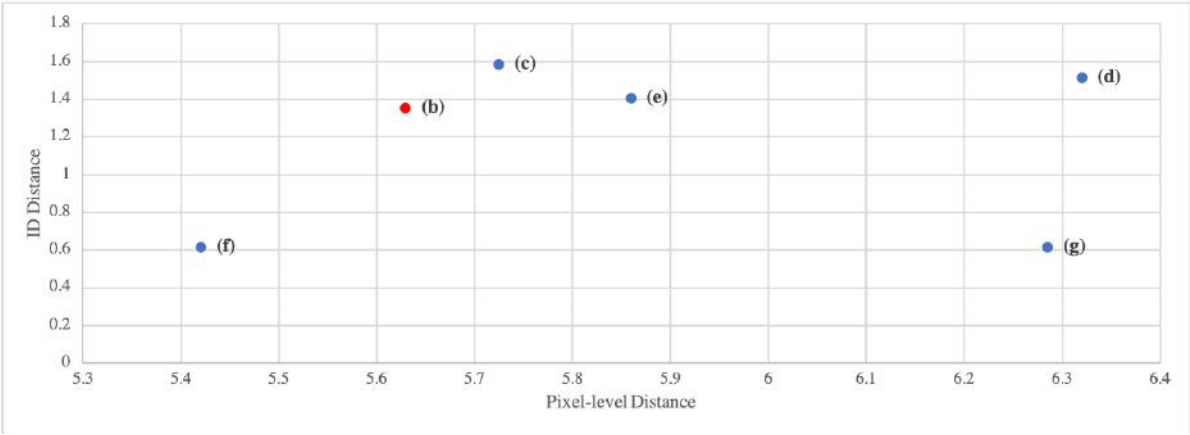
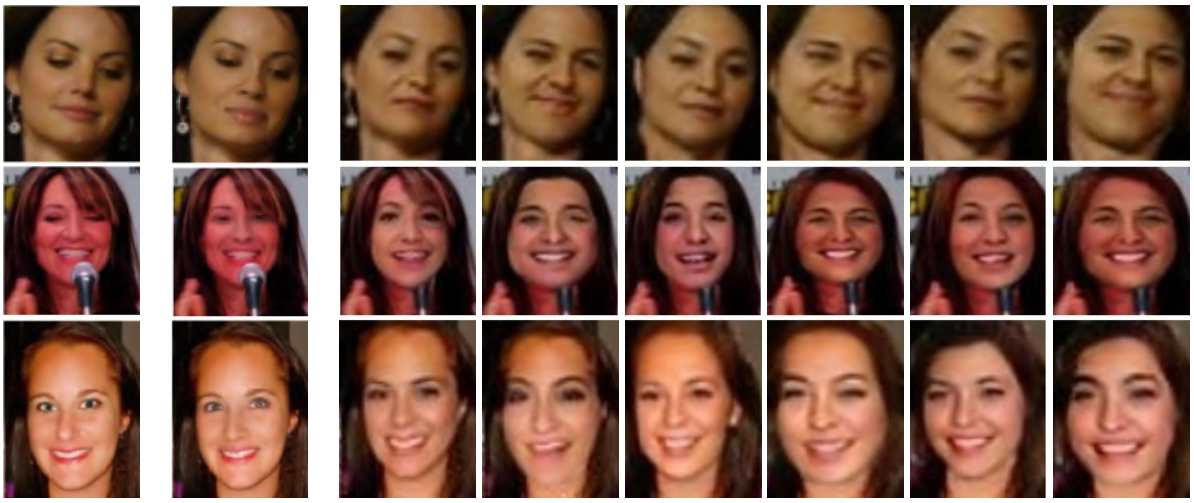


Figure 13. The mean pixel-level distance vs. the mean ID distance. The first should be low, while the second should be high. Shown are the methods of each column in Fig.12(b)–(g).

B. Additional Comparison with Previous Methods

We provide an extensive comparison with the work of [44]. In the paper we only included one of [44] generated outputs: the sample shown was the first output that gains <50% of recognition rates by an automatic face recognition algorithm, according to [44]. The work of [44] provides several models for different levels of de-identification. In Fig. 14, we present all faces from [44]. The reported recognition rate itself is given in Tab. 6.

As can be seen in the results of [44], the less recognizable the identity is, the less natural the face is. Note that: (1) our model provides for much stronger de-identification results, with the rank typically in the thousands, out of a dataset of 54,000 persons, as reported in the experiments section. (2) all models of the baseline method produce low resolution outputs (64×64) compared to our model's much higher resolution (256×256).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 14. A full set of results for the comparison with the work of [44]. (a) Source image, (b) our generated output, (c-h) generated outputs for [44] of different models. The work of [44] provides several models that provide for different levels of de-identification. As can be seen, models that gain a rate of $< 50\%$ of head obfuscation effectiveness by machine recognizers, provide less natural faces.

Row | Column	(c)	(d)	(e)	(f)	(g)	(h)
Row 1	70.8%	47.6%	36.6%	18.0%	22.5%	7.1%
Row 2	59.9%	26.3%	25.8%	12.7%	15.7%	7.2%
Row 3	59.9%	26.3%	25.8%	12.7%	15.7%	7.2%

Table 6. Head obfuscation effectiveness for [44]: recognition rates of machine recognizers (lower is better), as provided by [44]