LIVE FACE DE-IDENTIFICATION IN VIDEO

Anonymous authors
Paper under double-blind review

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 a 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.

The problem of video de-identification is a challenging computational 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. These factors include pose and expression, occlusion, illumination and shadow, and their dynamics.

In order to tackle these challenges, we introduce a novel encoder-decoder architecture. To the latent space, we concatenate the activations of the representation layer of a network trained to perform face recognition. The loss terms separate between low- and mid-level perceptual terms and high-level terms. The former are used to tie the output image to the input video frame, while the latter is used to obtain the desired modification. Additional losses include reconstruction losses, edge losses, and an adversarial loss. The network outputs both an image and a mask, which are used, in tandem, to reconstruct the output frame.

In contrast to the literature methods, which are limited to still images and often swap the image's 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 those videos has a similar appearance to the original person. 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.

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 (Goodfellow et al.,

Table 1: A comparison to literature de-identification methods. †The face is swapped with an average of a few dataset faces.

	(Newton, '05)	(Gross, '08)	(Samarzija, '14)	(Jourabloo, '15)	(Wu, '18)	Our
Preserves expression	-	-	-	-	-	+
Preserves pose	-	+	+	-	-	+
Generates new faces	-	†	-	†	+	+
Demonstrated on video	-	-	-	-	-	+
Demonstrated on a diverse dataset (gender, ethnicity, age, etc.)	-	+	-	+	-	+

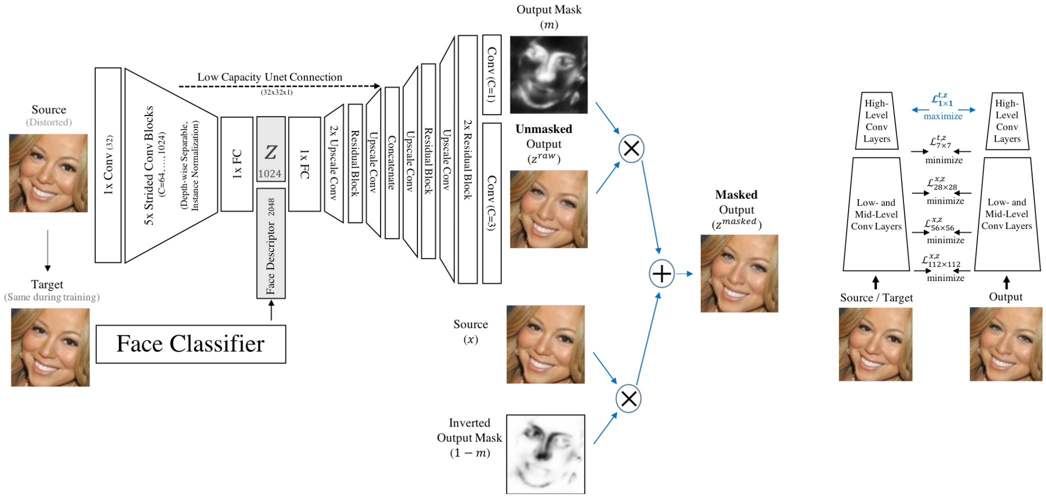
2014; Radford et al., 2015; Salimans et al., 2016) since their inception. Recently, high resolution natural looking faces were generated by training both the generator and the discriminator of the GAN progressively, starting with shallower networks and lower resolutions and enlarging these gradually (Karras et al., 2018).

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 (Kim et al., 2017; Yi et al., 2017; Benaim & Wolf, 2017; Liu et al., 2017). 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 (Taigman et al., 2017). 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 (Newton et al., 2005b; Gross et al., 2008), while more recent works rely on techniques for selecting distant faces (Samarzija & Ribaric, 2014) or averaging/fusing faces from pre-existing datasets Newton et al. (2005a); Jourabloo et al. (2015). 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 in the work of Wu et al. (2018). However, the experiments were restricted to a homogeneous dataset, with no apparent expression preservation within the results. See Tab. 1 for 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 (Blanz et al., 2004; Bitouk et al., 2008). 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 (Kemelmacher-Shlizerman, 2016; Nirkin et al., 2017). While these classical face swapping methods work in the pixel space and copy the expression of the target image, recent deep-learning based work swap the identity, while maintaining the other aspects of the source image (Korshunova et al., 2017). In comparison to our work, (Korshunova et al., 2017) 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 patches from the target person. Moreover, (Korshunova et al., 2017) is limited to stills and was not demonstrated on video.

The face swapping (FS) project (Faceswap, 2017) 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 input to the FS system, during training, is two large sets of images, one from each identity. Typically, in order to obtain good results, thousands of images from each individual with a significant variability in pose, expression, and illumination are used. In many cases, a large subset of the images of the source person are taken from the video



(a)

(b)

Figure 1: (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.

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 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 notably is the way in which augmentation is performed in order to train a more semantic encoder-decoder network. During 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 (Zhang et al., 2017). 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 (Johnson et al., 2016; Ulyanov et al., 2016a) that is based on the layers of a face-recognition network.

3 Method

Our architecture is based on an adversarial autoencoder (Makhzani et al., 2015), 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 by He et al. (2016b), trained over the VGGFace2 dataset of Cao et al. (2017).

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 a face detector, we employ dlib (King, 2009). (b) Multiple facial points are detected. In our implementation we use the 68 points provided by dlib. (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.

Training is performed on image datasets and not on video. 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 (Zhang et al., 2017) 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 that we do 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 actual application on unconstrained videos.

3.1 Network architecture

The architecture is illustrated in Fig. 1(a). The encoder is composed of a convolutional layer, followed by five strided, depth-wise separable (Chollet, 2017) convolutions with instance normalization (Ulyanov et al., 2016b). 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 (He et al., 2016a) 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 than the input channel size. Following an instance normalization and a LReLU (He et al., 2015) activation, the activations are re-ordered such 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 Unet connection (Ronneberger et al., 2015) 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.

3.2 Training and the Losses Used

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

$$\begin{aligned}\mathcal{L} =&\alpha_0\mathcal{L}_G + \alpha_1\mathcal{L}_R^{raw} + \alpha_1\mathcal{L}_R^{masked} + \alpha_2\mathcal{L}_x^{raw} + \alpha_2\mathcal{L}_y^{raw} + \alpha_2\mathcal{L}_x^{masked} + \alpha_2\mathcal{L}_y^{masked} \\ &+ \alpha_3\mathcal{L}_p^{raw} + \alpha_3\mathcal{L}_p^{masked} + \alpha_4\mathcal{L}^m + \alpha_5\mathcal{L}_x^m + \alpha_5\mathcal{L}_y^m, \qquad (1)\end{aligned}$$

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 perceptual loss terms are the only terms that differ between the two applications. 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, adversarial loss is imposed with a convex combination of example pairs (known as mixup) (Zhang et al., 2017) over a Least Square GAN (Mao et al., 2017) loss:

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

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:

$$\mathcal{L}_R^{raw} = \alpha_1\|z^{raw} - x\|_1 \qquad \mathcal{L}_R^{masked} = \alpha_1\|z^{masked} - x\|_1 \tag{3}$$

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:

$$\begin{aligned} \mathcal{L}_x^{raw} &= \alpha_2\|z'_x - x'_x\|_1 \qquad & \mathcal{L}_y^{raw} &= \alpha_2\|z'_y - x'_y\|_1 \\ \mathcal{L}_x^{masked} &= \alpha_2\|z'_x - x'_x\|_1 \qquad & \mathcal{L}_y^{masked} &= \alpha_2\|z'_y - x'_y\|_1 \end{aligned} \tag{4}$$

where x'_x is the derivative of the undistorted input image x along the x 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:

$$\mathcal{L}^m = \|m\|_1 \qquad \mathcal{L}_x^m = \|m'_x\|_1 \qquad \mathcal{L}_y^m = \|m'_y\|_1 \tag{5}$$

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 (Johnson et al., 2016) 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 constraining the high abstraction layer perceptual loss between the target and generated output.

Let $a^r_{n\times n}$ 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 application dependent 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 $\ell^{r_1,r_2}_{n\times n} = c_n\|a_{r_1,n\times n} - a_{r_2,n\times n}\|_1$, where c_n is a normalizing constant, corresponding to the size of the spatial activation map.

The perceptual loss is given by:

$$\mathcal{L}_p^c = \ell^{x,z^c}_{112\times112} + \ell^{x,z^c}_{56\times56} + \ell^{x,z^c}_{28\times28} + \ell^{t,z^c}_{7\times7} - \lambda\ell^{t,z}_{1\times1} \tag{6}$$

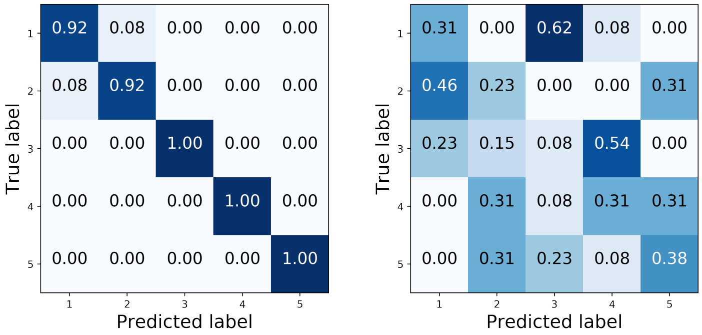
for c that is either raw or $masked$, and where $\lambda > 0$ is a hyperparameter, which determines the distance of the generated face's high level features from those of the target image.

The application of the multi-image perceptual loss during training is depicted in Fig. 1(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, as can be seen in Fig. 2. Note that we call it multi-image perceptual loss, since 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, similarly to other perceptual losses.

Table 2: User study. (a) Success rate in user identification of a real video from a modified one. Closer to 50% is better. (b) The confusion matrix in identifying the five persons for the real images (control). (c) The confusion matrix for identifying based on the de-identified images.

Video	Success rate
1	28.7%
2	66.7%
3	61.9%
4	52.4%
5	42.9%
6	47.6%
7	57.1%
8	71.4%
Average	53.6% ± 13.0

(a)



True label \ Predicted label	1	2	3	4	5
1	0.92	0.08	0.00	0.00	0.00
2	0.08	0.92	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	1.00

(b)

True label \ Predicted label	1	2	3	4	5
1	0.31	0.00	0.62	0.08	0.00
2	0.46	0.23	0.00	0.00	0.31
3	0.23	0.15	0.08	0.54	0.00
4	0.00	0.31	0.08	0.31	0.31
5	0.00	0.31	0.23	0.08	0.38

(c)

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.

4 Experiments

Training is done using the Adam (Kingma & Ba, 2016) 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 is 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 network was trained on a union of LFW (Huang et al.), CelebA (Liu et al., 2015) and PubFig (Kumar et al., 2009), totaling 260,000 images, the vast majority from CelebA. The identity information is not used during training. The model 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.

Sample results are shown in Fig. 3. In each pair of frames, we show the original frame, the target frame from which the identity was extracted, and the modified (output) frame. 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 available at `anonymous-deid-iclr2019submission.github.io` contains sample videos, with significant motion, pose, expression and illumination changes, to which our method was applied. It is evident that the method can deal with videos, without causing motion- or instability-based distortions. This is despite being strictly based on per-frame analysis.

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. The human observers ($n = 20$) were fully aware of the type of manipulation that the images had undergone. Still, the human performance is 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 10 videos (samples are attached as supplementary).

Another user study tested how identifiable the resulting images were. We considered images of five persons from a TV show and collected two sets of images: gallery and source. The source images were modified by our method using the same images also as target. As can be seen in the confusion matrix of Tab. 2(b) the users had no problem identifying the correct gallery image based on the source images. However, as Tab. 2(c) shows, post de-identification the answers had little correlation with the true identity, as desired.

In order to automatically quantify the performance of de-identification, we applied state-of-the-art face-recognition networks. Namely, the LResNet34E-IR and LResNet50E-IR networks of ArcFace (Deng et al., 2018). These networks were selected both for their performance and for the difference between these networks and the VGGFace 2 network, used as part of our network, in both training set and loss.

The results of the automatic identification are presented in Tab. 3. 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.

To emphasize the ability of identity-distancing, while maintaining pixel-space similarity, we compare our method to the work of Samarzija & Ribaric (2014). While that method 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 by Samarzija & Ribaric (2014), we compare our generated outputs (Fig. 4) by high-level perceptual distance from the source face, taking into account pixel-level similarity (Fig. 4). A comparison of the distance between the original and de-identified image for the two methods (Fig. 4(e)) reveals that our method results in lower pixel differences but with face descriptor distances that are as high.

A comparison with the work of Wu et al. (2018) is given in Fig. 5. Our results are at least as good as the literature ones, despite us having to run on the cropped faces extracted from the paper's PDF. Although Wu et al. (2018) presents visually pleasing results, unlike our work, they do not maintain low-level and medium-level features, including mouth expression and facial hair. Note that this previous work presents results on low-resolution black and white images only, with no pose variance.

To further demonstrate the robustness of our methods, we applied the technique to the images of the very difficult inputs from (Phillips et al., 2011), as copied directly from the sample figure there. As can be seen in Fig. 6, our method can deal with 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. 7.

An ablation analysis is shown in Fig. 8. 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.

A numerical analysis of the ablation analysis with the same options is given in Fig. 9. Each method is evaluated along two axis 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 plots show mean results obtained for our method (marked (b) to match Fig. 8) and the various ablation methods (marked (c)–(g)). As can be seen, our method maintains image similarity and also has a larger difference in ID 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 axis, since the de-ID performed is very weak with this variant.

For many applications, one would require a higher resolution face recognition. The supplementary video, for example, might look blurry at times, despite the share output of the generator due to the limited resolution. In order to overcome it, we trained a high-resolution (256x256) model that is obtained similarly with the following distinctions: (a) The decoder architecture is simplified and enlarged to be a lattice of 6x(Upscale block –> Residual block), (b) The batch size is set to 64, and (c) The model is trained for 80k iterations.

Training the second model, the resolution is improved, as can be seen in the supplementary video, while the de-identification affect remains as large, as can be seen in Tab. 4.

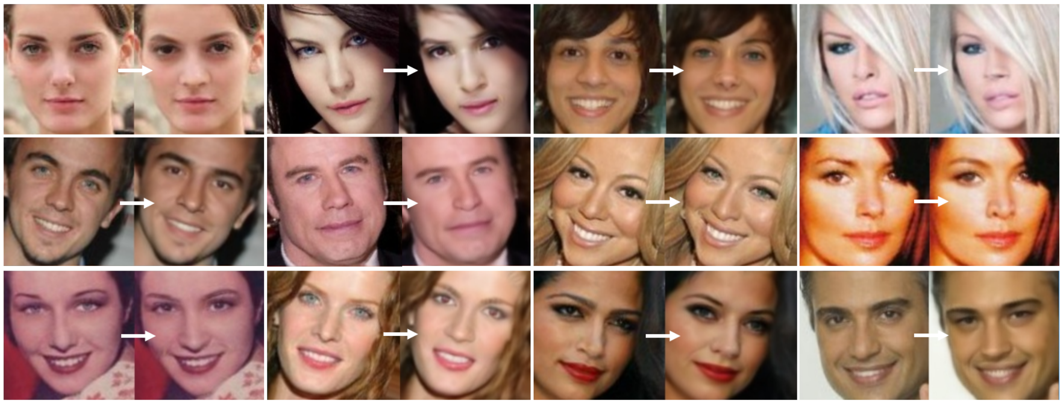


Figure 2: Training the de-identification network. Shown, in each pair, are the source (left) and output (right) images. During training (but not during test), the target image used is the same as the source one and the output maintains the low-level features of the source and distances the high-level features from it.

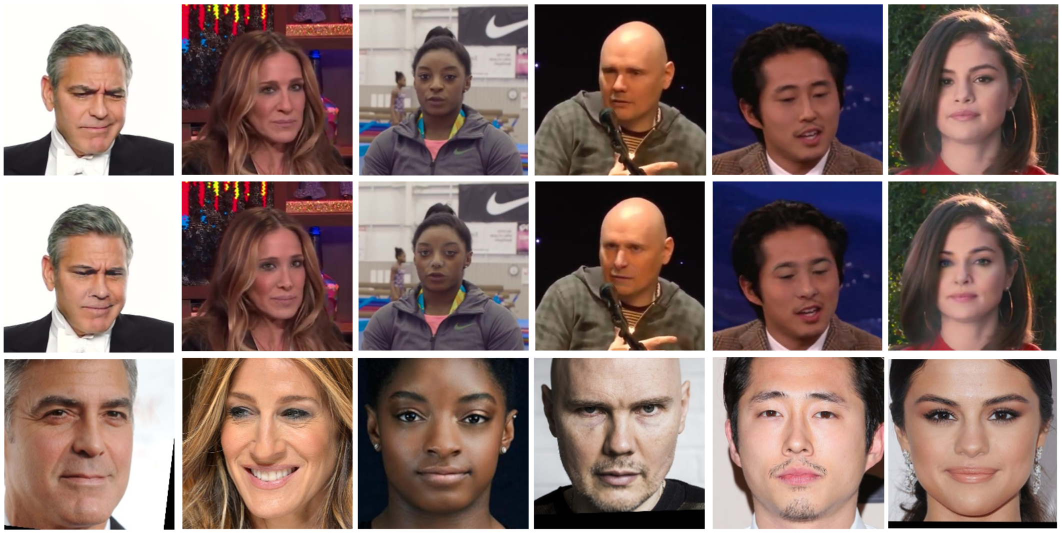
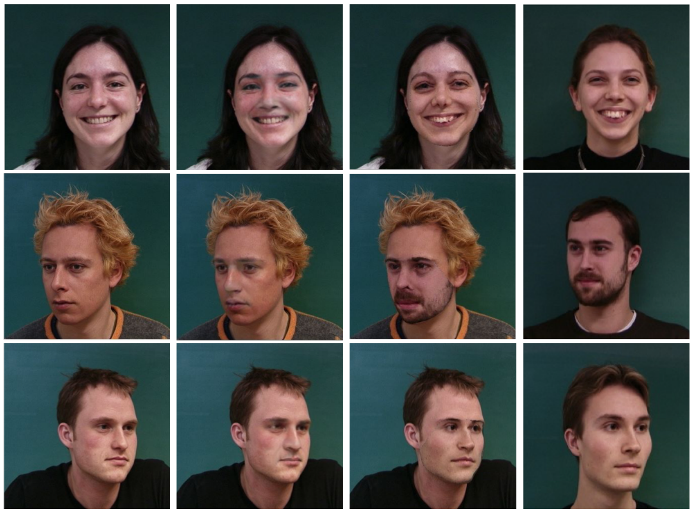


Figure 3: Sample results for de-identification (zoom). Triplets of source frame, converted frame and target are shown. The modified frame looks similar but the identity is completely different.

5 Conclusions

The appearance of a face is hard to model. For example, computer graphics animations avoid realistic faces in order to stay out of the uncanny valley. The recent GAN technology can create realistic high resolution (frontal) facial images, given a latent space vector z. However, the domain of faces is not completely covered: the recent generators suffer from both mode collapse and mode canceling, and for many face images I, one cannot find a latent vector z, such that a modern generator G would generate $G(z) = I$.

In this work, we employ an encoder-decoder architecture, which avoids the mode canceling pitfalls. However, unlike latent-space generators, not every mid-network activation leads to the generation of a valid face downstream. Despite this shortcoming, we show that an encoder-decoder architecture is flexible enough to support multiple conditional generation tasks.



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

	Distance	
Method	RGB values	Face desc.
Person in row 1		
Previous work	4.87	1.23
Our	3.18	1.40
Person in row 2		
Previous work	4.56	1.34
Our	1.95	1.34
Person in row 3		
Previous work	3.86	1.17
Our	1.90	1.20

(e)

Figure 4: Comparison with the work of Samarzija & Ribaric (2014) (taken from that paper's sample image). (a) Original image (also used for the target of our method). (b) Our generated output. (c) Result of (Samarzija & Ribaric, 2014). (d) Target used by Samarzija & Ribaric (2014). (e) A comparison of the distance between the original and de-identified image for the two methods. Our method results in lower pixel differences but with face descriptor distances that are as high or higher.

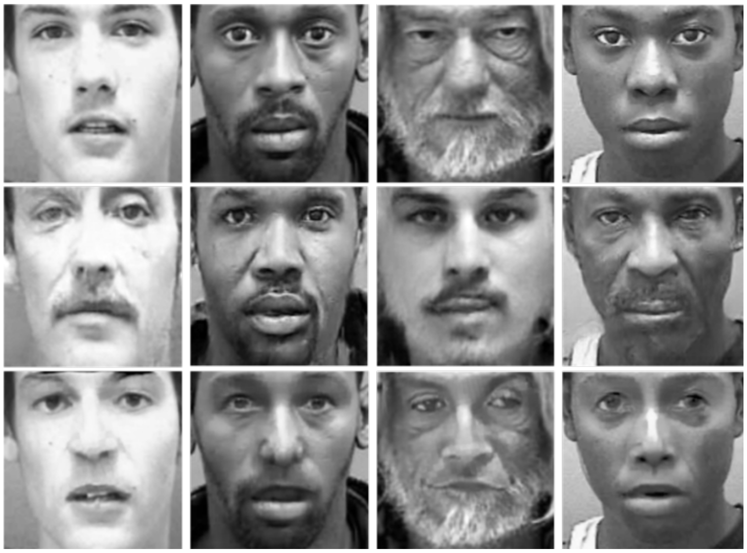


Figure 5: Comparison with Wu et al. (2018) (taken from that paper's sample image). Row 1 - Original images. Row 2 - results of Wu et al. (2018). Row 3 - Our generated outputs. The previous work does not maintain expression, pose, and facial hair.

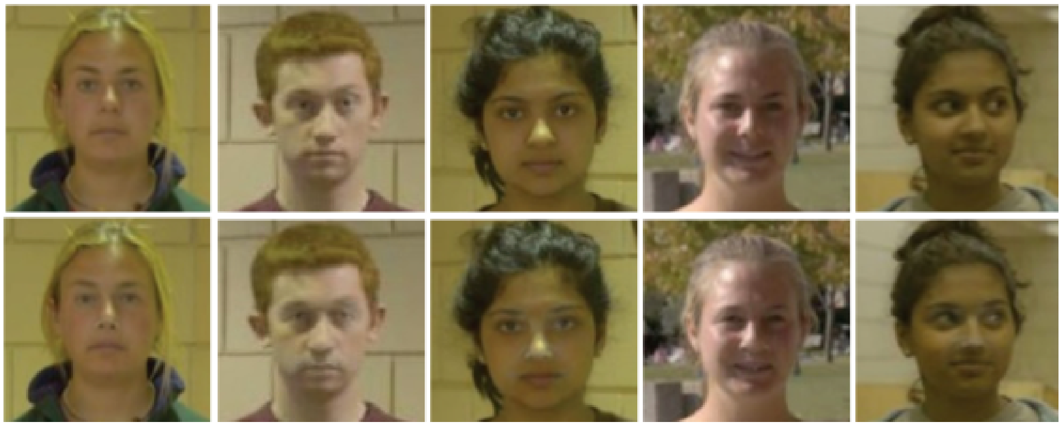


Figure 6: De-Identification applied to the examples labeled as very challenging in the NIST Face Recognition Challenge (Phillips et al., 2011) (taken from that paper's sample image, images are zoomed in).

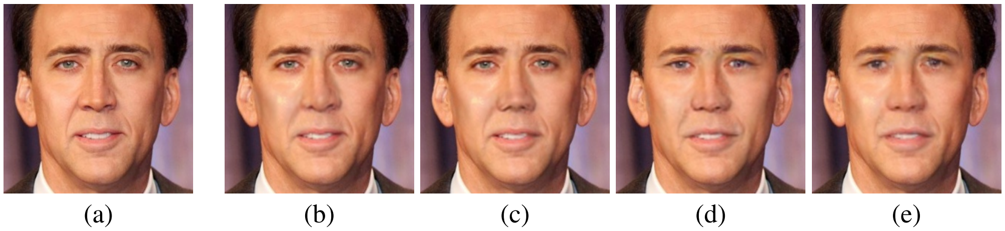


Figure 7: A sequence of images, generated with an incrementally (left-to-right) growing λ. The gradual identity shift can be observed. (a) Source. (b) $\lambda = -2 \cdot 10^{-7}$. (c) $\lambda = -5 \cdot 10^{-7}$. (d) $\lambda = -1 \cdot 10^{-6}$. (e) $\lambda = -2 \cdot 10^{-6}$.

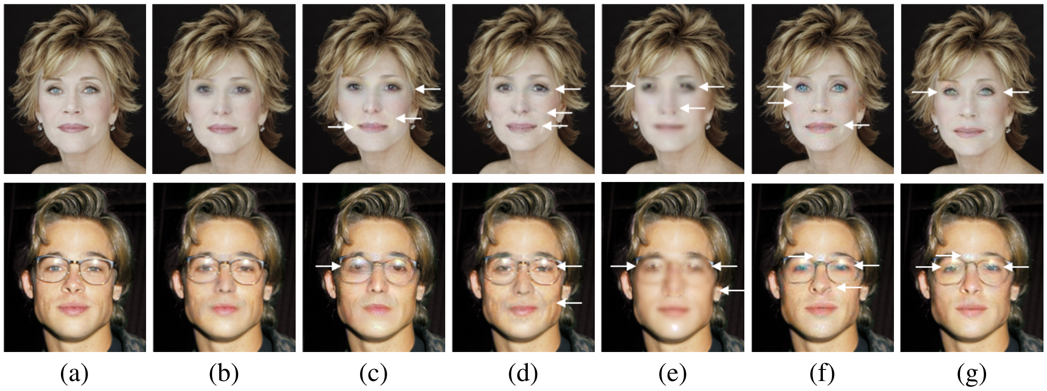


Figure 8: An ablation study. (a) Source image. (b) Our result. (c) No mask. *Bad face edge, artifacts near the mouth, glasses occlusion handled poorly.* (d) Adversarial loss on masked output only. *Artifacts around the right eye, green stripes near the mouth.* (e) No gradual λ strength in training. *Collapse into unnatural blurred face*. (f) Additional lower resolution output for the compound loss. *Weak de-id, checkerboard pattern near the nose, artifacts on the nose, between eyebrows and when handling occlusions.* (g) Weak λ and adversarial loss on masked output only. *Weak de-id, artifacts near the eyes and eyebrows.*

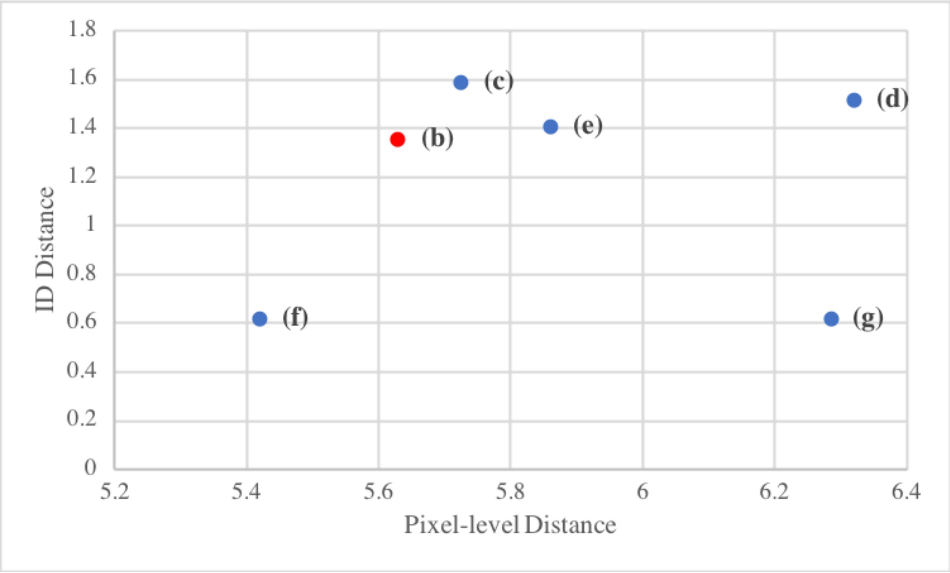


Figure 9: Quantitative ablation study results: for our method – marked as (b) – and the various variants in columns (c)–(g) of Fig. 8 we measure the mean pixel-level distance and the mean ID distance (as evaluated by the differences in the last layer of the VGGFace2 classifier), both in L1. The first should be low, while the second should be high. As can be seen in the results, the model we use (b) is very high in the ID distance, while considerably low in the pixel-level distance. The raw (unmasked) model (c) achieves an even higher ID distance, but this is anticipated, since it is not blended with the source image.

Table 3: Ranking of the true identity out of a dataset of 54,000 persons (SD=Standard Deviation). Evaluation is performed on the pre-trained LResNet34E-IR / LResNet50E-IR ArcFace networks.

	LResNet34E-IR				LResNet50E-IR			
	Original frames		De-Identified frames		Original frames		De-Identified frames	
Person	Median	Mean ±SD	Median	Mean ±SD	Median	Mean ±SD	Median	Mean ±SD
Simone Biles	1	6 ±91	1298	2038.7±2167	1	3 ±50	1730	2400.6±2142
Billy Corgan	1	147.4±615	3891	4112.6±2569	1	95.6±313	3156	3456.3±2601
Selena Gomez	1	1 ±0	2653	3017.7±1873	1	1±0	2256	2704±1873
Scarlett Johansson	1	1.3±4	9191	8146.6±2953	1	3.8±38.6	9012	7753.5±3112
Steven Yeun	1	1.1±1	7923	6115.4±4060	1	1.02±0.6	5806	4976.2±3167
Sarah J. Parker	1	1±0	980	1770.5±1787	1	1±0	679	1069.3±1096
Average	1	26	4322	4200	1	17	3773	3726

Table 4: Ranking results for the high resolution model. Show is the 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.

	Original frames		De-Id 256x256 frames	
Person	Median	Mean ±SD	Median	Mean ±SD
Simone Biles	1	3 ±50	1725	2223±1814
Billy Corgan	1	95.6±313	901	1334±1518
Selena Gomez	1	1 ±0	8058	8110±2186
Scarlett Johansson	1	3.8±38.6	4493	4830±2544
Steven Yeun	1	1.02±0.6	1069	1814±2544
Sarah J. Parker	1	1±0	408	620±665
Average	1	17	2776	3155

REFERENCES

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

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

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

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.

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

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

Faceswap. Github project, https://github.com/deepfakes/faceswap. 2017.

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.

R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Semi-supervised learning of multi-factor models for face de-identification. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

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

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

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

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.

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

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

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

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

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.

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

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2016.

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.

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

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

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

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

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

E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005a.

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, 2005b.

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.

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.

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

O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

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.

B. Samarzija and S. 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)*, 2014.

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

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

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

Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1806.08906*, 2018.

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

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

A De-Identification User Study

We attach the images as they were presented in the second user study. The users were asked to identify either the images on the 2nd row or the images on the 3rd row based on the gallery images in the first row. The users were shown all images at once, were given unlimited time, and were asked to perform the task as accurately as they could.

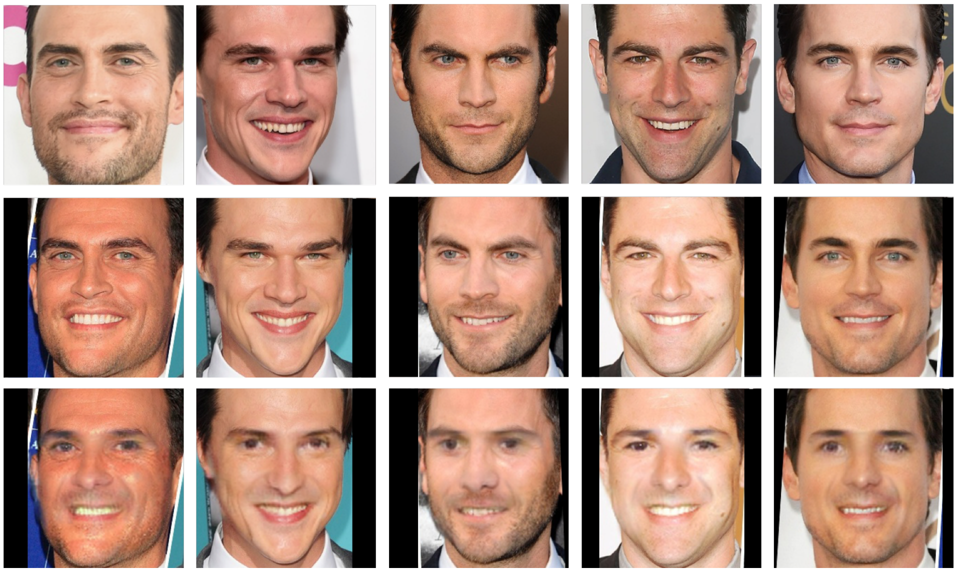


Figure 10: The images from the user study. Each column is a different individual. The first row are the gallery images, i.e, the album images the users were asked to select the identity from. The second row is the input image. The third row is the output of our method, i.e., the de-identified version of the second row.

B De-Identification Results Without the Zoom

For completeness, we append the original sized results, where no zoom-in was applied.

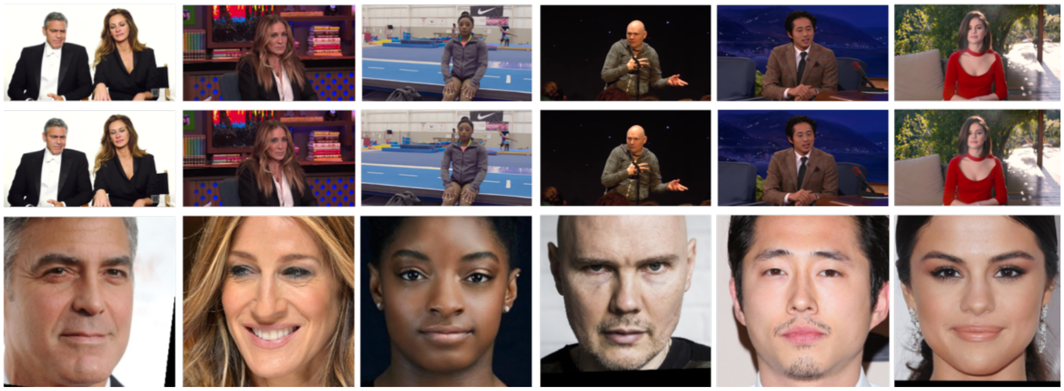


Figure 11: Same as Fig. 3 but without the zoom.

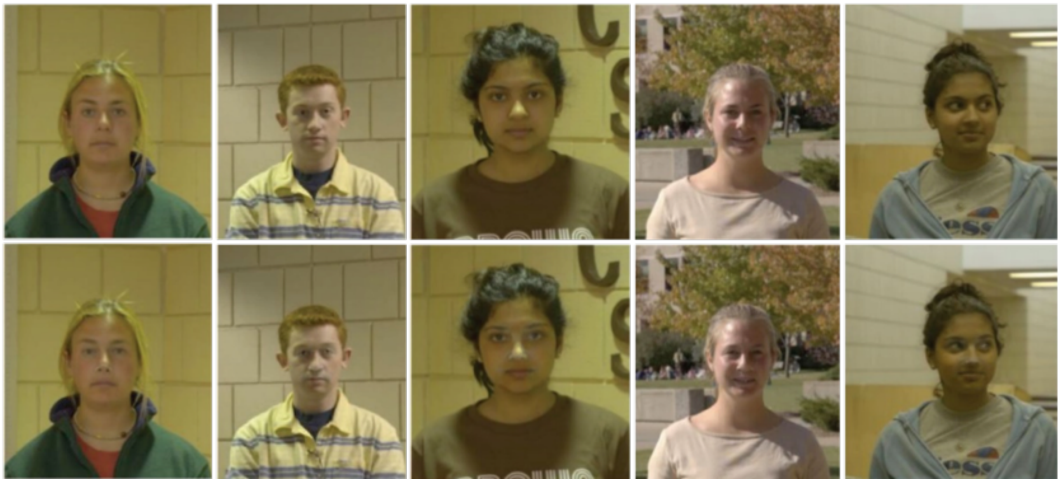


Figure 12: Same as Fig. 6 but without the zoom.