**Face Anonymization for privacy preservation using Generative Adversarial Networks**

*Shaik Mahaboob Subhani Sameer, Kanagala Bhanu Murthi,*

*DS Bhupal Naik, Asst.professor*

*Computer Science & Engineering Department*

*Vignan’s Foundation for Science Technology & Research Deemed to be University, Guntur*

**ABSTRACT**

As Computer Vision technology is getting advanced day by day, users are also captivated to use every technology for their routine. Generally, we use technology for having task done in a limited and appropriate time without keeping much effort. But there is an increasing concern of privacy in the society. Its everyone’s responsibility to have privacy from illegal technocrats. Usually, we use cameras or video recorders to get protected from illegal activities in public areas. It is high essential for not only get protected but also to have concern about their own privacy. In order to figure out the above concern we have proposed and developed a prototype i.e., Face Anonymization based on Generative Adversarial Networks. This keeps track of anonymized faces from original or real faces. In this paper we have defined different versions of Generative Adversarial Networks which are used to anonymize faces and also different from previous methods such as blurring, masking, etc. We have used a high-quality dataset which were aligned, scaled and centred faces. We have compared the results of each method and used evaluation of the quality.

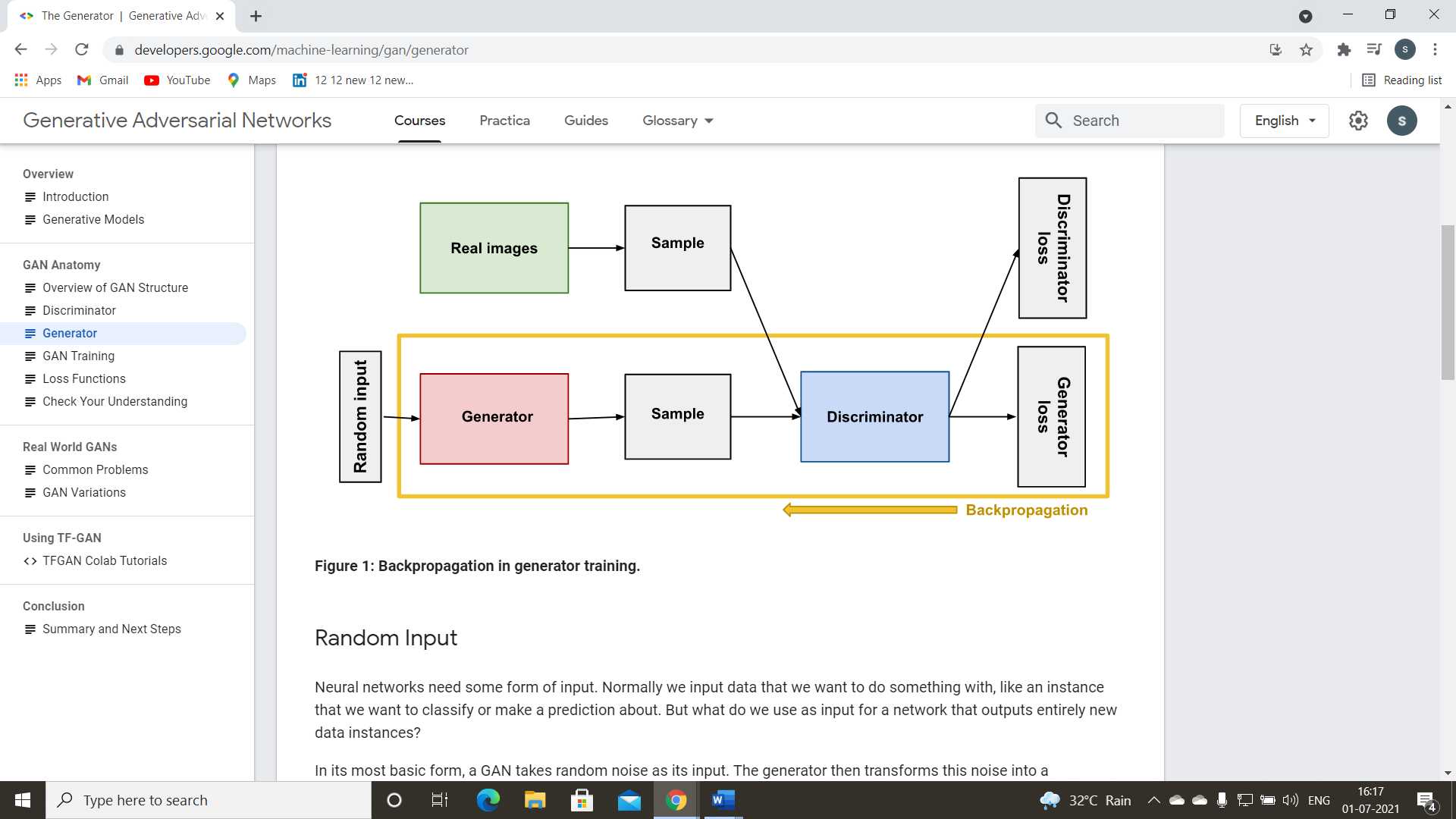
**Keywords:** Computer Vision, technocrats, Face Anonymization, Generative Adversarial Networks

1. **INTRODUCTION:**

Privacy is not something that we merely entitled to, it’s an absolute prerequisite. Without privacy there was no point in being an individual. Face Anonymization is a technique which is used to anonymize the original faces so that it reduces concerns regarding individual’s privacy. In this regard we need to be able to identify the faces in the videos or images so that we can anonymize those faces. After defining a dataset regarding the images to be anonymize, we can able to perform models on the dataset and get the results to be compared. We need to gather the images consists of faces which are aligned, scaled and centred. We can prefer existing datasets such as Celeba-dataset which consists of celebrity faces to get anonymized. We can have such real time datasets to perform the anonymization.

We have conventional methods such as blurring, pixeling, masking etc to get image anonymized but they add noise to the images at pixel-level. Due to this it significantly decreases the utility and level of privacy whereas the proposed model increases the utility and privacy. Generative Adversarial Networks consists of two parts such as Discriminator and Generator. Discriminator is like a classifier which distinguishes real faces from anonymized faces. Generator which creates anonymized faces by consolidating feedback produced by Discriminator. The training of Generator needs tighter amalgamation between generator and discriminator than the training of discriminator needed. For training of Generator, GAN requires random input, generator network, discriminator network, discriminator output, generator loss.

The generator network converts random input into a meaningful output while discriminator network just categorises the produced faces. Generator loss which castigated the generator due to the failure of fooling the discriminator.

 In order to reduce the error or loss in the output, we need to assign networks weights during the training of neural network. Actually, generator indirectly attached to

**Figure 1:** Overview of GAN Structure

the loss which we are demanding to affect. The additional block of network is essential to be involved in backpropagation which regulates each weight in the right direction through the calculation of weights influence in the output. But the influence of a generator weight turned on to the influence of discriminator weights it sustains into. Hence backpropagation begins at output and goes along back through discriminator into the generator.

The arrangement to compose training of generator kicks off with sample random uproar and generates generator output from sampled random noise. Then we'll get the discriminator in-case genuine or artificial classification for generator output. We can decide the loss from discriminator classification and backpropagate by means of both the discriminator and generator to gain gradients. We can use those gradients to alter only the generator weights.

Overall, the generator and the discriminator have diverse training processes. Training of GANs proceeds in alternating periods that is both discriminator and generator trains for one or more epochs. Those steps are repeated to carry on the training of both generator and discriminator networks.

1. **LITERATURE SURVEY:**

Privacy preservation of confront images has been studied from two aspects. ace images explain a lot of ancillary databases for which the user will not have consented. So as to guard such ancillary databases (soft biometrics), researchers have proposed various diverse methodologies.

In 2000, Boyle developed an algorithm to blur and pixelate the images in the video. In 2005, Newton have developed an algorithm for confront de-identification in video surveillance so as the confront recognition fails while maintaining other facial details. In 2006, Dismal have displayed that distorting image via blurring and pixelation technique delivers unpleasant results. Several researchers have additionally worked on the privacy of sentimental biometrics. In 2014, Othman and Ross have proposed attribute privacy maintaining technique, in which the soft biometrics attribute as an example gender is “flipped” while maintaining the identity for confront recognition. As an extension of this work, Mirjalili and Ross have proposed Delaunay triangulation, and convolutional autoencoders based methods to flip gender databases while maintaining confront identity in 2017.

In 2011, Suo have presented an image fusion framework in which the template of opposite gender confront image is taken for fusion with the candidate image while maintaining confront identity. In 2015, Jourabloo have developed an algorithm for de-identification of confront image while maintaining other attributes. For attribute preservation, it utilizes k images (motivated by k-Same) which shares the same attributes for fusion. In 2015, Sim and Zhang have proposed a technique which independently controls the identity change and maintains the other facial attributes. To anonymize the facial attributes, Rozsa have proposed deep learning-based exemplar for facial attribute prediction in 2017. As an extension of their employment Rozsa have utilized FFA and adversarial images are produced in which a facial attribute is flipped in 2016. They have observed that the few attributes are impacted while flipping the targeted attribute. For instance, while changing the ‘wearing lipstick’ attribute, other attributes that is ‘attractive’ and ‘heavy makeup’ are additionally flipped. While anonymizing attributes of faces, over there should be no disagreement between unique and anonymized faces. Perceptively anonymizing few and retaining few attributes require a “control” mechanism. For example, gender and expression can be anonymized while retaining chase and eye colour and other attributes that is attractiveness and hair colour may be in “do not care” condition.

The major limitation of those methods is that they're doing not address the most important two challenges mentioned above. Already algorithmic program primarily depends upon a candidate image which is fused with the 1st image. Rozsa have addressed this reissue but due to lack of any control mechanism, the other attributes are additionally suppressed, while suppressing one attribute.

1. **PROBLEM STATEMENT:**

we require a principled cheer to ‘anonymize’ one’s videos/images. Already anonymization methods add grievous down sampling or image masking, additionally as more sophisticated image processing strategies utilizing image segmentation. But such strategies remove scene explains within the images/videos in a try to save privacy, they're supported heuristics instead of being learned, and over there is no assure that they're optimal for privacy-protection. Moreover, they'll damage the following visual recognition performance cheers to loss of information Thus, a key challenge is generating an advance which will concurrently anonymize videos, while ensuring that the anonymization doesn't negatively bear on recognition performance. In this paper, we suggest a totally original principled advance for acquiring the video/image anonymizer.

We use an adversarial training strategy; i.e., we exemplar the training process as a fight between two competing systems: (1) a video/image anonymizer that modifies the 1st video to get rid of privacy-sensitive databases while maintaining scene notion performance, and (2) a discriminator that extracts privacy-sensitive databases from such anonymized videos. We use confront identity because the representative personal databases since confront is one of the strongest cues to deduce a person’s identity and use action detection because the representative scene notion task.

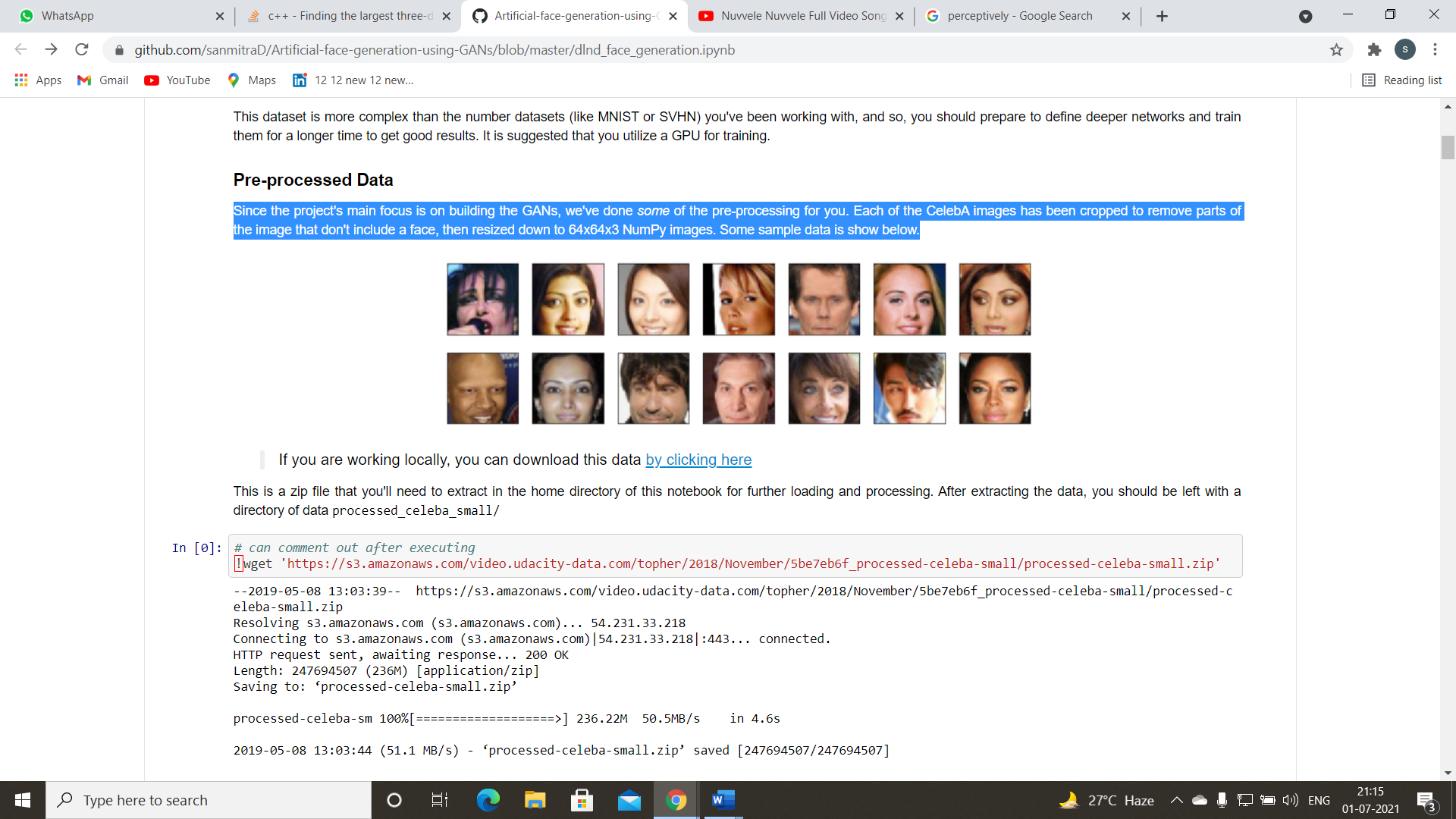
1. **PROPOSED WORK:**

In this paper, we'll clarify and direct a Deep Convolution Generative Adversarial Network (DCGAN) on a dataset of faces. Our goal is to get a generator network to produce new images of faces that see as realistic as possible.

This process will be smashed down into a series of missions from loading in databases to defining and training adversarial networks. Finally, we can imagine the results of your trained Generator to see how it performs; your produced samples should see love fairly realistic faces with limited amounts of noise.

**Dataset:** We will be using the CelebFaces Attributes to train your adversarial networks. This dataset is more complex than the number datasets (like MNIST or SVHN) we have been working with, and so, you should prepare to define deeper networks and train them for a longer time to get good results. It is suggested that you utilize a GPU for training.

**Pre-processed Data:** Since our main focus is on creating the GANs, we have done Several of the pre-processing for you. Each of the CelebA images has been cropped to remove parts of the image that don't add a face, then resized down to 64x64x3 NumPy images. Several sample databases is reveal underneath.

****

**Figure 2:** Sample images from Dataset

**Visualize the Dataset:** The CelebA dataset contains more than 200, 000 celebrity images with annotations. Since we're going to be producing faces, you won't require the annotations, we will only need the images. Note that these are color images with 3 color channels each.

**Loading the Dataset:**

There are a few other steps that we will need to transform this data and create a Data Loader

The following are the requirements which satisfies the data loader function.

* Our images should be square, Tensor images of size image size x image size in the x and y dimension.
* Our function should return a Data Loader that shuffles and batches these Tensor images.
* We can settle on any fair batch size parameter
* Our image size must be 32. Resizing the databases to a smaller size will make for faster training, while still generating convincing images of faces

Next, we can view some images. we should seen square images of somewhat-centered faces.

Notice that we'll require to transform the Tensor images into a NumPy type and transpose the dimensions to correctly show an image, but it will not be perfect.

We require to do a little of pre-processing; we know that the output of a tanh activated generator will acquire pixel values in a range from -1 to 1, and so, we require to rescale our training images to a range of -1 to 1. (Right now, they're in a range from 0-1.)

**Defining a model:**

A GAN consists of two adversarial networks such as Discriminator and Generator.

**Discriminator:** Our first task will be to define the discriminator. This is a convolutional classifier like we have built before, only without any maxpooling layers. To deal with this complex data, it's suggested we use a deep network with normalization. we are also allowed to create any helper functions that may be useful.

The following are the requirements which satisfies the Discriminator

* The inputs to the discriminator are 32x32x3 tensor images
* The output should be a single charge that will point to in-case a delivered image is genuine or artificial

**Generator:**

The generator should up sample an input and produce a new image of the same size as our training statistics 32x32x3. This should be mostly transpose convolutional layers with normalization applied to the outputs.

The following are the requirements which satisfies the Generator

* The inputs to the generator are vectors of several length z size
* The output should be an image of shape 32x32x3

**Initialization of weights:** To assist our models converge, we should initialize the weights of the convolutional and linear layers in our model.

All weights were initialized from a zero-centered Usual distribution with classic deviation 0.02. So, our next mission will be to clarify a weight initialization function.

The following are the requirements which satisfies the completion of weight initialization

* This should initialize only convolutional and linear layers
* Initialize the weights to a normal distribution, centered around 0, with a standard deviation of 0.02.
* The bias terms, if they exist, may be left alone or set to 0.

Now we need to calculate the losses for both types of adversarial networks.

**Discriminator Losses:**

For the discriminator, the perfect loss is the sum of the losses for genuine and artificial images, **d\_loss = d\_genuine\_loss + d\_artificial\_loss.**

Remember that we wants the discriminator to output 1 for genuine images and 0 for artificial images, so we require to establish the losses to mediate that.

**Generator Losses:**

The generator loss will gaze lookalike only with flipped labels. The generator's goal is to get the discriminator to believe its produced images are real.

**we may choose to use either cross entropy or a least squares error loss to complete the genuine\_loss and artificial\_loss functions.**

**Training:** Training will signify alternating between training the discriminator and the generator. We'll use our operates genuine loss and artificial loss to assist us to decide the discriminator losses.

* we should direct the discriminator by alternating on genuine and artificial images.
* Then the generator, which attempts to conjuring trick the discriminator and should have an objecting loss function.

We can set our number of training epochs and train our GAN.

GAN training proceeds in reciprocating periods:

1. The discriminator trains for one or more epochs.

2. The generator trains for one or more epochs.

3. Repeat steps 1 and 2 to persevere to direct the generator and discriminator networks.

1. **RESULT AND ANALYSIS:**

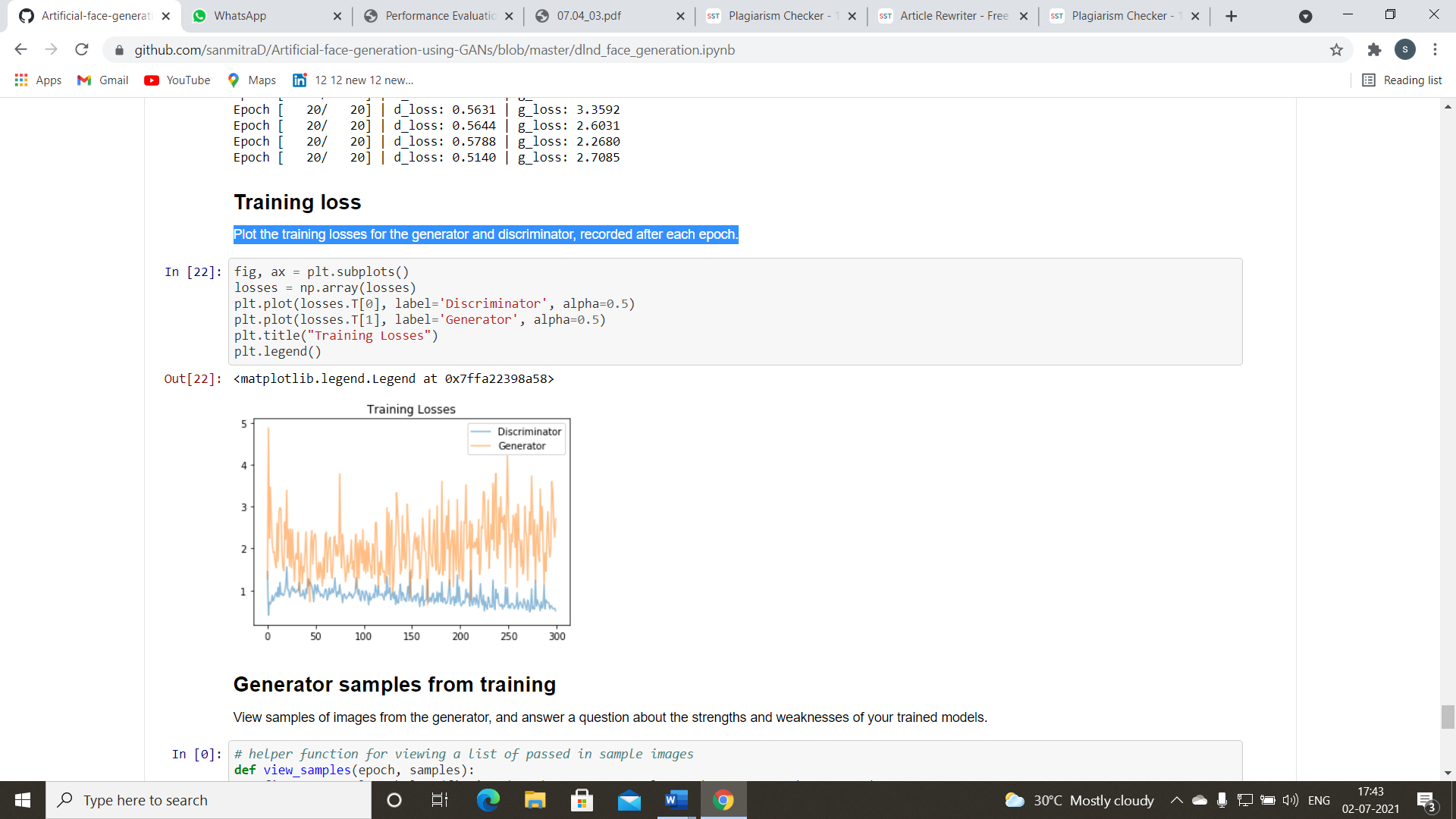
The factors which we need to consider are as follows:

* The dataset is biased; it's composed of "celebrity" faces that are mostly white.
* Model size; larger models have the opportunity to learn more features in a data feature space
* Optimization strategy; optimizers and number of epochs affect your final result

In this paper, we may say that the dataset is lightly biased. But, not to a large extent. we can understand it from the sample images printed at the top of the paper. That might be the reason why most of the generated images are mostly white.

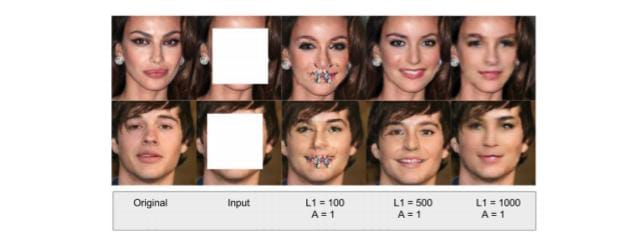
We even can observe that adding an extra layer and increasing the depths of convolution layers made the losses to converge to a better loss. The larger models have the opportunity to learn more features in the data feature space. Using Adam as optimiser yielded good results. The number of epochs also effected the final result. The training for larger number of epochs, making the size of model larger and using a more diverse dataset might improve this model.

**Training Loss:**

Plotting the training losses for the generator and discriminator, recorded after each epoch

**Figure 3:** Comparison: Training Losses

**Comparison** Both ’Adversarial Loss’ and ’Recreation Loss’ are necessary as they both serve their own purpose. Recreation loss helps keep the context and Adversarial loss helps generate realistic and good resolution images. But both the loss functions have their side-effects. Too much ’Adversarial loss’ leads to artifacts and too much ’Recreation loss’ leads to blurry images. Thus, it’s a trade-off between two of them. With better evaluation metric for quality of generated faces that sweet spot can be achieved.



**Figure 4:** Comparison: Joint Loss variations.

1. **CONCLUSION:**

We have designed a prototype to anonymize faces which is better than traditional anonymization techniques like blurring, pixelation and cutting of faces. Unlike traditional anonymization techniques our method produces anonymized faces that looks realistic and can be used for artificial intelligence applications thus it preserves the utility of the data. Through our experiments, we have demonstrated the efficacy and potential of using Generative Adversarial Network to anonymize faces with 80% Learning rate.

In particular, we have shown that formulating the task as a face completion task can yield anonymized faces that are aesthetically-pleasing and real looking than the orthodox face anonymization techniques and thus shows much promise for further development. We designed a joint loss function as an objective function to successfully train the face anonymizer neural network. We also designed a full end-to-end pipeline to anonymize multiple faces in a frame smoothly. As an additional contribution, we constructed a high-quality face data which contains aligned, scaled and centered faces. We have also identified several avenues for improving our current system.

**REFERENCES:**

[1] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian J. Goodfellow, Jean Pouget-Abadie. Generative adversarial nets. 2014.

[2] Soumith Chintala Alec Radford, Luke Metz. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015.

[3] http://www.gamingtechlaw.com/2016/10/privacy-gdpr-artificial-intelligence. html.

[4] Bradley Malin Ralph Gross, Edoardo Airoldi and Latanya Sweeney. Integrating utility into face de-identification.

[5] Reza Shokri Richard McPherson and Vitaly Shmatikov. Defeating image obfuscation with deep learning. 2016.

[6] Samuli Laine Jaakko Lehtinen Tero Karras, Timo Aila. Progressive growing of gans for improved quality, stability, and variation. 2017.

[7] Artur Luczak Volodymyr Turchenko, Eric Chalmers. A deep convolutional auto-encoder with pooling - unpooling layers in caffe. 2017.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[9] H. Su J. Krause S. Satheesh S. Ma-Z. Huang A. Karpathy A. Khosla M. Bernstein A. C. Berg O. Russakovsky, J. Deng and L. Fei-Fei. ImageNet large scale visual recognition challenge. 2015.

[10] A. Graves D. Jimenez Rezende K. Gregor, I. Danihelka and D. Wierstra. Draw: A recurrent neural network for image generation. 2015.

[11] W. Zaremba V. Cheung A. Radford T. Salimans, I. J. Goodfellow and X. Chen. Improved techniques for training gans. 2015.

[12] Jurgen Schmidhuber. Deep learning in neural networks: An overview. 2014