Selective Face Deidentification with End-to-End Perceptual Loss Learning

Blaž Meden*, Peter Peer* and Vitomir Štruc†
*Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail: {blaz.meden, peter.peer}@fri.uni-lj.si
†Faculty of Electrical Engineering, University of Ljubljana, Slovenia
E-mail: vitomir.struc@fe.uni-lj.si

***Abstract*—Privacy is a highly debatable topic in the modern technological era. With the advent of massive video and image data (which in a lot of cases contains personal information on the recorded subjects), there is an imminent need for efficient privacy protection mechanisms. To this end, we develop in this work a novel Face Deidentification Network (FaDeNet) that is able to alter the input faces in such a way that automated recognition fail to recognize the subjects in the images, while this is still possible for human observers. FaDeNet is based an encoder-decoder architecture that is trained to auto-encode the input image, while (at the same time) minimizing the recognition performance of a secondary network that is used as an so-called identity critic in FaDeNet. We present experiments on the Radbound Faces Dataset and observe encouraging results.**

I. Introduction

Artificial Intelligence (AI) is slowly becoming an everyday part of our lives. Recent developments in deep learning have pushed AI-based technology into information services, applications, gadgets, appliances, cars, and mobile platforms we use on a day-to-day basis. One area, where the current frontrunners in AI technology, i.e., deep models, have not yet made a significant impact is deidentification of personal data, and in particular, deidentification of facial images, which is a special kind of so-called Privacy Enhancing Technology (PET) that removes identity-related cues from the input imagery. Such technology is of paramount importance for ensuring privacy in services such as Google Street View or FourSquare, in multimedia data collections that are shared between government agencies, or online video-enabled chat rooms and video-conferencing apps, where vulnerable demographic groups, such as children or teenagers, are left exposed without suitable protective measures.

To address this gap, we introduce in this paper FaDeNet (Face Deidentification Network), a novel face deidentification approach based on convolutional neural networks (CNNs). We construct FaDeNet around so-called encoder-decoder networks, which have proven highly useful for different (conditional) image translation tasks, including semantic image segmentation [1], [2], [3]. Depending on the loss function used during training, encoder-decoder networks can be trained to modify (alter, degrade) selected image characteristics, while leaving others intact. We exploit this property in FaDeNet and train the network for facial deidentification by constructing an objective function composed of two terms. The first term is a pixel loss that forces the output image to be close to input and, thus, promotes information preservation, which is a desirable property of deidentification techniques. The second term is a recognition-oriented term that aims to *minimize* the recognition performance of a secondary CNN (pretrained for face recognition) that is used as a so-called identity critic during training and, thus, makes it difficult for automatic recognition techniques to identify the person in the generated output image. To achieve this *selective deidentification* we use an established face recognition model, i.e., InceptionV3 [4], in FaDeNet. As illustrated in Fig. 1, FaDeNet alters the input image in such a way that it becomes difficult for machine learning models to recognize, while the alterations are minute enough that humans are still capable of recognizing the individuals in the deidentified images.

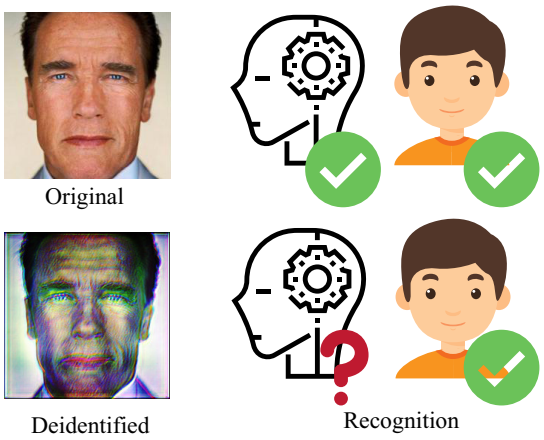

Figure 1: Our face deidentification approach, FaDeNet, takes a facial image as input (upper left) and alters it in such a way that it becomes difficult for machine learning models to recognize (lower left). The deidentified image is, however, still recognizable by human observers.

The main contributions of this work are:

- We introduce a deep face deidentification model called FaDeNet. The model relies on: *i)* an encoder-decoder network that generates deidentified facial images and *ii)* a pretrained face recognition model, i.e., InceptionV3, which serves as a constraint for the generative part of FaDeNet during training. The entire model is trained in

an end-to-end manner.
- We present an analysis of FaDeNet centered around reidentification experiments. We show that the deidentified images are of little use for face recognition models (even for models not used during training), but still carry identity information that is useful for human observers.

The rest of the paper is organized as follows. In Section II we cover related works in the area of face deidentification and briefly present some relevant deep learning models. In Section III we introduce FaDeNet for facial deidentification. In Section IV we describe the datasets used for our work, present training and implementational details, provide visual examples of facial images before and after deidentification and report results from reidentification experiments. We conclude our work in Section V with some directions for the future research work.

II. Related Work

In this section we describe the deidentification process and list key approaches that emerged in the field of facial deidentification. We also discuss recent deep models that are of relevance to this work.

A. Deidentification

Deidentification is defined as the process of concealing personal identifiers or replacing them with suitable surrogates in personal information in order to prevent the disclosure and use of data for purposes unrelated to the purpose for which the data was originally collected [5]. In image and video data, the process of deidentification is commonly related to deidentification of the facial region, which carries most of the identity-related information in the imagery.

Early deidentification techniques mostly included naive approaches, such as blacking-out, pixelation or blurring, which are generally considered as unsuitable for deidentification purposes [6], [7]. Blacking-out, for example, puts a black patch over the original face image to conceal identity, which guarantees anonymity, but also removes all non-identity related information – including characteristics that could be useful for further analysis (such as facial expressions). Pixelation and blurring are also considered unsuitable for deidentification, as they are prone to imitation attacks [8].

More recent face deidentification techniques from the literature try to overcome the above limitations by exploiting formal privacy models, such as k-anonymity, which provide formal guarantees regarding the anonymity of the deidentified data [9], [8], [10]. Methods from this group include the seminal k-Same approach [9] and related extension, such as k-Same-Model [11], k-Same-select [12], or the more recent k-Diff-furthest approach [13]. Most of these techniques rely on Active Appearance Models (AAMs) and ensure convincing visual deidentification results. Meden et al. [14] recently presented a CNN-based approach based on the k-anonymity model and showed that generative neural networks present a viable alternative to AAMs, which mitigates some of the problems associated with AAMs based deidentification. In this paper, we follow this line of work and also present a CNN-based face deidentification approach. Similarly to [14], our approach also tries to achieve privacy protection through the use of a generative model, however, our goal is to hide the identity of the individuals in the imagery only for machine learning models and not humans, which is conceptually very different from the goals of [14].

Our model is related to the work of Chriskos et al. [15], which deidentifies facial images using projections on hyperspheres. As with our approach, the deidentified images produced by the work in [15] are still recognizable by human observers, but represent a challenge to automated recognition algorithms. Another related approach is presented by Otman and Ross in [16]. In this work, the authors use morphing to hide gender information, while still preserving the identity of the individuals. In a follow up work, Mirjalili and Ross [17] described how to manipulate gender in face images while retaining biometric utility.

For a more comprehensive review of existing face deidentification techniques, the reader is referred to some of the recent surveys on this topic [5].

B. Deidentification with Deep Learning

Deep learning models have recently been shown to ensure state-of-the-art performance in a number of vision-related tasks [18], [19], [3] and have also been considered as a tool for deidentification. Meden et al. [20], for example, introduced a face deidentification approach using generative neural networks. Here, a deidentification pipeline is described that ensures identity protection through the use of a parametrized generative network. The network is able to produce alternative facial identities from a closed set of faces that can be used as surrogates during deidentification. The approach was later extended by integrating it into a formal k-anonymity based deidentification model, called k-Same-Net [14].

Chi and Hu [21] used facial identity preserving (FIP) features to preserve the aesthesis of the original images, while still achieving k-anonymity-based facial image de-identification and preserving desired utilities.

Brkić et al. [22] present a deidentification approach for soft and non-biometric traits (including facial synthesis, hairstyle and clothing) in video sequences. They use Deep Convolutional Generative Adversarial Networks (DCGANs) for face synthesis, recoloring scheme for clothing-color and hairstyle deidentification and perform rendering of extracted (segmented) hairstyles from existing facial images onto the synthesized faces.

Mirjalili et al. [23] present an privacy protection approach based on semi-adversarial learning which is similar to the approach presented in this work from a methodological point of view (using an auto-encoder architecture with additional classifier), but differs in the application domain (privacy protection of certain attributes, such as gender, race or age). The transformed facial images, produced by the approach from [23] can be used for face recognition, but are challenging for automated gender classification. Our approach, on the other hand, tries to hide identity information from the images that

is important from a machine learning perspective, while still preserving the ability for human recognition of the altered images.

C. Perceptual Learning and Adversarial Attacks

Recent machine learning techniques often define objective functions based on high-level image representations produced by deep learning models. These functions, commonly referred to as *perceptual loss functions*, are at the hart of many state-of-the art vision techniques and are also used in the work presented in this paper.

An example of perceptual learning is presented by Johnson et al. in [24], where a perceptual loss is used for image translation tasks and image super-resolution. A similar idea is presented for heterogeneous face recognition in by Sarfraz and Stiefelhagen in [25]. Here, the authors use an objective functions that try to minimize the difference between image features computed from facial images captured in different visual domains (visible light. vs. near infrared). Notably, our work uses a similar idea and also relies on a objective function computed from a pretrained CNN model, but uses a global objective (i.e., minimization of the recognition performance) instead of an objective defined over image features.

Our work can be seen to be related to the area of *adversarial attacks*, where the goal is to alter the input images is such a way that they become unrecognizable by machine learning models or force the models into incorrect predictions. The work from Nguyen et al. [26], for example, demonstrates how deep architectures can be fooled into making high confidence predictions for unrecognizable images, which might look completely noisy or unreal to a human observer. Similarly, Su et al. [27] describe how predictions of deep models can be altered by modifying only one pixel. Another example from this area is utilizing adversarial attacks [28] to force deep models into making incorrect predictions. The goal of our FaDeNet is similar to methods focusing on adversarial learning, but our model tries to do the opposite - making images unrecognizable for machine learning models, and less so for human observers.

III. Face Deidentification Network - FaDeNet

We now describe our selective face deidentification approach built around the Face Deidentification Network, FaDeNet. We start the section with a high-level overview of FaDeNet and then discuss the architecture of FaDeNet's components as well as the training procedure and loss used during model learning.

A. Overview of FaDeNet

As illustrated in Fig. 2 FaDeNet consists of two main parts, which we refer to as:

- *A transformer network* which tries to alter (deidentify) the input image $\mathbf{x}$ in such a way that the difference between the input and output image is minimized. The minimization procedure is needed to ensure that as much of the available information of the input as possible is preserved in the deidentified image. If we denote the output of the transformer network as $\mathbf{y}$, then the input-output mapping $f_{\theta_{TN}}$ can be defined as:

$$\mathbf{y} = f_{\theta_{TN}}(\mathbf{x}), \tag{1}$$

  where θ_{TN} denotes the set of model parameters that need to be learned during training.
- *An identity critic*, a CNN model pretrained for face recognition that is used as a recognition constraint. It serves as the identity classifier upon which a perceptual loss can be defined. The goal of the perceptual loss is to maximize the difference between identity labels obtained from the input image $\mathbf{x}$ and deidentified image $\mathbf{y}$. The identity critic is used as a constraint for the training procedure and is not needed during run-time. If we denote the identity critic as $f_{\theta_{IC}}$, then the identity labels $\mathbf{x}_{id}$ (or better said the class-probability distribution) can be computed from an input image $\mathbf{x}$ as follows:

$$\mathbf{x}_{id} = f_{\theta_{IC}}(\mathbf{x}), \tag{2}$$

  where θ_{IC} again stands for the set of model parameters.

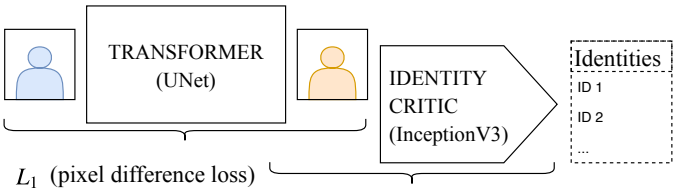

Figure 2: Illustration of FaDeNet. The model uses an encoder-decoder network (i.e., U-Net - the transformer) to selectively deidentify the input image, and a pretrained recognition model (i.e., InceptionV3 - the identity critic) to ensure that the deidentified images are not useful for automatic face recognition. The recognition model is not used during run-time.

When implementing FaDeNet, we use a U-Net [1] architecture for the transformer network and separately trained InceptionV3 [4] model as our identity critic. The U-Net architecture is chosen due to it's popularity, state-of-the-art performance for various vision-oriented tasks and the fact that an implementation is readily available[1]. The InceptionV3 model, on the other hand, is selected because of its light-weight nature, which allows us to train FaDeNet efficiently without significant computational overheads. Next we describe the transformer network and identity critic in detail.

B. The Transformer Network – U-Net

The transformer network (illustrated in Fig. 2) is a standard U-Net [1], a state-of-the-art architecture initially developed for fast and precise segmentation of images. The U-Net architecture relies on an encoder-decoder structure that consists of a contractive and an expansive path. The contractive path

[1] https://github.com/zhixuhao/unet

follows the typical architecture of a convolutional encoder, consisting of repeated 3×3 convolutions followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for down-sampling. Each down-sampling operation doubles the number of feature channels. In the expansive path, every step consists of an up-sampling operation of the feature maps, followed by a 2×2 convolution (also denoted as "up-convolution") that halves the number of feature channels. The expansive path also concatenates appropriately cropped feature maps from the corresponding contracting path at each level. The concatenated feature maps are followed by two 3×3 convolutions and a ReLU activation. The final layer uses 1×1 convolutions to map each 64-component feature vector to the desired number of classes. The architecture has 23 convolutional layers in total [1]. To make the architecture suitable for our purposes, we modify the last layer of the network, so it generates 3-channel color images at the output.

C. The Identity Critic Network – InceptionV3

For the identitiy critic, we use the InceptionV3 [4] model, pretrained for face recognition. It needs to be noted that the parameters of the InceptionV3 model are not learned during FaDeNet training, instead the model is used with frozen weights and is exploited only as a recognition constraint (i.e., for the definition of a perceptual loss) when training the transformer network.

The InceptionV3 architecture is based on the idea of incorporating multiple smaller models inside a bigger network. Each of these smaller models, referred to as inception modules, consist of multiple parallel pathways and each of them uses different convolution and pooling layers to recover local features via smaller convolutions and high abstracted features with larger convolutions. Specifically, a single inception module consists of 4 pathways: a single layer with 1×1 convolution, two layered 3×3 and 1×1 convolutions, two layered 5×5 and 1×1 convolutions and two layered 1×1 convolutions, followed by 3×3 max pooling layer. Each inception module achieves a significant amount of dimensionality reduction via a filter concatenation layer, which combines all 4 parallel pathways into a single feature map of reduced size and applies a ReLU activation before passing it into the next inception module. The resulting network model is deeper and more complex than many competing architectures, but still has a relatively small number of parameters and lower computational complexity than most competing models. No fully-connected layers are used in InceptionV3, Instead, the last convolutional map is subjected to channel-wise global average pooling, and the average activation values of each of the 2048 channels are typically used as the feature vectors of the input image and/or the input to a softmax classifier during training [19].

D. Loss and Training

The goal of the FaDeNet training procedure is to learn the parameters of the transformer network θ_{TN} using a combined pixel-level and perceptual loss, where the pixel-level loss is defined over the inputs and outputs of the transformer network and the perceptual loss is defined over the outputs of the softmax classification layer at the top of the InceptionV3 model.

Consider a training set of N facial images $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ belonging to M identities. The end-to-end FaDeNet training procedure can then be defined based on the following two objective functions:

- *The pixel-level loss* $\mathcal{L}_1(\theta_{TN})$ that penalizes the difference between the input and output images of the transformer network and forces the transformed (deidentified) image to be close to the input (making it recognizable to human observers):

$$\mathcal{L}_1(\theta_{TN}) = ||\mathbf{x}^{(i)} - f_{\theta_{TN}}(\mathbf{x}^{(i)})||^2, \tag{3}$$

  where $\mathbf{x}^{(i)}$ stands for an image from $\mathcal{X}$ and $||\cdot||$ denotes the L_2 norm.
- *The perceptual loss* $\mathcal{L}_2(\theta_{TN})$ that penalizes the cross entropy between the identity labels corresponding to the input and output images predicted by the identity critic and consequently (indirectly) encourages the identity features corresponding to the input and output images of transformer network to be different (making the images unrecognizable to machine learning models). The perceptual loss is defined as:

$$\mathcal{L}_2(\theta_{TN}) = -\sum_{c=1}^{M} \mathbf{x}_{id}^{(i)} \log \mathbf{y}_{id}^{(i)}, \tag{4}$$

  where $\mathbf{x}_{id}^{(i)}$ denotes the ground truth class probability distribution (a one-hot vector) of the input image $\mathbf{x}^{(i)}$ and $\mathbf{y}_{id}^{(i)}$ denotes the output probability distribution produced by the softmax layer of identity critic based on the deidentified facial image $\mathbf{y}^{(i)}$.

Finally, the combined loss $\mathcal{L}_{total}$ used for FaDeNet training is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_1 - \lambda\mathcal{L}_2, \tag{5}$$

where λ is loss weighting parameter with value $\lambda = 0.001$. We minimize $\mathcal{L}_{total}$ on our training set $\mathcal{X}$ using error back-propagation. Note that the minus in the above equation ensures that we minimize the recognition performance on the deidentified training images. Before calculating $\mathcal{L}_1$, we normalize all training images from $\mathcal{X}$ onto the unit circle using the Euclidean norm.

IV. Experiments and Results

This section presents experiments to evaluate our (selective) transformative deidentification approach. We first discuss the datasets used for network training and experimentation, we describe the network training and the data augmentation techniques used and lastly present visual results along with a qualitative analysis which suggest that our proposed deidentification approach is indeed effective against automatic deep recognition models while preserving the possibility of recognition by humans.

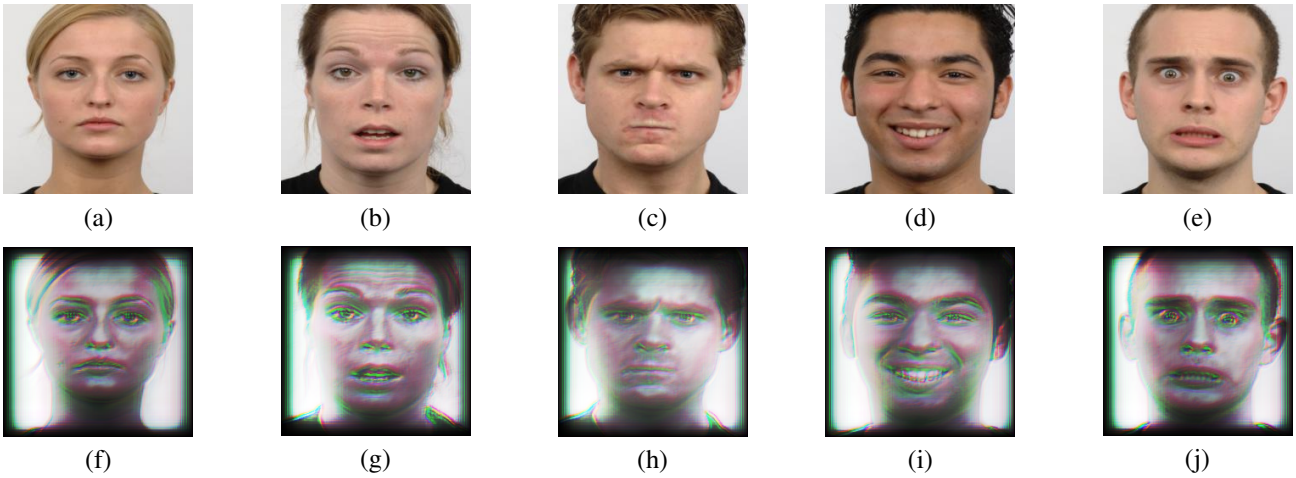

Figure 3: Visual examples of original and deidentified pairs of faces. Images (a) – (e) show the original input images and images labeled (f) – (j) show deidentified images generated by FaDeNet. Note that the deidentified images are visually still very close to the originals from a human perspective, but are very different in the domain of extracted identity features, when a pretrained InceptionV3 face recognition network is used for feature extraction. The results are best viewed in color.

Method	**Rank-1 ($\mu \pm \sigma$) – orig.**	**Rank-1 ($\mu \pm \sigma$) – deid.**	**Rank-5 ($\mu \pm \sigma$) – orig.**	**Rank-5 ($\mu \pm \sigma$) – deid.**
Inception V3 [4] (id. critic)	0.353 ± 0.015	0.023 ± 0.005	0.946 ± 0.028	0.119 ± 0.007
VGG-Face [29]	0.902 ± 0.024	0.058 ± 0.012	1.0 ± 0.019	0.254 ± 0.026

Table I: Recognition performance over five repetitions of recognition experiments with images form the Radbound Faces Dataset before and after deidentification. Two different deep recognition approaches are included in the comparison: our identity critic network InceptionV3 and the independently trained VGG-Face (trained by Grm et al. [19]).

Figure 4: Close-up detail of a texture rich region in the deidentified image. We can clearly see that introducing the perceptual loss term during training results in modified color channels in the deidentification proces. This, in turn, reduced the similarity of the extracted features in comparison with the features obtained from the original images.

A. Datasets

We use the XM2VTS dataset [30] to train FaDeNet end-to-end. The dataset consists of good quality, mostly frontal face images with neutral facial expressions, taken against a uniform background. There is a total of 2360 face images of 295 subjects in the dataset (eight images per subject, shot in multiple sessions).

We use the Radbound Faces Dataset [31] as our test image set, due to the small number of identities included, high quality image with good alignment with eyes facing towards the camera, providing uniform image background and very good illumination conditions. The dataset contains facial images of 67 subjects with eight different facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise, contempt and neutral) per subject. The complete dataset captures faces in three different gaze directions and from five camera angles under all eight facial expressions. From these images, we select only the frontal images displaying 57 adult subjects with 8 different facial expressions for our test set in the recognition experiments. There is a total of 456 images in the dataset, and we equally divide them into 228 gallery images and 228 probe images).

B. Training details

First the identity critic network of FaDeNet is trained as identity classifier with a batch size of 32 over 40 epochs, using the Adam optimizer with default parameters.

After training the identity critic, the transformer network of FaDeNet is trained with a batch size of 32 over 40 epochs, using the Adam optimizer with default parameters and the learning rate of $lr = 0.00001$. The dimensions of the input and output images are fixed to 224×224 pixels, due to the

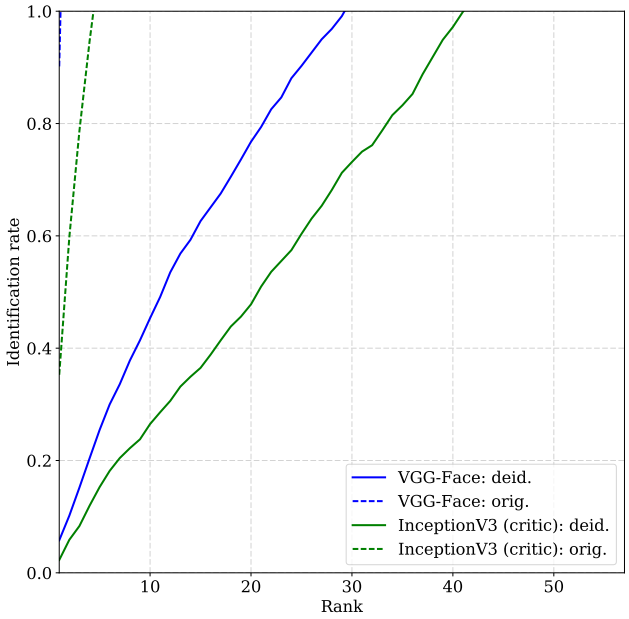

Figure 5: Recognition experiments: average Cumulative Match Characteristics (CMC) curves over five repetitions of the recognition experiments on the original and deidentified images of the Radbound Faces Dataset with two recognition techniques. Results indicate that unaltered identities are correctly recognized most of the time with both techniques considered, while the performance is severely degraded after performing selective deidentification.

usage of same size input and output dimensions while training identity critic network.

To avoid over-fitting, we perform various data augmentation tasks during training. Specifically, we use random horizontal flips with a probability of 0.5. Additionally we apply cropping and padding in the range $[-5\%, 10\%]$, using the `imgaug` augmentation library[2]. Lastly, we add some affine transformations to the sequential augmentation in form of random scaling with a scaling factor in the range $[80\%, 120\%]$ in both image directions, and random rotations in the range of $[-5°, 5°]$. Such augmented images enable the transformer network to converge quicker and produce better resulting output images.

C. Visual Examples

Some visual examples of image pairs before and after the deidentification are displayed in Fig. 3. As we can see, the transformer network mixes the color channels, so the output appear as selectively perturbed color-like images with visually recognizable facial structures while modifying only local patterns in such way, that the InceptionV3 recognition model is unable to infer the correct identity. In other words, the network alters the image in such way, that recognition is still possible for a human observer, but it becomes significantly difficult when recognition is done based on extracted features generated by a pretrained recognition network, such as InceptionV3 or VGG-Face as we show in the next section.

In Fig. 4 a facial patch from a texture rich region is enlarged to highlight the effect of the transformer network. We can see how edges and other high frequency details are altered when the patch is examined up-close. If we observe the details carefully, we can distinguish between displacements and color alterations of each of the image channels. We presume that these displacements are the main reason for facial deidentification in the feature space, as it is indicated in results in the next Subsection IV-D.

D. Recognition experiments

We quantify the efficiency of our deidentification approach on automated recognition models through recognition experiments with images from the Radbound Faces Dataset [31]. We randomly split the 456 images of 57 identities from the dataset during each experimental run into probe image set consisting of 4 randomly selected facial images per each identitiy to be deidentified (total of 228 images) and the remaining set of 4 images per identity to serve as the gallery (total of 228 images).

We perform identification experiments with the constructed probe and gallery sets, repeating this process five times and report the identification performance in the form of the average rank one and rank five recognition rates (Rank-1 and Rank-5) and corresponding standard deviations computed over the five experimental repetitions. Experiments are conducted before and after deidentification to assess the performance level of deidentification by the proposed approach.

The results of the recognition experiments on the unaltered and deidentified Radbound Faces Database are shown in Figure 5 in the form of Cumulative Match Characteristics (CMC) curves and in Table I as mean Rank-1 and Rank-5 rates with corresponding standard deviations.

We use two recognition models for the experiments: i) InceptionV3 that was already used during FaDeNet training, and ii) an independently trained CNN face recognition model VGG-Face (taken from [19]). The latter model is used to evaluate the generalization capability of the deidentification approach. We use features from the 6-th dense layer when evaluating VGG-Face, which produces a feature vector of 4096 values. When evaluating InceptionV3 we obtain features from the last flattening layer of size 2048, which is placed in between last average pooling and final softmax layer. Once the features are extracted, we rely on the cosine distance to measure the similarity between feature vectors and conduct recognition.

As can be seen, both of the two evaluated recognition approaches (our identity critic and independently trained VGG-Face) achieve high recognition rates on unaltered images, which suggests that the original images are relatively easy to identify. On the other hand, the recognition performance is severely impacted by our deidentification approach, as it can be seen in Figure 5.

All in all, the empirical evaluation suggests that the risk of reidentification for images deidentified with our approach is severely degraded. The recognition performance of all

[2] https://github.com/aleju/imgaug

tested recognition techniques has significantly dropped after performing deidentification.

V. Conclusions

Our results indicate that introducing an additional critic network into the training process of a transformer network for face deidentification results in a deep model, that can deidentify input images in such way, that the processed images are still recognizable to human observers, however the features obtained from these altered images (by using deep recognition networks) are scrambled to a degree that hinders the possibility of re-identification by state-of-the-art automated methods based on deep learning. In our opinion, this research opens new opportunities for research on deep learning based deidentification as well as offers new insights and questions about the interpretability of deep features produced in such models.

Acknowledgements

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Program P2-0250 (B) Metrology and Biometric Systems, the ARRS Research Program P2-0214 (A) Computer Vision, and the RS-MIZŠ and EU-ESRR funded GOSTOP. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[2] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Ž. Emeršič, P. Peer, V. Struc, S. V. A. Kumar, and B. S. Harish, "Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 742–747.

[3] Ž. Emersic, L. L. Gabriel, V. Struc, and P. Peer, "Convolutional encoder-decoder networks for pixel-wise ear detection and segmentation," *IET Biometrics*, vol. 7, no. 3, pp. 175–184, 4 2018.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2818–2826.

[5] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131 – 151, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596516300856

[6] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000, pp. 1–10.

[7] C. Neustaedter, S. Greenberg, and M. Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 1, pp. 1–36, 2006.

[8] R. Gross, L. Sweeney, J. Cohn, F. De la Torre, and S. Baker, "Face de-identification," in *Protecting privacy in video surveillance*. Springer, 2009, pp. 129–146.

[9] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, Feb 2005.

[10] R. Gross and L. Sweeney, "Towards real-world face de-identification," in *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*. IEEE, 2007, pp. 1–8.

[11] R. Gross, L. Sweeney, F. de la Torre, and S. Baker, "Model-based face de-identification," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, June 2006, pp. 161–161.

[12] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *Proceedings of the 5th International Conference on Privacy Enhancing Technologies*, ser. PET'05. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 227–242. [Online]. Available: http://dx.doi.org/10.1007/11767831_15

[13] Z. Sun, L. Meng, and A. Ariyaeeinia, "Distinguishable de-identified faces," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 04, May 2015, pp. 1–6.

[14] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer, "k-same-net: k-anonymity with generative deep neural networks for face deidentification," *Entropy*, vol. 20, no. 1, p. 60, 2018.

[15] P. Chriskos, O. Zoidi, A. Tefas, and I. Pitas, "De-identifying facial images using singular value decomposition and projections," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 3435–3468, Feb. 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-4069-8

[16] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015, pp. 682–696.

[17] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 564–573, 2017.

[18] K. Grm, S. Dobrisek, and V. Struc, "Deep pair-wise similarity learning for face recognition," in *2016 4th International Conference on Biometrics and Forensics (IWBF)*, March 2016, pp. 1–6.

[19] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *IET Biometrics*, vol. 7, no. 1, pp. 81–89, 2017.

[20] B. Meden, R. C. Mallı, S. Fabijan, H. K. Ekenel, V. Štruc, and P. Peer, "Face deidentification with generative deep neural networks," *IET Signal Processing*, vol. 11, no. 9, pp. 1046–1054, 2017.

[21] H. Chi and Y. H. Hu, "Face de-identification using facial identity preserving features," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2015, pp. 586–590.

[22] K. Brkić, T. Hrkać, Z. Kalafatić, and I. Sikirić, "Face, hairstyle and clothing colour de-identification in video sequences," *IET Signal Processing*, July 2017. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-spr.2017.0048

[23] V. Mirjalili, S. Raschka, A. M. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *Proc. of 11th IAPR International Conference on Biometrics (ICB 2018)*, February 2018.

[24] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 694–711.

[25] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," *International Journal of Computer Vision*, vol. 122, no. 3, pp. 426–438, 2017.

[26] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," 2015.

[27] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *CoRR*, vol. abs/1710.08864, 2017. [Online]. Available: http://arxiv.org/abs/1710.08864

[28] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[29] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[30] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity, "Face verification competition on the xm2vts database," in *Proceedings of the 4th International Conference on Audio- and Video-based Biometric Person Authentication*, ser. AVBPA'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 964–974. [Online]. Available: http://dl.acm.org/citation.cfm?id=1762222.1762347

[31] O. Langner, R. Dotsch, G. Bijlstra, D. Wigboldus, S. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition&Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.