EPD-Net: A GAN-based Architecture for Face De-identification from Images

Alakh Aggarwal
CSE Department
IIT (BHU), Varanasi
Varanasi, India
alakh.aggarwal.cse15@iitbhu.ac.in

Rishika Rathore
Chemical Engineering Department
IIT (BHU), Varanasi
Varanasi, India
rishika.rathore.che17@iitbhu.ac.in

Pratik Chattopadhyay
CSE Department
IIT (BHU), Varanasi
Varanasi, India
pratik.cse@iitbhu.ac.in

Lipo Wang
School of EEE
Nanyang Technological University
Singapore
ELPWang@ntu.edu.sg

***Abstract*—Nowadays huge amount of crowd data captured by surveillance cameras gets shared publicly in the form of images or videos through television or the internet. Although many of these videos are meant to provide public security, they also lead to a widespread concern towards privacy protection, since a lot of personal information about subjects gets revealed through this video/image data. Hence, "de-identifying" people (i.e., obscuring identity information) captured by the surveillance cameras is of utmost importance for providing privacy along with security. Traditional identity obfuscation techniques such as blurring, warping, and filtering lead to the loss of vital non-biometric information. More recent k-same-based as well as generative model-based de-identification techniques eliminate the above problem to a certain extent. Still, the visual quality of the generated images produced by these methods is not realistic. Also, these approaches are unable to maintain structural integrity and cannot preserve the required non-biometric information at a high resolution. As an improvement, in this paper, we propose a new network termed as EPD-Net and train it with suitable loss functions to maximize the emotion similarity and minimize the identity similarity. Experimental results verify the effectiveness of our approach and its superiority over other popular face de-identification techniques.**

***Index Terms*—De-identification, Identity Obfuscation, Emotion Preservation, Generative Adversarial Networks, Privacy Preservation**

I. Introduction

In the present days, a large volume of data gets shared over the internet across different parts of the world. In this process of transfer and share of visual data, leak of privacy becomes the primary concern, since these data reveal the biometric traits of the subjects without their consent. Most real-life applications working on crowd data focus on providing public security, where knowledge of the identity of individuals has lower significance, and capturing other non-biometric information such as race, gender, facial expression/emotion have a higher significance. For example, consider a public hospital scenario where the number of patients is large, in comparison to the number of hospital staff, and monitoring the condition of every patient constantly becomes very difficult. The facial expression of patients captured by the surveillance cameras installed in the hospital corridors and/or patient cabins provide vital clues about the physical and mental conditions of the patients. By observing this data the hospital authorities can take appropriate measures and/or provide necessary treatments, if required. However, these videos often reveal the identity of the patients captured in the videos. From the above discussion, it is clear that there is a need for de-identifying patients' identity, while simultaneously preserving their emotions. To the best of our knowledge, there exists no work in the literature that solves this problem effectively except the work in [1].

Fig. 1: Sample images from RafD (Radboud Faces Dataset) consisting of 67 subjects belonging to different classes, namely, Caucasian male, Caucasian female, Moroccan male, Kids male, Kids female, etc.

978-1-7281-9615-2/20/$31.00 ©2020 IEEE

Although the objective of this work is similar to that of ours, the images generated by this method lack in quality, and the identifiable characteristics of a face are not eliminated well. The de-identified faces can be easily identified visually and this makes it unsuitable for application in real-life.

In this paper, we aim to improve the existing solutions to the problem of face de-identification, while simultaneously preserving emotion (or, facial expression) information. GANs [2] have been extensively used in recent years for several image generation or translation tasks such as image de-fencing [3], colorization [4], etc. The present work is an application of image to image translation, where a facial image is provided as input, and a de-identified version of the same face is obtained as the output. Motivated by the recent success of conditional GANs (in short, cGANs) in image translation tasks [5], we also use this category of the network in the present work. Specifically, we propose a new network architecture termed as EPD-Net (Emotion Preserving De-identification Network), that consists of a pix2pix GAN [5] along with two additional auxiliary deep networks, namely the identity and emotion verificators. When a face image is input to the EPD-Net, the network does certain computations in its layers to generate a realistic de-identified face image that preserves the emotional characteristics of the input face and eliminates identity characteristics to the extent possible.

The rest of the paper is organized as follows. A thorough survey of related literature is discussed in Section II. Section III introduces the proposed EPD-Net network along the loss functions used in training the model. Extensive experimental evaluation and comparison with some competing approaches are presented in Section IV. Section V finally concludes the paper and points out possible future scopes for research.

II. Related Work

Privacy protection in images/videos is very important since most online images and surveillance videos inadvertently disclose the identities of subjects captured. In the past, research work on de-identification has been mostly focused on developing algorithms for facial identity obfuscation such as [6]–[12], and only a few gait/silhouette de-identification techniques have been proposed, such as [13], [14]. The gait/silhouette de-identification algorithms commonly use only blurring operator to de-identify a subject silhouette, which cannot always guarantee anonymity. Since the theme of the paper is developing algorithms on face de-identification by preserving emotional characteristics, here we focus on describing the research trend on face de-identification, and skip discussions related to gait/silhouette de-identification.

Previous techniques used for de-identification were simplistic in nature. They were mostly aimed at concealing detailed texture information present on the face by the means of blurring [6], warping [7], [8], pixelation [15], and other image processing-based techniques. This category of these approaches is termed as non-formal, since these do not provide any formal proof of anonymity. As an improvement, k-same family of approaches was developed later, which has been used for de-identification tasks in various studies such as [9]–[11]. These approaches, on the surface, work by finding cluster centers of k-clusters from the data set and replacing the cluster instances with the cluster centers. Although the drawback of uncertain anonymity can be successfully handled by the means of the k-same family of approaches, the resulting de-identified images look highly non-realistic, and bear ghostly appearances, due to the averaging operation used in computing the cluster centers. With the introduction of Generative Neural Networks (GNNs), learning based techniques were developed to generate better quality de-identified faces, for example, [1], [12].

In [12], Meden et. al. proposed a generative neural network (GNN) based approach in which proxy clusters are generated using the k-same algorithm and the de-identified faces are generated by the GNN. The complete de-identification network has been termed as the k-same-net, and although it improves over the above-mentioned non-formal techniques, the generated images still look non-realistic. Moreover, each of the k-same family of de-identification approaches, including k-same-net, suffer from the following two shortcomings: (i) The k-same approach assume that each subject is only represented once in the data set, but this may not be true in practice. The presence of multiple images from the same subject, or images sharing similar biometric characteristics can lead to lower levels of privacy protection, (ii) k-same operates on a closed set of facial images and produces a corresponding de-identified set, which is not applicable in situations that involve the processing of individual images, or sequence of images outside the domain of data set.

More recent approaches employ the generalization power of Generative Adversarial Networks (GANs) [2] to de-identify faces. The basic architecture of a GAN consists of a generator network that generates an output based on the input data's probability distribution function and a discriminator that distinguishes the generated output from the input data. Typically, a GAN is trained in multiple iterations using Minimax algorithm, and this imparts an ability to the generator network to accurately fit the input probability distribution function. In recent years, researchers have aided the generation and discrimination process of GANs with slightly different but similar concepts. For example, the work in [1] employs a GAN along with a verificator network to obtain a similarity measure between the features of the generated image and the desired features, as well as a regularizer network to compute the structural similarity between the two images. The objective function for training the GAN accounts for the combined loss given by the verificator and the discriminator networks. However, the objective function in [1] does not take into account emotion preservation. Hence, the de-identified images generated by this approach cannot guarantee emotion preservation effectively. Our observation is that the images generated by [1] also do not look significantly realistic (results are shown in Section IV).

In this paper, we propose an improvement to the work in [1], by adding an additional verificator network to preserve

the emotion information at a higher resolution compared to [1]. Also, we propose to use a deep classification network as an identity verificator network instead of the Siamese network as in [1]. This is since the function of the identity verificator is to predict if the generated face corresponds to a particular identity. This can be effectively done if the network is trained to learn a mapping from an input face to its corresponding identity. The Siamese network, on the other hand, does not take into account the ground truth label information, and predicts the similarity of the generated face with the original face based on the similarity in their appearances. The use of the classification network as an identity verificator also helps in obfuscating identity information better than that [1].

The main contributions of the work are as follows:

- Developing an improved network termed as EPD-Net for face de-identification that preserves emotion characteristics of a face effectively while simultaneously obfuscating the facial identity information.
- Improving the identity verificator network used in a previous study [1] by employing a deep convolution classification network instead of Siamese network.
- Extensive experimental evaluation and comparison with competing approaches.

III. Proposed Approach

The face de-identification process can be viewed as a mapping from an input face image x to its corresponding de-identified image $\hat{x}$ using a mapping function f such that $f(x) = \hat{x}$. In an ideal situation, the mapping function f must be non-invertible, i.e., it should not be possible to recover the original face from the de-identified face. But as discussed in Section I, obfuscation of identity features is not the only goal of this work. Along with identity obfuscation, emotion loss should be minimized to the extent possible. Generative Adversarial Networks (GANs) are a popular choice for researchers to generate images following a specific distribution pattern. Since the objective of the present work is generating face images by satisfying a given set of conditions, it appears that a suitable GAN-based architecture can be employed to learn the mapping function f. The basic version of GAN consists of two networks: (a) a generator network which tries to learn the input data distribution, and (b) a discriminator network which predicts whether an input image belongs to the input distribution (i.e., real data) or the distribution learned by the generator network (i.e., fake data). However, using this basic structure of GAN, it is not possible to achieve the dual objective of identity obfuscation and emotion preservation. If a set of auxiliary features needs to be employed while simultaneously learning the input data distribution, then conditional GANs (in short cGANs) [16] seem to be the best choice. In a cGAN, separate loss functions are designed corresponding to each auxiliary information using either standard metrics or neural networks, and the combined loss from all the auxiliary networks/metrics are jointly minimized along with the discriminator loss function.

To construct the EPD-Net, we consider two separate deep networks along with a pix-2-pix GAN and design loss functions to satisfy the following conditions: (a) the generated image must be free from identifiable features present in the input face, and (b) the generated image must preserve the emotion of the input face. We term these auxiliary networks as *identity verificator* and *emotion verificator* networks, and these produce scores depicting the identity similarity and emotion dissimilarity between the input and the generated de-identified faces, respectively. The architecture of the proposed network can be schematically represented using Figure 2. With reference to the figure, the network has the following

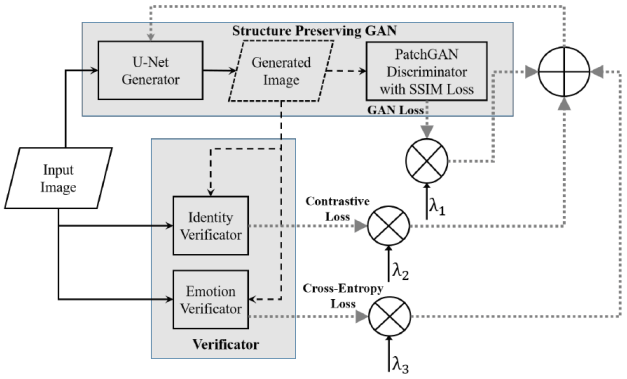

Fig. 2: An Overview of the proposed network architecture

modules: (a) a cGAN that is based on pix-2-pix network as described in [5], and (b) identity and emotion verificator networks, as described later. The module (a) consists of a generator and a discriminator. The generator generates an image and the discriminator outputs a score (shown as *GAN Loss*) depicting whether the generated image belongs to the probability distribution of the gallery set or not. The verificator networks compute two loss metrics, shown as *Contrastive Loss* and *Cross-Entropy Loss*, that respectively indicate the extent of identity similarity and emotion dissimilarity between the input and the generated images. A weighted combination of these loss values at a particular epoch is fed back to the generator, which then generates an improved face image in the subsequent epoch. The combination tallies more accurately with the different conditions or constraints. In the figure, λ_1, λ_2, and λ_3 represent the weighting factors corresponding to *GAN Loss*, *Contrastive Loss*, and *Cross-Entropy Loss*, respectively.

A. Structure Preserving GAN

The GAN architecture and the training algorithm used for face de-identification are described in detail next. Unlike the previous work on face de-identification, we focus on generating more realistic face images with distinct facial features (i.e., eye, nose, mouth, etc.) and clear face boundaries. For this, a structural similarity loss term is also optimized along with the standard discriminator loss functions in the objective function. The pix-2-pix GAN consists of an U-Net generator

[17] and a variant of the PatchGAN discriminator, which is a fully convolutional network formed by the *Conv-BatchNorm-ReLU* blocks as described in [5]. The loss function for training the model is computed at the final layer of this discriminator. The generator is fed with a random noise vector (say z), derived from noise distribution given by P_z. Let us also assume that the input data distribution is given by P_{data}, and the distribution learned by the generator at a particular epoch by P_{gen}. For each input image x in the training set, the generator performs a mapping of the form $G : (x, z) \mapsto \hat{x}$, where $x \in P_{data}$ and $\hat{x} \in P_{gen}$. The GAN is trained in multiple epochs and the training process stops when the two distributions P_{gen} and P_{data} become similar.

As the generator continues to learn the input distribution in a better way at each epoch, the discriminator network also keeps on improving its prediction power so that it can correctly classify a given image into the appropriate class: i.e., real or fake. As already explained before, this is a binary classification task where the two classes correspond to samples from the distributions P_{data} and P_{gen}. Let us assume that the ground truth label corresponding to P_{data} is [0,1] while for P_{gen} it is [1,0]. The mutual conflict between the two networks helps in gradual improvement of the distribution function learned by the generator, as well as the prediction capability of the discriminator. On completion of the training phase, the generator learns the distribution P_{data} accurately and it attains the capability of generating images that look similar to the original data.

Suppose, at a particular epoch, the discriminator is trained with N training patterns, and let the i^{th} pattern be denoted by y_i, where y_i is the input image. The discriminator distinguishes the probability distribution in which y_i belongs, such that $y_i \in P_{data}$ or $y_i \in P_{gen}$. The mini-max loss function to be optimized for the generator-discriminator network can be represented by (1):

$$\begin{aligned} &L_{PatchGAN}(\{y_1, y_2, ...y_N\}, D(x, y)) = \\ &\mathbf{E}_{x,\hat{x}\in P_{data}(x,\hat{x})}[log(D(x, \hat{x}))]+ \\ &\mathbf{E}_{x\in P_{data}(x,\hat{x}),z\in P_z(z)}[log(1 - D(x, G(x, z)))], \end{aligned} \tag{1}$$

where $\mathbf{E}$ denotes the expectation operator. We observe that training the cGAN with the loss function of (1) alone fails to preserve the emotion and facial structure effectively. The rendered image often looks unrealistic and ghostly. To preserve the above non-biometric features at a high resolution, and generate an aesthetically pleasing de-identified face, we add a structural similarity loss term (denoted by $L_{ssim}(x, \hat{x})$ [18]) between the input face (x) and the generated face ($\hat{x}$) along with the loss function of (1). This loss term is computed by taking into consideration the similarities in the contrast, luminance and structure between x and $\hat{x}$, and is a combined measure of these three. If the luminance, contrast and structure similarities are respectively denoted by $l(x, \hat{x})$, $c(x, \hat{x})$, and $l(s, \hat{x})$, then the structural similarity index (SSIM) [19] between two input images x and $\hat{x}$ is mathematically defined as

$$SSIM(x, \hat{x}) = l(x, \hat{x})^{\alpha}.c(x, \hat{x})^{\beta}.s(x, \hat{x})^{\gamma}, \tag{2}$$

where,

$$l(x, \hat{x}) = \frac{2\mu_x\mu_{\hat{x}} + c_1}{\mu_x^2 + \mu_{\hat{x}}^2 + c_1}, \tag{3}$$

$$c(x, \hat{x}) = \frac{2\sigma_x\sigma_{\hat{x}} + c_2}{\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2}, \tag{4}$$

$$\text{and } s(x, \hat{x}) = \frac{\sigma_{x\hat{x}} + c_3}{\sigma_x\sigma_{\hat{x}} + c_3}. \tag{5}$$

In the above expressions, μ_x and $\mu_{\hat{x}}$ denote the average intensities of the inputs x, $\hat{x}$, σ_x^2 and $\sigma_{\hat{x}}^2$ refer to the variances in the intensities of x and $\hat{x}$, and $\sigma_x\sigma_{\hat{x}}$ denotes the covariance of x and $\hat{x}$. Constants c_1, c_2 and c_3 are defined as follows: c_1=$(k_1L)^2$, c_2=$(k_2L)^2$, c_3=$c_2/2$, k_1=0.01, k_2 = 0.03 where L is the dynamic range of the pixel values. For example, the value of L is 255 for 8 bit gray-level images. The structural similarity loss metric $L_{ssim}(x, \hat{x})$ is computed from (2) according to the following expression:

$$L_{ssim}(x, \hat{x}) = \frac{1}{2}(1 - SSIM(x, \hat{x})). \tag{6}$$

If L_{cGAN} denotes the loss function to be minimized for the generator and the discriminator networks without enforcing any condition, then L_{cGAN} can be mathematically represented as follows:

$$L_{cGAN}(x, \hat{x}, D(x, \hat{x})) = L_{PatchGAN}(x, \hat{x}) + \Phi L_{ssim}(x, \hat{x}). \tag{7}$$

where, Φ is a positive constant in the range [0,1]. Since there is a trade-off between effective de-identification and structural similarity preservation, the value of Φ must be chosen carefully to maintain a proper balance between the two. We observe that a value of Φ equal to 0.25 can help in achieving the desired balance.

B. Verificator Networks

As shown in Figure 2, we employ two different verificator networks to minimize identity similarity and maximize emotion similarity. These networks are discussed next.

1) Identity Verificator: VGG-16 Convolutional Neural Network [20] is employed as an identity verificator which takes as input the original image x and the generated image $\hat{x}$ and computes the similarity in the identity features between these images. Let us consider that there are M faces in the gallery set and the VGG-16 identity verificator network has M output nodes corresponding to each identity. If I_x denotes the ground-truth identity label for person x, (where x = 1, 2, 3, ..., M), then I_x is a feature vector of dimension M, where

$$I_x(i) = \begin{cases} 1 & \text{if } i = x \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The network is initially trained on the gallery set to learn the mapping function: $V : x \mapsto I_x$. Once this verificator is trained accurately, its weights are freezed, and next this trained

network is used in conjunction with the cGAN (discussed in Section III-A) to train the generator to produce de-identified faces.

Let us consider that at a particular iteration, the generator outputs image $\hat{x}$ for a certain input x. With reference to Figure 2, $\hat{x}$ is given as input to the identity verificator network along with the ground truth identity label I_x. As a response to input $\hat{x}$, the identity verificator outputs $V_{\hat{x}}$, (where, $V_{\hat{x}} = V(\hat{x})$, i.e., it is the vector formed by concatenating the values at the output nodes on presentation of input $\hat{x}$). In any de-identification task, the objective is to maximize the distance between I_x and predicted identity features $V_{\hat{x}}$. To do this, a loss function (denoted as $L_{contrastive}$) is computed at the output layer of the identity verificator. $L_{contrastive}$ is mathematically defined as follows:

$$L_{contrastive}(V_{\hat{x}}, I_x) = max(0, \alpha - ||I_x - V_{\hat{x}}||_2)^2, \quad (9)$$

where α is a regulatory parameter. The expression (9) ensures that the value of $L_{contrastive}$ lies between 0 and α.

2) Emotion Verificator: Similar to the identity verificator network (described in Section III-B1), we employ Xception Convolutional Neural Network [21] model to find the similarity of the emotion features between the input and the generated face images. Let S denote the mapping function learned by this network for mapping an image x to its corresponding emotion class represented by E_x. When the generated image $\hat{x}$ corresponding to input x is fed to the same Xception network, it generates an output vector $S(\hat{x})$. Now, if $L_{crossentropy}(S(\hat{x}), E_x)$ denotes the loss function associated with the emotion verificator, then a measure of dissimilarity between the two emotions is obtained by applying a cross-entropy loss function as shown in (10):

$$L_{crossentropy}(S(\hat{x}), E_x) = -(E_x log(S(\hat{x})) + (1 - E_x) log(1 - S(\hat{x})). \quad (10)$$

C. De-identification with Emotion and Facial Structure Preservation

Combining all the losses with weighted-sum, we get our final objective function

$$\begin{aligned} &L_{face}(x, G(x,z), D(x, G(x,z))) = \\ &\lambda_1 L_{cGAN}(x, G(x,z), D(x, G(x,z))) + \\ &\lambda_2 L_{contrastive}(V(G(x,z)), I_x) + \\ &\lambda_3 L_{crossentropy}(S(G(x,z)), E_x), \end{aligned} \quad (11)$$

where λ_1, λ_2, λ_3 are pre-defined constants ($0 \leq \lambda_1, \lambda_2, \lambda_1 \leq 1$), and the optimal generator output G^* is found using the minimax function as follows:

$$G^* = arg \min_G \max_D L_{face}(x, G, D). \quad (12)$$

IV. Experiments and Results

A. Data Set and System Description

Our algorithm has been implemented on a system with 64 GB RAM, one i9-18 core processor, and three GPUs, out of which one is Titan Xp with 12 GB RAM, 12 GB frame-buffer memory and 256 MB BAR1 memory, and the other two are GeForce GTX 1080 Ti with 11 GB RAM, 11 GB frame-buffer memory and 256 MB of BAR1 memory.

We evaluate our algorithm on two popularly used data sets, namely the Radboud Faces Database (RaFD) [22] and Kaggle emotion detection data set. The RaFD data consists of 8040 facial images corresponding to 67 subjects belonging to five categories: Caucasian male, Caucasian female, Caucasian kid male, Caucasian kid female, and Moroccan Dutch male. For each subject, the RaFD data preserves different emotions including anger, disgust, fear, happiness, sadness, neutral, etc. Each emotion is accompanied by three gaze directions and all the snapshots are taken from five different angles. The emotion verificator model described in Section III-B is trained with the RaFD data set to detect the following emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The learning rates used for training the Identity and Emotion verificator networks are set to 1e-6 and 1e-4, respectively.

To evaluate the effectiveness of our approach, we make a thorough experimental analysis and also perform a comparative evaluation with two popular state-of-the-art face de-identification methods (namely, PPGAN [1] and K-SameNet [12]). In Figure 3, we show some results obtained by applying the proposed de-identification algorithm on sample face images from the RaFD data. The first row in the figure represents the original face images, while the second row shows the corresponding de-identified faces.

It can be visually observed from the figure that in addition to obfuscation identity by a considerable extent, the EPD-Net also successfully preserves the emotion characteristics on the faces.

Before presenting further quantitative results, we define two metrics that have been used to evaluate the effectiveness of de-identification algorithms, namely, the de-identification rate and the switching de-identification rate. De-identification rate refers to the percentage of subjects for which identity obfuscation in the generated image is successful, whereas switching de-identification rate refers to the percentage of subjects for which the generated face's identity fails to match the identity of any subject in the gallery set. For effective de-identification, both de-identification and switching de-identification rates must be high. These metrics can be computed from the VGG-16 output score by considering a threshold parameter (say, μ), such that a verification score greater than μ implies de-identification (or, switching de-identification) is successful. We observe that a value of μ=0.35 results in a self de-identification rate of 69.90%, and a switching de-identification rate of 86.08%, which can be regarded as significantly good de-identification performance.

Next, we perform a comparative performance analysis of our approach with PPGAN [1] and K-SameNet [12] in terms of face recognition misclassification rate. To study the effect of using the SSIM loss term during training the de-identification model, in the same experiment we have also evaluated the performance of our approach without considering the SSIM

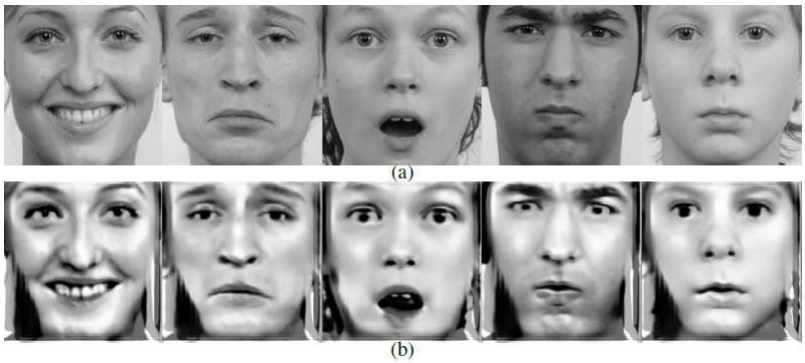

Fig. 3: (a) Sample Face Images from RaFD data, (b) corresponding de-identification results obtained by applying our approach

loss term. Results are shown in Figure 4 for ranks 1 to 5, by means of Cumulative Match Characteristic (CMC) curves. The prediction of the VGG-16 network is employed to plot these curves, i.e., if the class predicted by the VGG-16 network is different from the actual class, then misclassification occurs. It is imperative that the algorithm with higher misclassification rate should be regarded as a better de-identification algorithm. It is observed from the figure that the proposed method surpasses all the other approaches used in the comparative study in terms of misclassification rate for the different rank values. Although Rank 1 misclassification rate for each approach is similar, our approach outperforms the other methods as well as the approach without employing the SSIM loss function for each of the higher rank values. This proves that the identity obfuscation potential of the proposed approach is superior to the competing approaches.

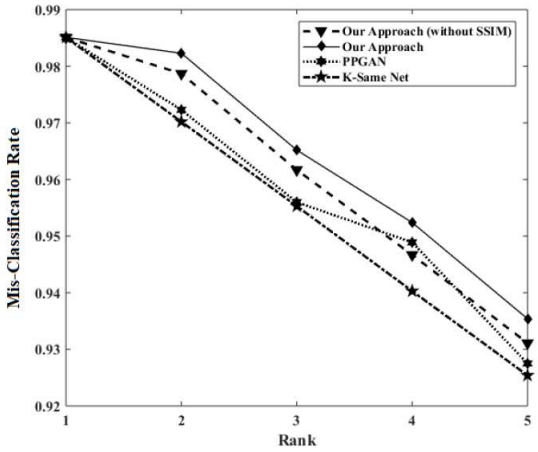

Fig. 4: Rank based mis-classification rate for our approach (with and without SSIM), PPGAN and K-Same Net

Next, we compare the performances of the different methods used in the previous experiment in terms of emotion preservation capability. Results are shown in Table I by means of averaged mean-squared error metric. For each test face image, we find the emotion features by passing the image through the Xception network, and next compute the squared difference between the two feature vectors. Finally, the mean of the square-differences from all the faces is presented in Table I. As expected, it is observed that due to the inclusion of the emotion verificator network in the complete architecture (refer to Figure 2), the generated image is able to preserve the emotion of the input image at a higher resolution compared to other state-of-the-art approaches, thereby resulting in a lower mean-squared error-rate. It can also be seen from the table that the mean-squared error of our approach increases by a factor of 0.0142 when SSIM loss is not included while training the GAN (refer to Equation 7). Thus, it can be concluded that apart from generating faces with proper facial features, the SSIM loss term also significantly helps in preserving the emotion present in the input face.

Method	Mean-Squared Error
Our Approach	0.0099
Our Approach (without SSIM)	0.0241
PPGAN	0.0257
K-same	0.0238

TABLE I: Emotion mean-squared error

In the next experiment, we study the extent to which structural similarity is preserved in case of the different approaches. Results are shown in Figure 5, by means of bar diagrams in which each bar represents the average SSIM index obtained after comparing the generated image with the original image for all the faces. It can be seen from the figure that, PPGAN [1] has marginally better performance (by a factor 0.001) compared to our approach. This is since while minimizing the identity similarity, the proposed network also obscures the color and contrast information present in the input face to a certain extent, which accounts for a slightly low SSIM score. Despite the marginally better performance of [1] over our approach in terms of SSIM score, its emotion preservation capability, as well as misclassification rates are inferior compared to our approach as already observed from Figure 4 and Table I. Thus, in terms of effective de-identification with emotion preservation, our method performs the best among

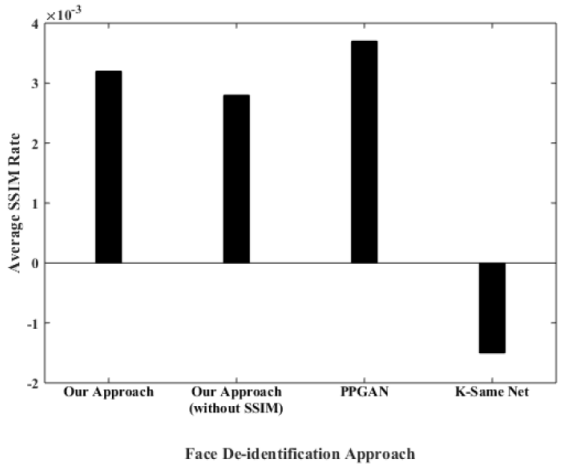

Fig. 5: Average SSIM rate given by the different face de-identification approaches

other existing learning-based approaches.

From the thorough experimental analysis, it can be concluded that the proposed approach performs de-identification effectively and has superior performance compared to other state-of-the-art face de-identification techniques. The present work can be easily extended to perform de-identification with the preservation of other non-biometric features like gender, skin-tone, race, etc., by adding separate verificator networks.

V. Conclusions and Future Work

In this paper, we propose a new architecture for face de-identification using conditional Generative Adversarial Network. The network successfully preserves emotion, but obscures the identifiable characteristics in a given face image. Extensive experimental evaluation show that our approach outperforms other state-of-the-art face de-identification algorithms in terms of identity obfuscation and emotion preservation capabilities. Although we have worked with emotion preservation only, our approach can be conveniently extended to retain any required non-biometric information along with simultaneous removal of biometric features. In future, the work can be extended to perform face de-identification in videos. The proposed de-identification technique will have significant use in commercial applications, e.g., study of customer emotional feedback for products, whereby, maintaining privacy, behavioral-based insider threat detection, and several others.

Acknowledgements

The authors also express their sincere gratitude to NVIDIA for supporting their research with a Titan Xp GPU.

References

[1] Y. Wu, F. Yang, and H. Ling, "Privacy-protective-gan for face de-identification," *arXiv preprint arXiv:1806.08906*, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[3] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[4] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 151–166, Springer, 2017.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[6] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view.," in *ICCV*, pp. 2373–2380, 2009.

[7] H. Sohn, W. De Neve, and Y. M. Ro, "Privacy protection in video surveillance systems: Analysis of subband-adaptive scrambling in jpeg xr," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 2, pp. 170–177, 2011.

[8] L. E. Schulz and E. B. Bonawitz, "Serious fun: preschoolers engage in more exploratory play when evidence is confounded.," *Developmental psychology*, vol. 43, no. 4, p. 1045, 2007.

[9] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," in *null*, p. 161, IEEE, 2006.

[10] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *International Workshop on Privacy Enhancing Technologies*, pp. 227–242, Springer, 2005.

[11] R. Gross, L. Sweeney, F. De La Torre, and S. Baker, "Semi-supervised learning of multi-factor models for face de-identification," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[12] 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.

[13] P. Korshunov and T. Ebrahimi, "Using warping for privacy protection in video surveillance," in *Digital Signal Processing (DSP), 2013 18th International Conference on*, pp. 1–6, IEEE, 2013.

[14] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299–310, 2011.

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition.," in *bmvc*, vol. 1, p. 6, 2015.

[16] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

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

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 13, no. 4, pp. 600–612, 2004.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

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