DeepPrivacy: For Face De-identification

Avinash Shanker
University of Texas Arlington
avinash.shanker@mavs.uta.edu

Puneeth Gopi
University of Texas Arlington
puneeth.gopi@mavs.uta.edu

Abstract

People get their pictures taken on Google street view or when they pass by a journalist giving a report in public places, questions concerning the privacy of people visible arises. Among those image sources exposed to the public with or without our awareness, a considerable number of them contain our identity especially the bio-metric information. To solve these issue, we implement a model which can change a persons face to look like a completely different person, thus protecting their privacy. The model is trained with combination of generative adversarial networks(GAN) and autoencoders. We ensure anonymity by synthesizing GAN generated images. The generated faces are used to de-identify subjects in images or video, while preserving non-identity-related aspects of the data and consequently enabling data utilization [12].

1. Introduction

Beneficial from the blooming development of media and network techniques that makes huge amount of images more approachable, image analysis techniques bear its prosperity in the past decade and brings unprecedented convenience to our daily life. Among those image sources exposed to the public with or without our awareness, a considerable number of them contain our identity especially the bio-metric information [4]. Not only will the unprotected exposing cause the leak of privacy, the common approaches of protection like blurring and pixelization may also not be satisfied in thwarting face recognition software.The other extreme side is that we simply mask identity area off, which is perfect for identity removal. But it causes serious loss of data utility in application of visual understanding as the scene information is changed with objects removal [6]. Thus it is critically important to build a framework that can properly de-identify the privacy information from the image while keeping its utility at the same time [3].

Specially for the face de-identification problem, the dilemma is that on the one hand, we want the de-identified image to look as different as possible from the original image to ensure the removal of identity; on the other hand, we expect the de-identified image to retain as much structural information in the original image as possible so that the image utility remains. The Generative Adversarial Networks (GANs) provide an inspiring framework on generating sharp and realistic natural image samples via adversarial training. GAN can be naturally used for the face de-identification as it can generate new samples from the gallery following original input data distribution. We use this generated image to super impose on the target image or video thus changing the face of the target.

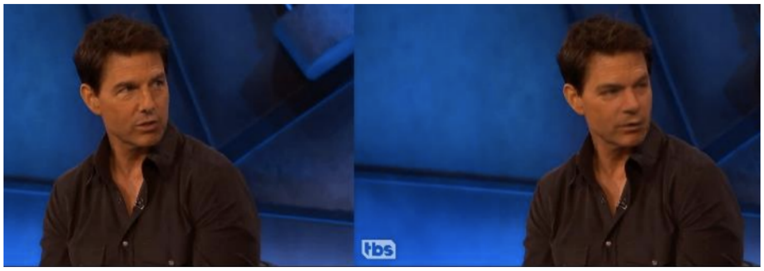

Figure 1. Image of Tom Cruise De-identified

2. Methodology

The main objective of the architecture is to provide image to image translation from two different faces. This involves generating and superimposing faces which can be done using autoencoders and generative adversarial networks.[1].

2.1. Autoencoders

Autoencoders are artificial neural networks(ANNs) which consists of an encoder and a decoder. The input goes to the encoder which results in a lower dimensional representation of the image as seen in the image. The encoder is constructed such that the number of nodes in progressing layers decreases and therefore results in capturing the latent features. This middle layer is latent space representation of the input image [2].

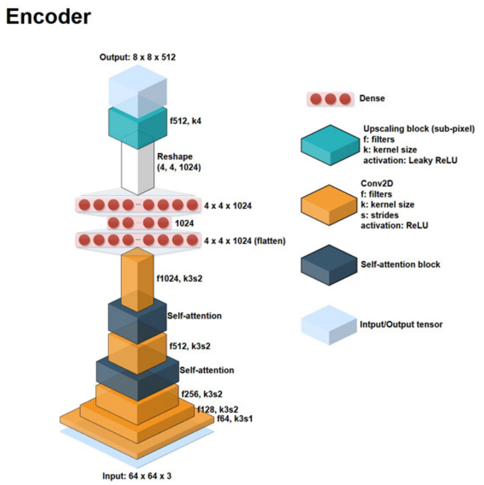

Figure 2. Encoder Architechture

The decoder then takes the output of the encoder and reconstructs the input image using the latent features. This is a lossy process, but allows the network to learn the latent features, specifically the facial features which we will need for replacement.

Autoencoders are utilized in deepfakes by training the same encoder on all possible input faces and learn the specific latent features. The autoencoders are trained so that all faces share the same encoder but difference decoders. This ensures that during the training phase, the encoder is able to learn the facial features and the specific decoders are able to construct the respective faces.

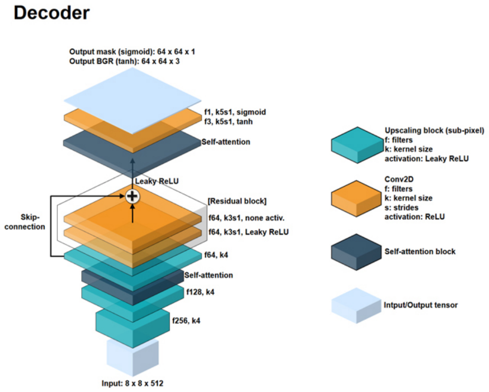

Figure 3. Decoder Architecture

2.2. Generative Adversarial Networks (GANs)

Generative Adversarial Networks or GANs for short, are a specific type of deep learning network consisting of two networks, namely the generator network and the discriminator network [9]. The role of the generator network is to generate new data instances through a deep learning network , for example, a U-network which utilizes convolution followed by de-convolution with the goal of generating passable target faces without being caught [8].

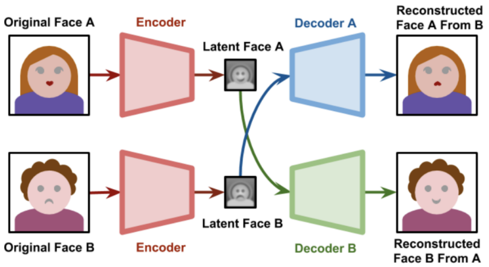

Figure 4. Decoder Architecture

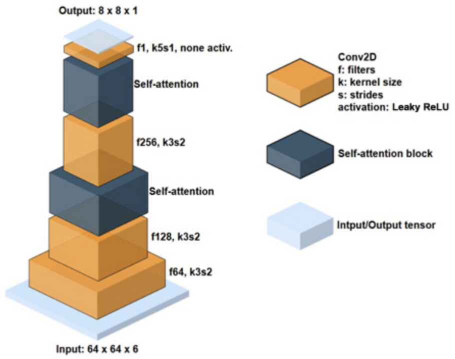

Figure 5. Discriminator network then evaluates the authenticity of output images from the generator network and decides whether each instance of data it reviews, belongs to the actual training data set or not

In terms of deepfakes, the generator takes in face of person A and returns a B-like image as part of a semi-supervised learning process. This generated B like image is fed into the discriminator alongside a stream of real B images taken from the actual dataset [7]. The discriminator takes in both the real and fake images and returns the probabilities, with 1 representing a true image and 0 representing a fake image. The generator network and discriminator network compete against each other to minimize the loss and moving towards the goal of generating images indistinguishable from real images.

The discriminator network serves as the counterpart to the autoencoder to generate the output images.

3. Optimized techniques and deep learning

3.1. Extract faces from videos

When extracting the faces from the input video for training the architecture, it is important to recognize the importance of face alignment and cropping. Using a sophisticated CNN such as Multi-task Cascaded Convolutional Networks (MTCNN) allows us to extract images within a set number of frames and utilize VGGFace [10] for removal of any noise so as to crop the face as much as possible which allows the deep learning network to identify the facial features better and allow for more precision, thereby looking smoother in the final output video.

3.2. Pre-process images for facial features

Face tracking and alignment with Kalman Filter in combination with MTCNN allows for stable detection of the face and consistent face alignment. The Kalman filter smoothens the boundaries of the frame positions and helps to remove jitter on the resulting face swap image [13].

VGGFace perceptual loss helps in improving direction of eyeballs to be more realistic and consistent with input face. It also smoothes out artifacts in the segmentation mask, resulting higher output quality.

Attention mask predicts an attention mask that helps on handling occlusion, eliminating artifacts, and producing natrual skin tone.

3.3. Combination of GANs and Autoencoders

A combination of a generative adversarial network and autoencoders is tried to take the best of both networks. In this combined architecture called a Self-Attention Generative Adversarial Network(SAGAN), we utilize the autoencoder as the generative part of the GAN, thereby utilizing the latent space representation for more efficient generation of images. This would be the final architecture for training on the final image dataset generated from the input videos. The discriminator network serves as the counterpart to the autoencoder to generate the output images. building a combined architecture such as SAGAN which utilizes both an autoencoder network as well as a GAN structure allows the combined model to generalize better and take the best of both networks when training [11].

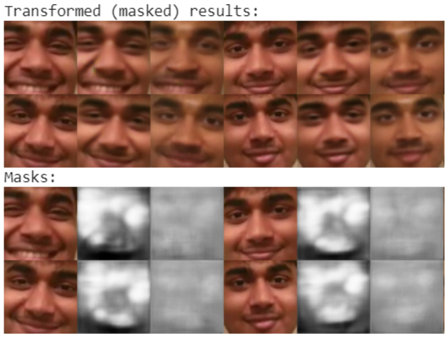

Figure 6. Illustration of face de-identification

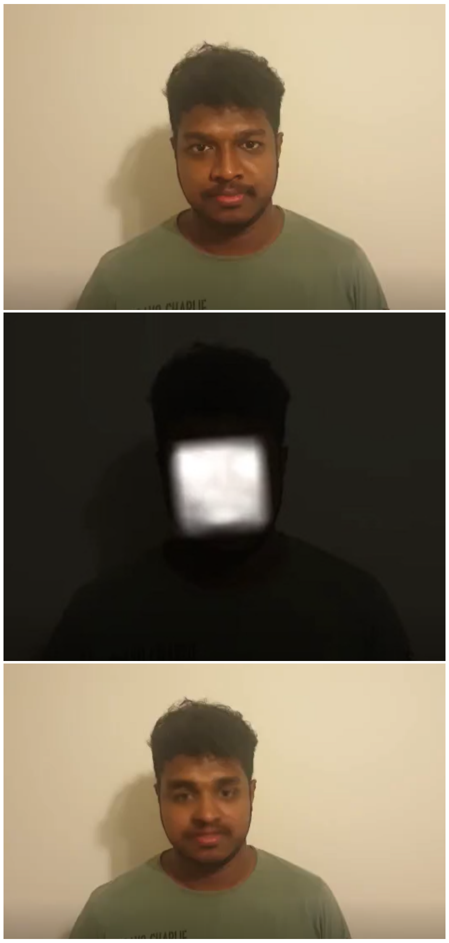

Figure 7. Original image, Masking, de-identified image

4. Results

Fig 6 Shows the visualization of correspondence between original image and de-identified images. The bottom two rows shows the original image being masked. The top two rows shows the transition from the original image to the de-identified image.

In figure 7, the top image is the original image that needs to be de-identified. The middle image shows the region where the mask is being applied. The bottom image is the final de-identified output which is the combination of original image, super-imposed with the GAN generated image. The face in this image is clearly different from the original image [5].

Since the model uses resnet50 is computationally expensive, the model was trained progressively till 20k

Loss	Percent
Adversarial loss	7.49%
Reconstruction loss	30.10%
Edge loss	27.22%
Perceptual loss	16.41%

Table 1. Generator losses after running for 20,000 epochs

epochs(20K epoch took 9hrs) for which we got the following satisfactory results. The loss obtained is shown in table1. If trained further we would have obtained better loss values.

5. Conclusion

In this paper, we present a face de-identification framework, DeepPrivacy, which generates deidentified output according to a single input. We explicitly integrate the de-identification metric into the objective function to ensure the privacy protection. Meanwhile, we try to preserve visual similarity as much as possible to retain data utility by adding a regulator. In the results, we demonstrate the effectiveness of proposed method in terms of privacy protection, utility preservation, and visual similarity.

6. Individual Contributions

Puneeth - worked on generating GAN image and training.

Avinash - Worked on MTCNN to extract face images and binary mask datasets.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 1

[2] J. H. Bappy, C. Simons, L. Nataraj, B. Manjunath, and A. K. Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 1

[3] 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*, pages 1–10. ACM, 2000. 1

[4] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328. IEEE, 2017. 1

[5] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[6] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008. 1

[7] R. Chesney and D. K. Citron. Deep fakes: a looming challenge for privacy, democracy, and national security. 2018. 2

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2019. 2

[9] H. Hukkelås, R. Mester, and F. Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*, pages 565–578. Springer, 2019. 2

[10] rcmalli. keras-vggface. https://github.com/rcmalli/keras-vggface/, 2019. 3

[11] shaoanlu. faceswap-gan. https://github.com/shaoanlu/faceswap-GAN, 2019. 3

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

[13] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427. IEEE, 2017. 3