Centro de Informática
UFPE

Pedro Kempter Brant

An Analysis of CIAGAN Use on Face Expression Preservation

Federal University of Pernambuco
secgrad@cin.ufpe.br
www.cin.ufpe.br/~graduacao

Recife
May, 2022

Pedro Kempter Brant

An Analysis of CIAGAN Use on Face Expression Preservation

A B.Sc. Dissertation presented to the Center of Informatics of Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Bachelor in Computer Engineering.

Concentration Area: *Computational Intelligence*
Advisor: *Veronica Teichrieb*
Co-Advisor: *Lucas Silva Figueiredo*
Willams de Lima Costa

Recife
May, 2022

FICHA

BANCA

I dedicate this dissertation to my family, my friends and my girlfriend Mariana.

ACKNOWLEDGEMENTS

Gostaria de agradecer aos meus orientadores, Will, Lucas e VT por me guiarem em todo o processo de pesquisa e escrita do projeto. Também quero estender esse agradecimento a todos que fazem o Voxar Labs ser o que é, um laboratório de ponta que me acolheu e me fez um pesquisador e ser humano melhor.

Sobre a federal e o centro de informática, agradeço aos funcionários, professores, colegas e amigos que, desde 2016, fizeram parte da minha formação profissional. Foram muitos momentos difíceis e desafiadores, mas aprendi e cresci muito com todo o processo. Também estendo aos amigos e professores do CFC.

A minha família, agradeço ao meu pai, Maurício, meu avô Moacyr e meu tio avô Nelson por me inspirarem a fazer engenharia. Agradeço a minha mãe, Dominique, pelo carinho e apoio de sempre e aos meus avós Pedro e Carin, que, embora não tenha conhecido, me ensinaram ser uma pessoa honesta e trabalhadora. Agradeço a minha irmã, Nana, a Diego, Marie e Hase pelo companheirismo durante as noites de fazendo trabalhos da faculdade e, finalmente, a minha avó Madalena por todos os momentos, especialmente os sábados que passamos juntos.

Finalmente, agradeço a minha namorada, Mariana, por todo apoio, carinho, respeito e amizade que estamos construindo juntos.

“Somewhere, something incredible is waiting to be known."

–Carl Sagan

ABSTRACT

Data privacy is an increasing concern in the Big Data era. For instance, deep learning models, such as Facial Expression Recognition (FER), need personal information as inputs, precisely photos or videos, to extract emotion expression. Although these systems demand sensitive data, these applications do not have to violate people's identity and privacy. Anonymization techniques can be applied before the face analysis. Thus, the desired application can still process the essential data while the user's ID is not compromised. To preserve the user's ID, an emerging group of deep neural networks called Generative Adversarial Networks can, among other applications, generate entirely new faces from a given photo. This work analyzes a GAN anonymization technique, CIAGAN, precisely on facial expression preservation. Our experiments use a facial expression database to be anonymized, RAF-DB, and a FER network, DAN, to evaluate CIAGAN's ability to preserve the original face expression. The results show that only 32.45% of the expressions were preserved when each face was classified pre and post anonymization. Moreover, we trained a CIAGAN's model using Mediapipe's facial landmarks instead of those recognized by Dlib, and compared qualitatively with the pre-trained model. Although not fully refined, the trained model gives insights to improve the CIAGAN on preserving facial expressions and raises a discussion on exploring new guiding information on Conditional GANs to preserve useful information after the anonymization process.

Keywords: GANs. Expression Preservation. Anonymization.

RESUMO

A privacidade dos dados é uma preocupação crescente na era do Big Data. Por exemplo, modelos de aprendizado profundo, como o Facial Expression Recognition (FER), precisam de informações pessoais como entradas, precisamente fotos ou vídeos, para extrair emoção. Embora esses sistemas exijam dados confidenciais, essas aplicações não precisam violar a identidade e a privacidade das pessoas. Técnicas de anonimização podem ser aplicadas antes da análise da face. Assim, a aplicação desejada ainda pode processar os dados essenciais enquanto o ID do usuário não é comprometido. Para preservar o ID do usuário, um grupo emergente de redes neurais profundas chamado Generative Adversarial Networks pode, entre outras aplicações, gerar rostos inteiramente novos a partir de uma determinada foto. Este trabalho analisa uma técnica de anonimização GAN, CIAGAN, especificamente no que diz respeito a preservação da emoção facial. Nossos experimentos usam um banco de dados de expressão facial a ser anonimizado, RAF-DB, e uma rede FER, DAN, para avaliar a capacidade do CIAGAN em preservar a expressão facial original. Os resultados mostram que apenas 32,45% das expressões foram preservadas, comparando cada face classificada pré e pós anonimizada. Além disso, treinamos um modelo do CIAGAN usando os pontos de referência faciais do Mediapipe em vez dos reconhecidos pelo Dlib e comparamos qualitativamente com o modelo pré-treinado. Embora não totalmente refinado, o modelo treinado fornece insights para melhorar o CIAGAN na preservação de expressões faciais e levanta uma discussão sobre a exploração de novas informações orientadoras sobre GANs condicionais para preservar informações úteis após o processo de anonimização.

Palavras-chave: GANs. Preservação da Expressão. Anonimização.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI	Artificial Intelligence
CDF	Cumulative Distribution Function
CelebA	Large-scale CelebFaces Attributes
CGAN	Conditional Generative Adversarial Network
CIAGAN	Conditional Identity Anonymization Generative Adversarial Network
DAN	Distract your Attention Network
DNN	Deep Neural Networks
EDSP	European Data Protection Supervisor
EU	Europe Union
FER	Facial Expression Recognition
GAN	Generative Adversarial Network
GDPR	Europe's General Data Protection Regulation
LGPD	Lei Geral de Proteção de Dados
LSLF	Least-Squares Loss Function
ML	Machine Learning
MMDA	Multimodal Discriminant Analysis
RAF-DB	Real-world Affective Faces Database

LIST OF SYMBOLS

D	Discriminator Function
G	Generator Function
x	Data
y	Auxiliary information
z	Random Variable

CONTENTS

1

INTRODUCTION

1.1 MOTIVATION

The big data era has brought many possibilities to process and explore the vast amount of data collected. However, it also raises the concern related to security, and privacy Khanan *et al.* (2019). Thus, projects should be aware of those concerns and work on protecting this data, especially those that use sensitive personal content such as people's faces and emotions.

Projects such as smart cities are revolutionizing cities; the need to understand human behavior is overgrowing to allow intelligent systems and decision-makers to understand how people relate to smart spaces. In addition, given how these systems rely heavily on personal data, awareness regarding protection and privacy is being raised by the population involved.

One of the biggest projects in this field, Sidewalk Toronto sid (2021) has ended in May 2020 due to economic uncertainty related to the COVID crisis. However, it has helped develop a new model of inclusive urban development. A public consulting made by Forum Research, one of Canada's leading survey research firms, infers that 60% do not trust Sidewalk Labs to collect data on its residents, with 39% not having any trust at all. However, one-third do trust Sidewalks Labs to collect data on its residents, with 11% having a lot of trust tor (2019).

Besides public opinion, in her book "Affective computing" Picard (2003), discourse about the importance of computers having the ability to recognize, understand or even have and express emotions. He also discusses and criticizes its consequences:

> "Emotions, perhaps more so than thoughts, are ultimately personal and private. Providing information about the most intimate motivational factors and reactions. Any attempts to detect, recognize, not to mention manipulate, a user's emotions thus constitutes the ultimate breach of ethics and will never be acceptable to computer users. Attempts to endow computers with these abilities will lead to widespread rejection of such computer systems and will help promote an attitude of distrust to computers in general."

Hence, it is important to disassociate the emotional expression from the user's ID. Once the emotion felt in determining public space can be relevant to decision-makers, whom it comes

from does not need to be exposed.

Privacy is more than a philosophical or moral question. For example, in the field of law, the Europe's General Data Protection Regulation (GDPR) Recital 1 - Fundamental Right to Data Protection says: Data protection is a fundamental right. Citizens of EU Member States owe this right, in part, to the Charter of Fundamental Rights of the EU and the Treaty on the Functioning of the EU. In Brazil, the "*Lei Geral de Proteção de Dados (LGPD)*" has a similar purpose, which endorses the increasing relevance, not only by the population itself but by the government to protect its citizen's data.

Approaches that use video or images captured from public environments are especially dangerous once they have a severe chance of exposing someone's identity without consent. In order to avoid it, some approaches have been developed over the years (Figure 1 Newton *et al.* (2003)). The most naive techniques either try to mask the face partially (Figure 1 b, c, d, e), pixelation (Figure 1 g), or add random noise (Figure 1 m, n). One may think it is pretty challenging to re-identify the person, but while it can be difficult for humans, it can be a simple task for some face recognition algorithms. Besides it, some useful information is lost within the identity (Eg. emotion).

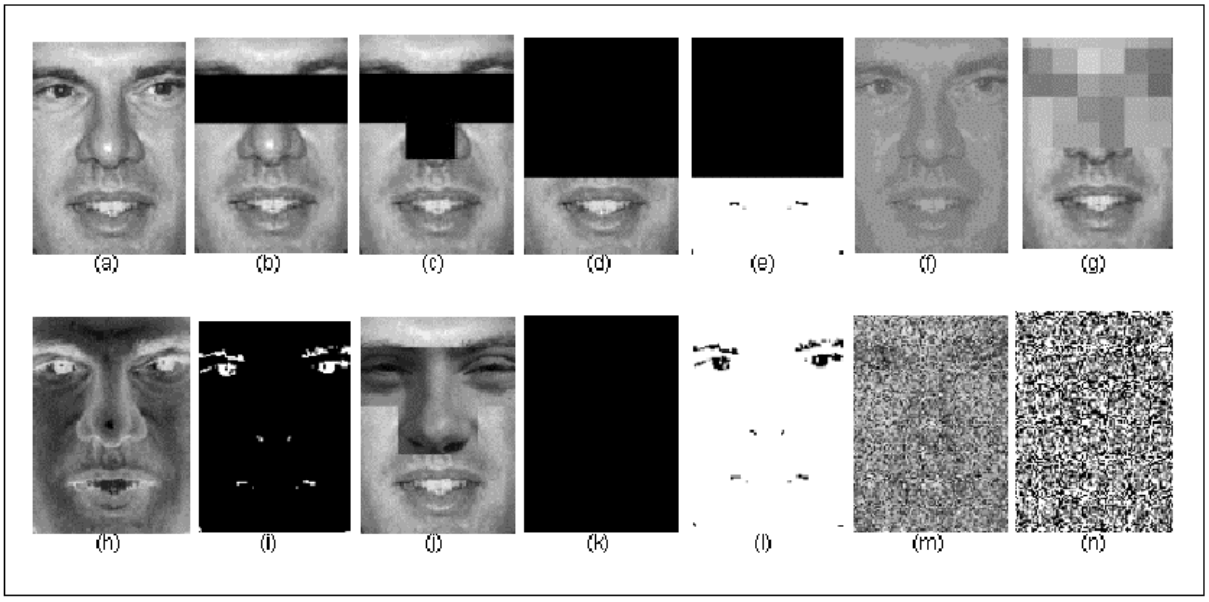

Figure 1: Newton, Elaine Sweeney, Latanya Malin, Bradley. (2003). Preserving Privacy by De-identifying Facial Images.

Unlike the previous methods, the use of Generative Adversarial Network (GAN) can generate a completely new face to substitute the original one. This framework has a lot o potential as the AI pioneer Yann LeCun, who oversees AI research at Facebook, has called GANs "the most interesting idea in the last ten years in machine learning."

Nevertheless, the problem of losing expression traces or other useful information can persist. Hence a method that generates the face based on the original landmarks entitled Conditional Identity Anonymization Generative Adversarial Network (CIAGAN) Maximov *et al.* (2020) has the potential to solve the problem: anonymize someone's face while preserving his/her original facial emotion expression. Therefore, this work proposes to analyze CIAGAN in the context of not only Identity anonymization but in preserving emotion.

In Brazil, the concessionaire "ViaQuatro" used cameras for advertising purposes on São Paulo's subway and was fined by justice by BRL 100,000 fin (2018). They were fined for not having asked for the consent of passengers to obtain biometric data. The company still needed to have presented clear information about data collection and processing. This endorces the necessity of having systems capable of anonymizing faces, preferlably with the hability to preserve many facial attributes, possibly gender, race, age or other useful information that could be used in advertisings or even social researches. Once the person ID is not revealed by adding an anonymization system, those advertisings or researches could collect information in public environments without inflicting any legal clause related to privacy.

1.2 OBJECTIVES

1.2.1 General Objective

The **general** objective of this work is to analyze how the anonymization model CIAGAN Maximov *et al.* (2020) would perform in terms of maintaining the original face expression while succeeding in preserving the face identity.

1.2.2 Specific Objective

Those are the **specific** objectives of this work:

- To reproduce qualitatively the results obtained on the original CIAGAN material;
- To evaluate the effectiveness of CIAGAN in preserving the original face expression based on Facial Expression Recognition (FER) methods;
- To propose a modification on CIAGAN focused specifically on improving the facial expression preservation;
- To compare the proposed modifications with the initial results;

1.3 DOCUMENT STRUCTURE

This work is divided as follows: chapter 2 contains the definitions of some theoretical background needed to comprehend this work, such as Anonymization, GANs, CGANs, FERs. chapter 3 mentions and describes some related works, what they proposed, and their solutions. chapter 4 explains the methodology used to do the proposed analysis. chapter 5 describes the experiment and the results obtained. Finally, chapter 6 concludes the work and suggests future works.

2

THEORETICAL BACKGROUND

This chapter describes some theoretical background relevant to the comprehension and development of this work. This chapter is divided into two subtopics: the first is relative to security and privacy concepts. The second gives a brief definition of some core concepts in Artificial Intelligence (AI) and other technologies used in this work.

2.1 PRIVACY

According to Raghunathan in his book "The Complete Book of Data Anonymization: From Planning to Implementation" Raghunathan (2013) :

> "data anonymization is the process of de-identifying sensitive data (prevent someone's identity from being revealed) while preserving its format and data type."

Depending on the chosen algorithm, the masked data can be a sequence of either realistic or a random data. Polonetsky *et al.* (2016) published a visual guide to practical data de-identification, as we show in Figure 22, to elucidate the different concepts related to privacy preservation. The methods: Pseudonymization, De-Identification, and Anonymization were displayed on a table regarding how they affect personal data.

In Table 1 the identifiers are divided into direct: the ones which can discover someone's identity without any extra data, and indirect: the data which, with some extra data, can be used to identify the person.

Table 1: Examples of Indirect and Direct Identifiers

Direct Identifiers	Indirect Identifiers
	Place of work
Name	Job Title
Address	Salary
Postal Code	Employment History
Telephone Number	Medical Diagnosis
Photograph or Image	Geolocation
	Device ID

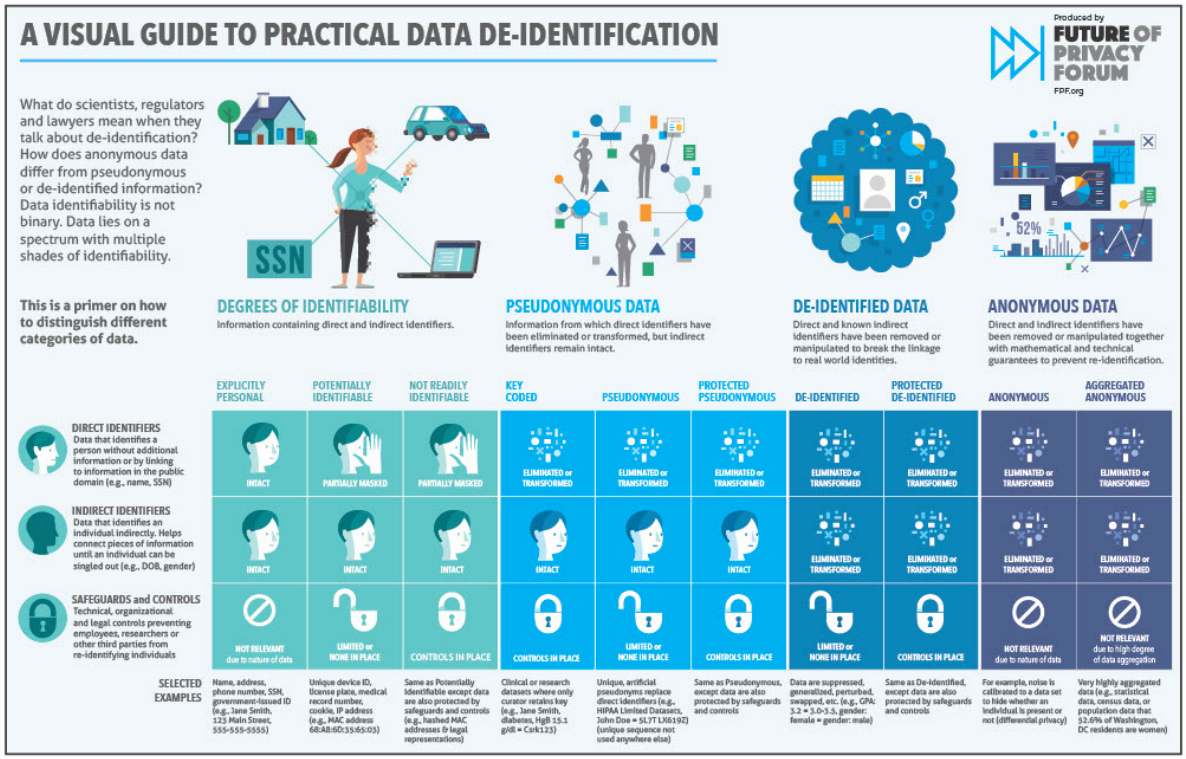

Figure 2: Polonetsky, Jules and Tene, Omer and Finch, Kelsey (2016) Visual De-Identification Guide

2.1.1 Pseudonymization

Direct identifiers have been eliminated or transformed, like fake names, but the indirect identifiers stay the same. Thus, a third party could still identify someone by accessing other datasets and overlapping other attributes. E.g., use the user's document ID instead of the real name.

2.1.2 De-Identification

Direct and known indirect identifiers have been removed or manipulated. It is safer than pseudonymization, but data can still be crossed with external content to determine someone's identity.

2.1.3 Anonymization

Besides removing or manipulating direct and indirect identifiers, mathematical and technical models guarantee their effectiveness in avoiding re-identification.

2.2 ARTIFICIAL INTELLIGENCE

McCarthy defines AI "It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable." In other words, AI is the field of science where computers learn how to perform activities otherwise limited to humans and maybe, overpass our limitations. As a subfield of AI, Machine Learning (ML) can learn to recognize patterns and predict future outcomes.

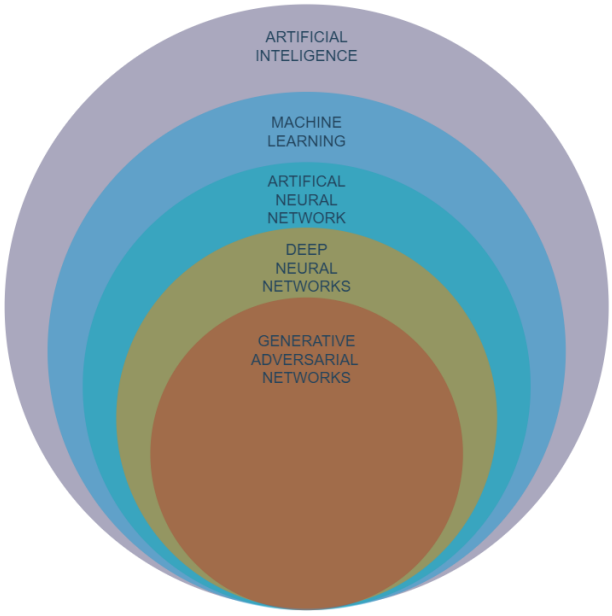

Figure 3: Artificial Inteligence, Machine Learning and Deep learning relationship

2.2.1 Artificial Neural Networks

Artificial Neural Networks are a class of ML methods inspired by the human brain. They are composed of layers, each layer containing artificial neurons or nodes. Each node has

a weight and receives input from the previous layers (since the input layer) and can pass the information towards if the activation function conditions are fulfilled until the output layer. With enough data and iterations, the loss function is minimized, and thus the weights of the nodes are optimized. The trained network is then ready to return the desired output, given an input of similar distribution of the training dataset. Deep Neural Networks (DNN), is a subfield of Neural networks, in which vast amounts of data are used in Neural networks with more than one hidden layer.

2.2.2 GAN

GAN were proposed by Goodfellow *et al.* as a framework for estimating generative models. It consists in two models trained simultaneously, but with opposite objectives: while G tries to generate samples as close as possible of the dataset distribution, D tries to identify whether the image came from the original dataset or from G. In Figure 4, G tries to generate handwriting numbers, and D tries to determine which images are indeed sent by the training set.

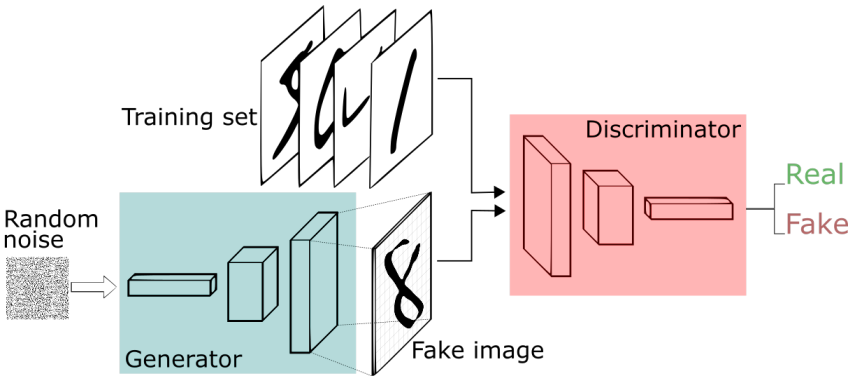

Figure 4: GAN architecture

More specifically, the main idea behind GANs is that the generator receives a random variable and tries to find a "transform function." This function can be understood as an inverse of the Cumulative Distribution Function (CDF). Once CDF is a function of a random variable with the domain interval [0,1], the "transform function" would take a distribution of a random variable and reshape than into new distribution (Figure 5), similar to the dataset distribution dlb (2022).

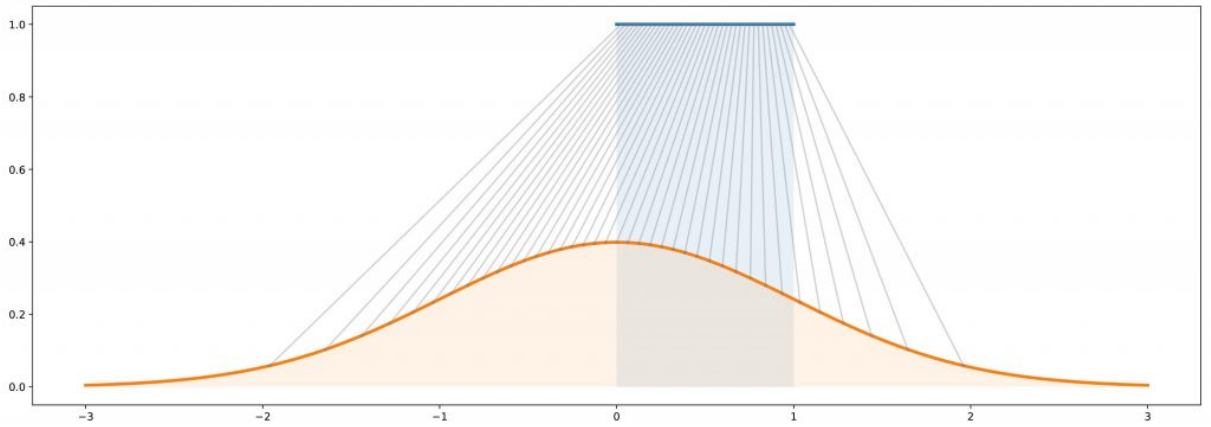

Figure 5: Transform Function idea used on Generator

Nevertheless, what if the true dataset distribution is not known? Possibly, the true distribution does not even exist as it should be overly complex. Then, only approximated distributions could be achieved. The discriminator models a discriminative function to classify whether the received image comes from the dataset or the generator (Figure 6).

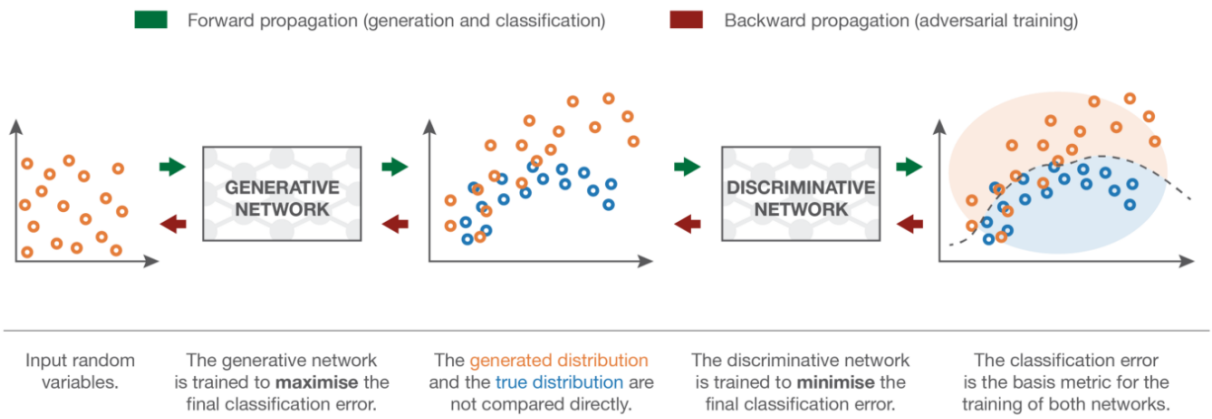

Figure 6: GAN training iteration

With $D(x)$ receiving a data x and returning the chance of it being from the original dataset and $G(z)$ receiving a random variable z and generating a "transform function" to data space. The loss function we want to minimize is described in Equation 2.2.2 and is the sum of the error of D when classifying if x belongs to the original distribution plus the error of D classifying the generated data from G as part of the generated dataset distribution (false). G tries to minimize the difference between the generated and real data, and D tries to maximize the distinguishability between true and generated data.

$$\min_G \max_D V(D,G) = \mathbf{E}_{x\sim pdata(x)}[\log D(x)] + \mathbf{E}_{z\sim pz(z)}[\log(1-D(G(z)))], \tag{2.1}$$

2.2.3 CGANs

Mirza & Osindero introduced Conditional Generative Adversarial Network (CGAN), a similar yet powerful addition to traditional GANs. Auxiliary information y now guides the generator and the discriminator. By considering y as a condition, the network can be directed to generate data under certain restrictions or conditions delimited by y. The Equation 2.2 are the same as Equation 2.2.2, with the only difference being the conditional of y.

$$\min_G \max_D V(D,G) = \mathbf{E}_{x\sim pdata(x)}[\log D(x|y)] + \mathbf{E}_{z\sim pz(z)}[\log(1 - D(G(z|y)))], \tag{2.2}$$

2.2.4 FERs

FER is a subarea of Affective computing, a field that focuses on detecting, recognizing, and simulating human affects (feeling, emotion or mood). According to the European Data Protection Supervisor (EDSP) and European Data Protection Supervisor *et al.* (2021), in its TechDispatch, which provides descriptions of new technology and possible impacts on privacy and the protection of personal data, FER can be defined as:

> "FER is the technology that analyses facial expressions from both static images and videos in order to reveal information on one's emotional state."

The analysis of emotion occurs in three steps: face detection, followed by facial expression detection, and finally, emotional state based on expression classification (Figure 7).

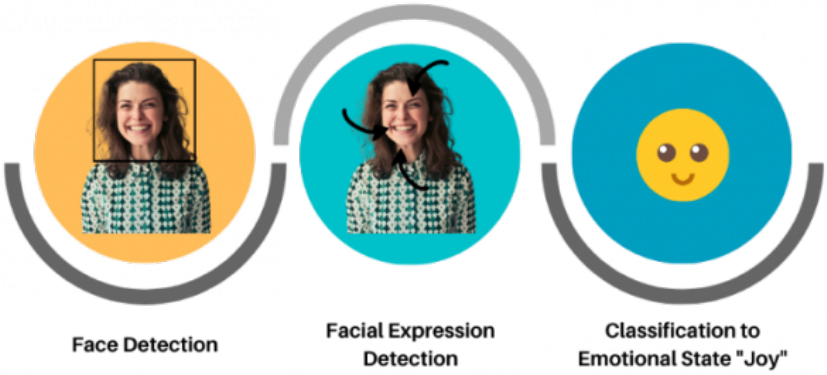

Figure 7: FER steps

3

RELATED WORKS

This chapter reunites some related works in de-identification that preserve emotional facial features.

In an early work for face de-identification with emotion preservation, Agarwal *et al.* propose an approach based on improving stylegan Karras *et al.* (2018). In their work entitled "Privacy preservation through facial de-identification with simultaneous emotion preservation" first, the desired face F1 has some emotional features extracted using mini-Xception convolutional neural network to an "emotion vector". Then, the emotion vector is compared and designate to the closest cluster. The closest cluster process is repeated now with face pose (orientation). The closest face in the cluster, denoted F2, is then used as input along F1's "emotion vector" to generate a new face that has features from F2 but preserves emotional aspects from F1.

In Figure 8 the first row shows the original faces, the second row shows proxy faces used as input on the GAN model, and, in the third row, the final result.

Figure 8: Agarwal Results

The following work tries to preserve only the essential features to determine the emotion. In "Preserving Privacy in Image-based Emotion Recognition through User Anonymization" Narula *et al.* (2020), the authors examine the interplay between emotion-specific and user identity-specific information. They propose an anonymization model based on a convolutional neural network (a network usually used to analyze images). The authors train their model in an adversarial manner that minimizes emotion classification loss and maximizes identification loss. This architecture could also be used to anonymize speed instead of images. However, the framework proposed was a bit expensive computationally and storage-wise once the model had many parameters to adjust. Nonetheless, it proves the concept that utility information, such as emotion, can be preserved, despite the original face being anonymized.

In Figure 9 from top to bottom, we have the original face; the face only with the attributes needed to recognize the emotion, the prior work on the area, and the proposed model, which freezes the initial layers of the emotion model, reducing the numbers of iterations needed to classify emotions.

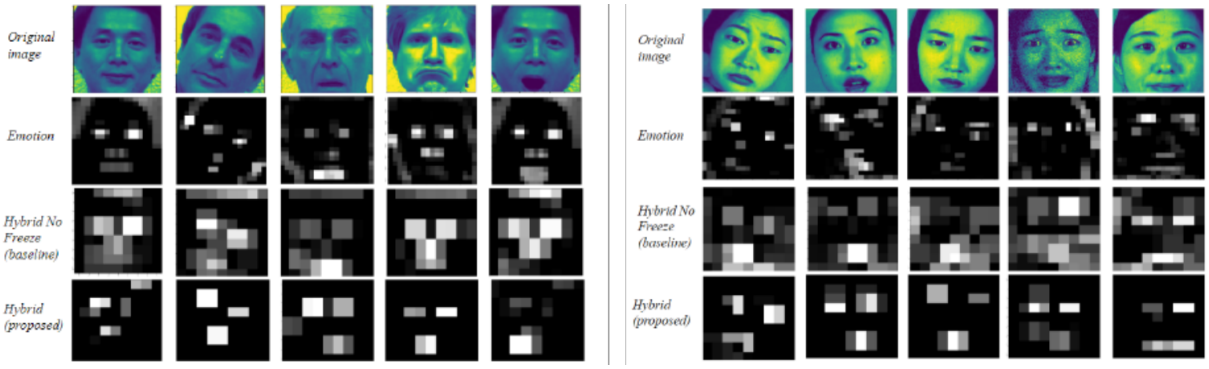

Figure 9: Narula Results

Although not explicitly preserving the facial expression, the authors could preserve many facial features and change precisely one desired attribute. In "Controllable Face Privacy" Sim & Zhang (2015), the authors use Multimodal Discriminant Analysis (MMDA) to create an orthogonal subspace in gender (G), race (R), and age (A). A given image Y is decomposed in Y = G,R,A,S parameters. Any of these attributes can be manipulated individually given a degree of intensity (Figure 10). The opposite is also available; a new face can be generated given the desired attributes. For example, if it is desired to maintain gender, race, and age and de-identify the person, the residual parameter (S) can be manipulated while the others are preserved. Although all images use people on neutral faces, they could not be used emotion-wise as it is today. This can be very helpful to computational analysis that needs these attributes to be computed unaltered.

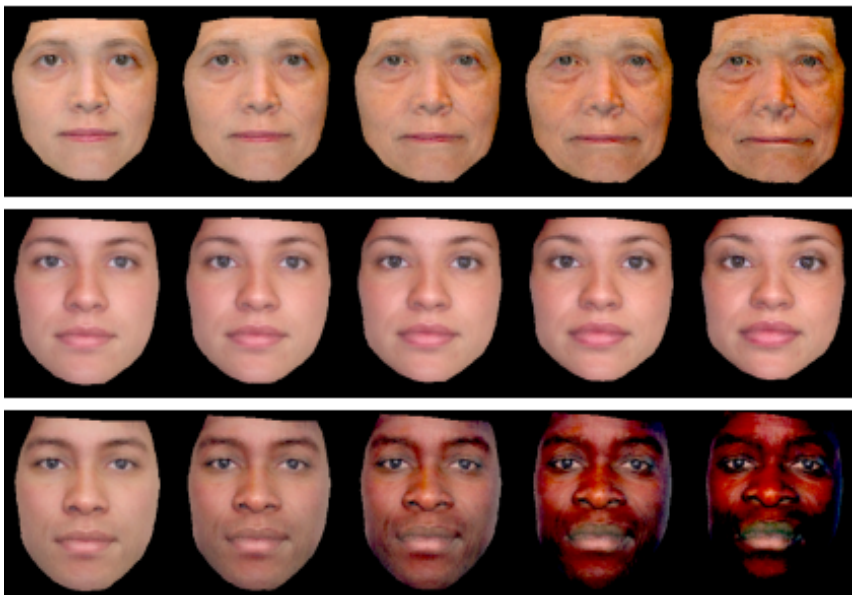

Figure 10: Sim and Zhang Results

4

METHODS

In this work, we propose an analysis of the usage of the CIAGAN Maximov *et al.* (2020) model in terms of expression preservation. A dataset focused on Facial Expression Recognition (FER) named RAF-DB Li *et al.* (2017) was used for the analysis. First, the model was evaluated by Distract your Attention Network (DAN) Wen *et al.* (2021) to classify the emotional expression present on each image and used as the baseline for comparison. Next, the CIAGAN model anonymizes the same dataset, and DAN re-evaluated the image results. Finally, the results were compared in a confusion matrix considering all seven labels' emotions available on the dataset (Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral).

4.1 ANONYMIZATION

CIAGAN was the chosen method to analyze emotion preservation because it preserves mouth expression through conditioning the GAN . As with any CGAN (see 2.2.3), CIAGAN receives as input (Figure 11) both the image to be transformed and a condition for the generated output image.

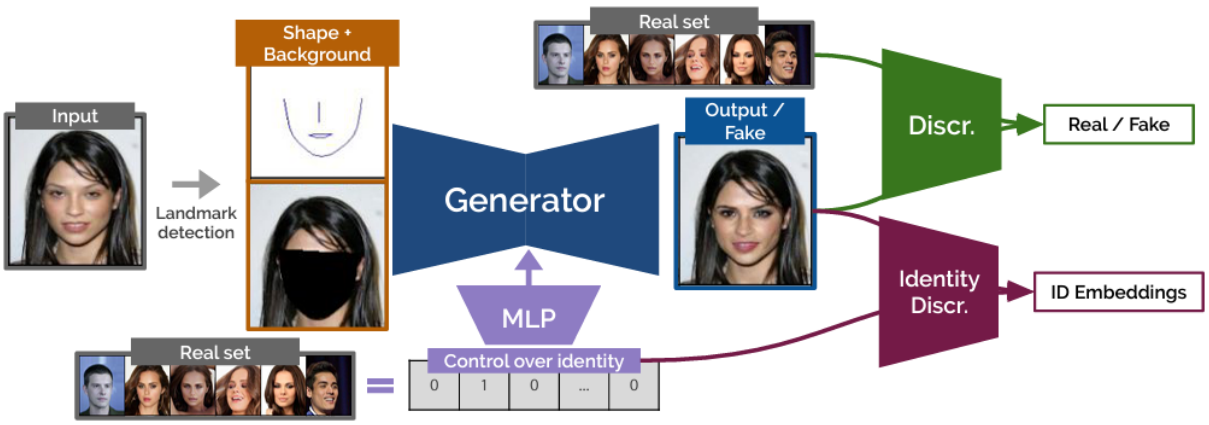

Figure 11: Ciagan Architecture

4.1.1 Inputs

Before the image is anonymized, the image has its landmarks removed with the Dlib method King (2009), and the face is removed and substituted with a black mask. It passes little identity information while conditioning the generator to preserve the pose for posterior tracking applications, besides giving temporal consistency for video anonymization.

4.1.2 Conditions

The landmarks extracted from Dlib are also passed as one of two crucial conditions alongside the image masked. Dlib can extract 68 landmarks from the face, but only a part of it is being used. The proposed CIAGAN method gets the face contour and the nose for the pose preservation and the mouth region, which can preserve some emotional information, like smiling or laughing. The eye's landmarks are removed to give more freedom to the model generator.

An ablation study has been done considering the whole face as input instead. The detection rate and image quality deteriorated slightly in the author's experiments. Besides, it would not preserve some useful face expressions used in FERs.

The second crucial condition is the desired Identity, chosen at random. It guides the generator to use some of its desired identity features. Without it, using only the landmarks as a condition, the model would quickly overfit and generate faces similar to the training set deteriorating the anonymity goal. The generated face identity should not be the same as any real identity.

4.1.3 Loss Function

Unlike the cross-entropy loss that mostly penalizes wrongly classified samples, Least-Squares Loss Function (LSLF) Mao *et al.* (2017) was the chosen Loss Function once also it penalizes generated samples that are classified correctly but are not close enough to the real ones. It helps to keep fake samples and get them closer to the decision boundary.

4.2 EVALUATION METRICS

DAN Wen *et al.* (2021) was the method chosen to evaluate the expression preservation before and after the anonymization process. During the development of this work, DAN is the current state of the art (best accuracy rate) in the RAF-DB dataset raf (2022). They achieved the best result by having multiple non-overlapping local attentions on each face analyzed. They demonstrate that multiple face clues should be considered to classify facial emotion efficiently, as can be seen in the four red areas analyzed in each face (each row) to classify the emotion in Figure 12.

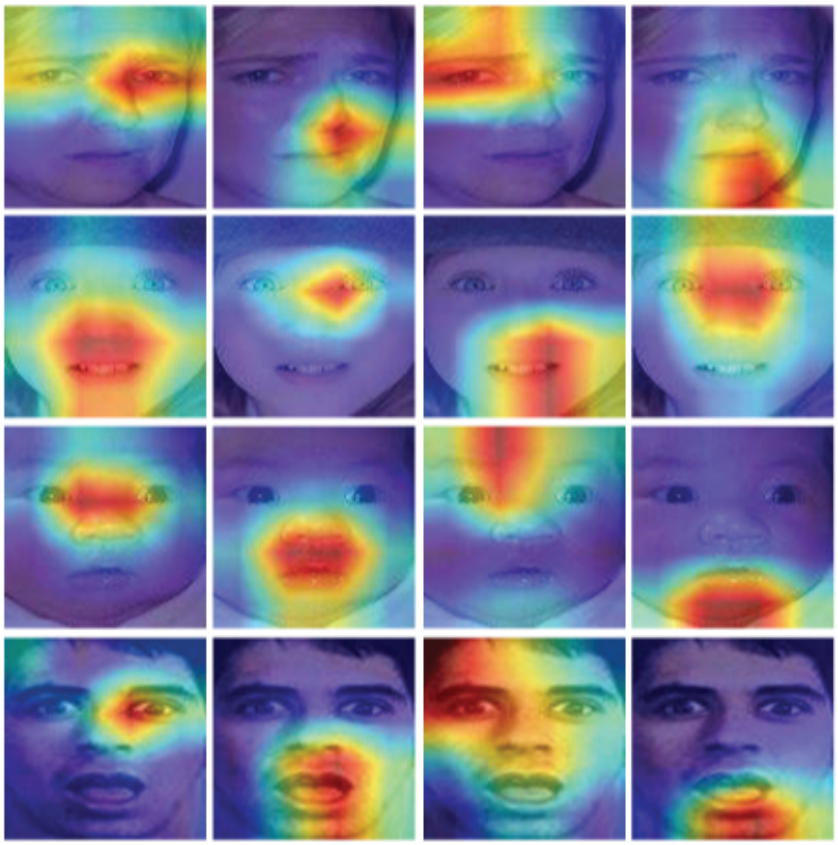

Figure 12: Multiple local attentions used on DAN

4.3 DATASETS

Two datasets were used in this work. The first one, Large-scale CelebFaces Attributes (CelebA) Liu *et al.* (2015), is a large dataset that includes more than ten thousand celebrities, with a total of more than two hundred thousand photos. The faces are aligned, and the dataset covers a considerable pose variation. The pre-trained model used in experiments 1 and 2 and the proposed model were trained with 1200 identities containing more than 30 photos each.

The second dataset used was Real-world Affective Faces Database (RAF-DB), a sizeable facial expression dataset with third thousand facial images labeled in 7 distinct emotions. It is one of the most used datasets to test FERs and was chosen to test the capacity of CIAGAN in preserving the original facial emotion in experiment 2.

4.4 LANDMARK'S FACIAL DETECTOR

4.4.1 Dlib

The Dlib King (2009) is a widespread Toolkit used in a variety of projects, mainly because of being an open-source project and because of its quality and facility to use. It contains components for linear algebra, image processing, data mining, data structures, machine learning, and others areas. For example, CIAGAN's implementation has used a built-in landmark detector to recognize 68 facial landmark points as in Figure 13. Then, a partition of the landmarks was used as input to guide the newly generated face.

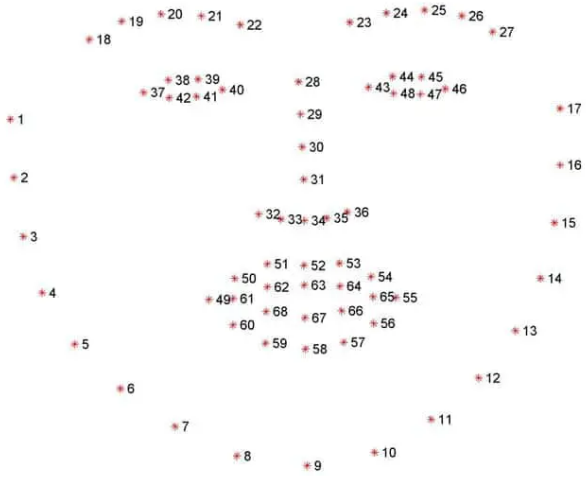

Figure 13: Dlib landmarks

4.4.2 Mediapipe Face Mehs

Google developed Mediapipe's open-source toolkit Grishchenko *et al.* (2020) to cover machine learning solutions. The kit has functions to detect objects, faces, iris, hands, and body pose. For example, we used the Face Mesh function to substitute Dlibs landmarks on the model training. The Face Mesh 14 is a 3D mesh mapped with 468 points on the face.

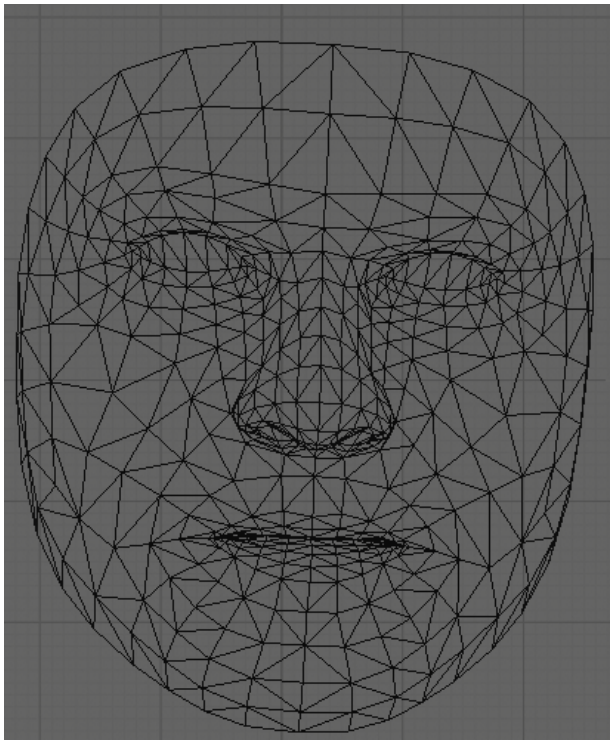

Figure 14: MediaPipe Face Mesh landmarks

5

EXPERIMENTS AND RESULTS

This chapter describes three experiments made in this work. In the first experiment, a qualitative analysis is done on the pre-trained model, considering two datasets. The second experiment evaluates the facial emotion preservation of CIAGAN on two datasets. Finally, the third experiment shows insights and initial attempts to train the model aiming to augment the facial emotion preservation pos anonymization.

The pre-trained model, used in experiments one and two, was downloaded directly from the CIAGAN's author [1]. The model used in experiment three was trained for one day and 8 hours on a desktop computer with the following specifications: ubuntu, NVIDIA RTX 3080 Ti with 12GB VRAM.

5.1 EXPERIMENT 1: QUALITATIVE ANONYMIZATION ANALYSIS OF PRE-TRAINED MODEL ON DISTINCT DATASETS

We elaborated experiment 1 to be a visual analysis of CIAGAN's pre-trained model. It consists of two steps: the first step is reproducing the anonymization with the pre-trained model on CelebA, the dataset used in the model's training. The second step is to test the anonymization model on the RAF-DB dataset, with seven distinct facial expression labels.

[1] https://github.com/dvl-tum/ciagan

Figure 15: Celeba original images (left) and anonymized (right)

The first step results are displayed in Figure 15. On the left, we have the original faces of the CelebA dataset. The faces resulting from the anonymization process of CIAGAN's pre-trained model are displayed on the right. Overall, the model can anonymize faces with significant quality, blending the new faces into the scene.

Figure 16: RAF-DB original images(left) and anonymized (right)

The results of the second step are shown in Figure 16. In red are the original faces from the RAF-DB dataset. In green, the faces that the CIAGAN's pre-trained model anonymized. The RAF-DB dataset collected its images from the internet. Some of them are distorted , some are cut, in black and white , or have an inferior image quality.

5.2 EXPERIMENT 2: FACE EMOTION PRESERVATION ANALYSIS ON PRE-TRAINED MODEL

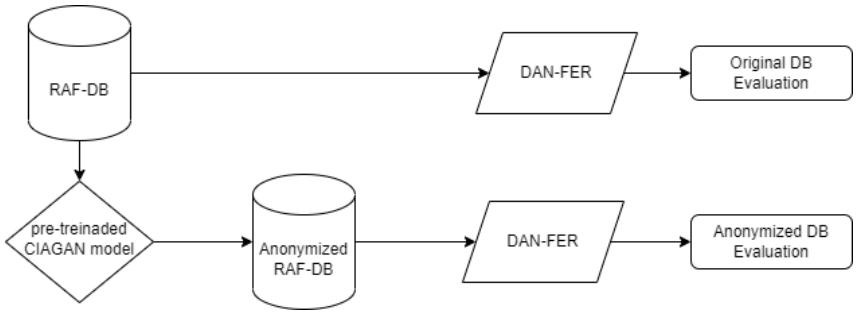

Figure 17: CIAGAN's Model Evaluation On Expression Preservation

This experiment is visually explained in (Figure 17) and occurs as follows: we deploy the DAN model to classify the RAF-DB dataset and use these annotations as a baseline of emotion classification. Next, we deploy the CIAGAN model with the pre-trained weights (with the dlib landmark detector) and anonymize the images on the RAF-DB dataset. Finally, we deploy DAN once again to evaluate the anonymized dataset and compare the classification of both datasets.

5.2.1 Results

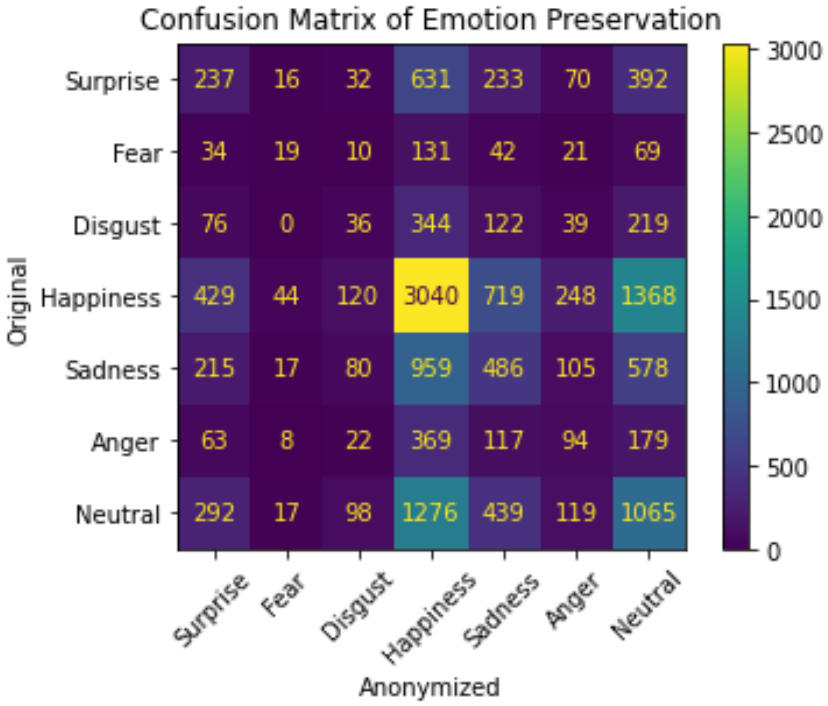

Figure 18: Confusion Matrix of Experiment 2

Comparing image by image DAN's classification pre and post anonymized, the overall facial expression preservation score was 32.45%.Figure 18 shows the confusion matrix resulting from the classification of the original dataset (x-axis) and the anonymized (y-axis). The RAF-DB dataset includes seven different expression labels: Surprise, Fear, Disgust, Happiness, Sadness, Anger, and Neutral. The main thing to point out is that Happiness was the class with an accuracy of 50.93%. We raised to hypothesis to explain it:

- Hypothesis 1: In CIAGAN's architecture (see Figure 11), the network receives as input the background of the face and the shape, captured by dlib's landmarks. The shape extracted from the face only gets the mouth, nose, and face contour to condition the generated face. Once no hint about the eyes or eyebrows is given, the generator has no clue how to maintain the whole facial expression. Therefore, the FER (see Figure 12), which analyses multiple face areas to determine the expression, gets the mouth region right, but the eyes region is not guaranteed to preserve the original facial expression.
- Hypothesis 2: CIAGAN is trained based on CelebA. This dataset contains more than two hundred thousand photos of celebrities, with the majority smiling or being neutral while having their photos taken. So, the network possibly favored this emotion when generating a new anonymized face.

5.3 EXPERIMENT 3: QUALITATIVE ANONYMIZATION ANALYSIS ON MODEL VARIATION

Experiment 3 was formulated from hypothesis 1, that the lack of significant landmarks was the main reason for the low results in facial expression preservation. Therefore, we trained the model substituting the Dlib's face contour, nose, and mouth for the mediapipe's landmarks Grishchenko *et al.* (2020), face contour, nose, mouth, eyes, and eyebrows.

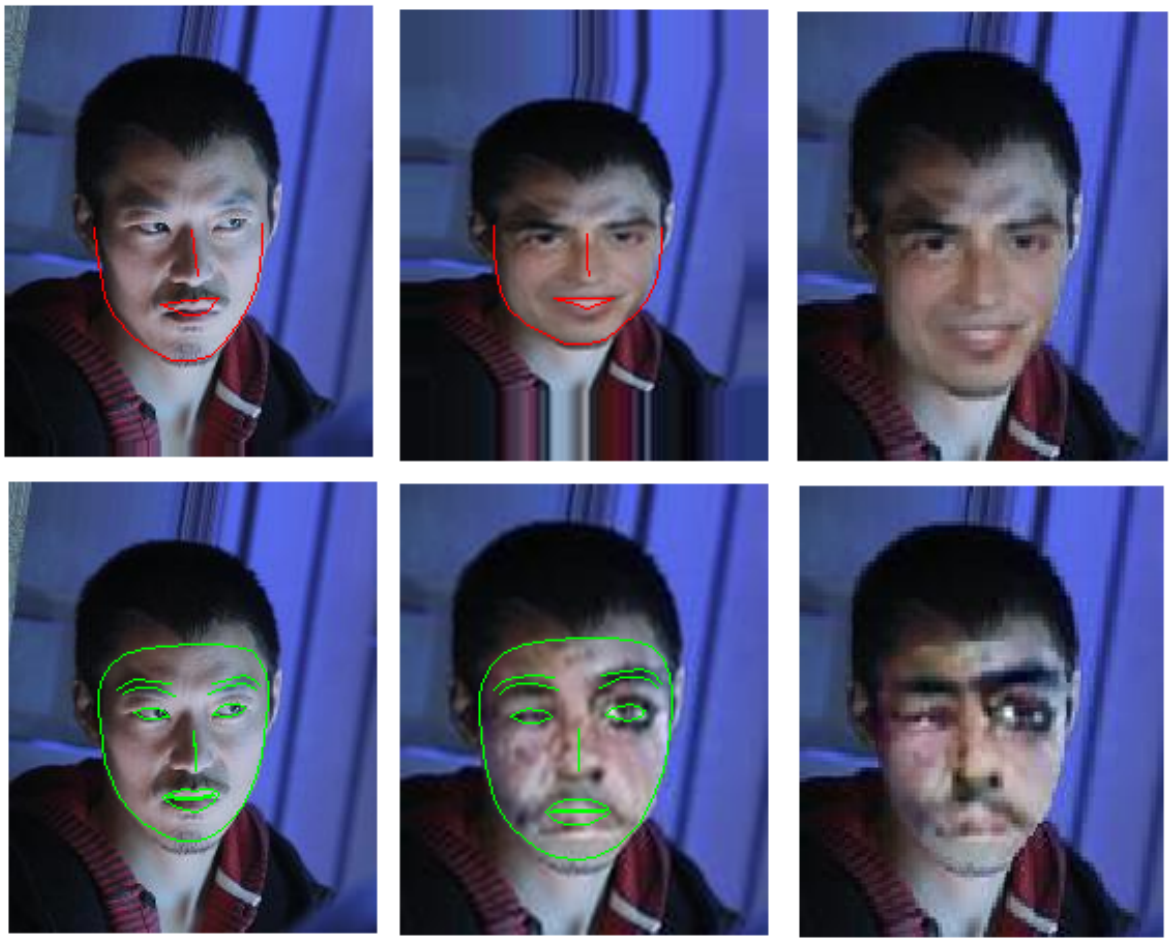

Figure 19: Comparison beetween, Dlib (red) and mediapipe (green) landmarks

In Figure 19, from left to right are displayed the original face with the landmarks detected, the anonymized face with the landmarks, and, finally, just the anonymized face. The red lines correspond to the pre-trained proposed model by CIAGAN that uses Dlib King (2009) to detect the face landmarks, while the green ones are our proposed model using mediapipe's landmarks Grishchenko *et al.* (2020).

Even though the final result (rightmost superior) of the original CIAGAN model has a better definition, it changes the original facial expression. This mistake is caused because Dlib could not correctly capture the expression on the original face; as a result, the anonymized face is smiling while the real one is not, leading to a miss emotional classification.

Although not yet fine-tuned visually, our approach used an advanced landmark detector. Therefore the landmarks were closer to the original face, especially on the mouth and eyebrows.

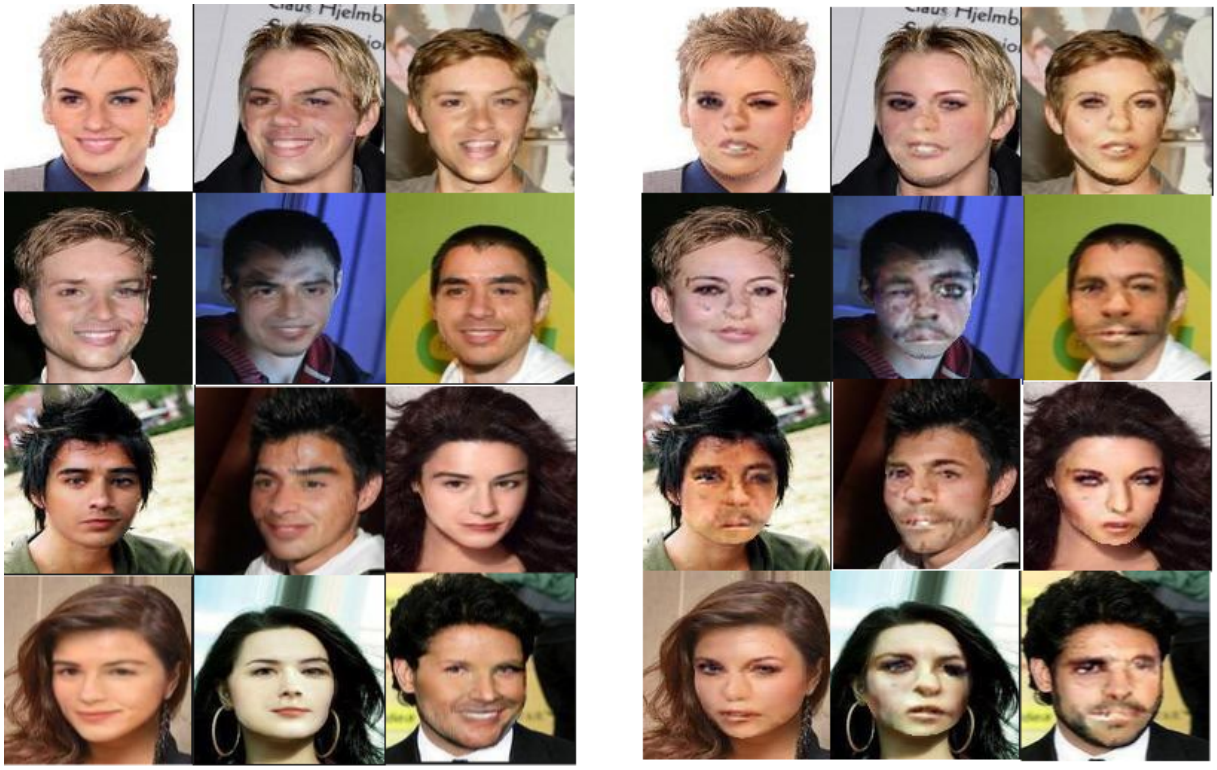

Figure 20: Celeba Anonymization Original (left), proposed (right)

5.3.1 Limitations

Figure 21: RAF anonymized by our proposed model

Regardless our approach had decent results in Anonymizing the CelebA dataset (Figure 20), it was not robust enough to be used in a distinct dataset from the trained one. Unfortunately, because of the unsatisfactory result in RAF-DB (Figure 21), experiment 2 could not be reproduced with our approach. Therefore, a more profound study could be done in the future to find the proper parameters to train the network to perform better in this or other FERs datasets.

6

CONCLUSION AND FUTURE WORKS

In this work, a study of use was elaborated. The original model proposed in CIAGAN's research paper was explained and evaluated according to the ability to generate new anonymized faces preserving or not the original facial expression. Unfortunately, the study has shown that although the author's initial concern about preserving smiles or laughs, this ability was not enough to preserve other kinds of emotions satisfactorily.

Two hypothesis were elaborated to improve these results: firstly, The conditional landmarks used with Dlib was not detailed enough. Secondly, related to the dataset CelebA used, which could biases the happiness facial expression on the model. A new model was trained, considering mediapipe's landmarks as input. This model was compared qualitatively with the original approach. Although not yet fine-tuned, our approach could preserve landmarks closer to the original CIAGAN model in some faces. This result suggests that facial expressions could be better preserved with more detailed landmarks models as input.

For future works, we would like to further investigate the network to, not only focus on improving facial expression preservation, but also guarantees the initial image quality. We would like to test other FERs datasets and others FERs as well to further analyse the CIAGAN. We also would like to test the first hypothesis and use a different dataset on training.

In addition, we encourage future studies on the trade off between preserving useful facial information precisely and the actual face de-identification accuracy. Besides only preserving emotional expressions, preserving other important facial information presented on the original face should be a concern as well as not deprecating the actual de-indentification capacity.

REFERENCES

(2018). Justiça multa ViaQuatro em R$ 100 mil por biometria facial no metrô de SP viaquatro's fines. https://www.terra.com.br/noticias/tecnologia/justica-multa-viaquatro-em-r-100-mil-por-biometria-facial-no-metro-de-s ffd3745fc49456834e8f4252a7034ed5p0xkbfoj.html. Accessed: 22-05-10.

(2019). Support for Sidewalk Toronto Mixedforumresearch. http://poll.forumresearch.com/post/3002/sidewalk-toronto-july-2019. Accessed: 22-04-11.

(2021). Sidewalk Torontosidewalk. https://www.sidewalklabs.com/toronto. Accessed: 22-04-11.

(2022). Data Science Academy. Deep Learning Bookdl book. https://www.deeplearningbook.com.br/. Accessed: 22-05-10.

(2022). Facial Expression Recognition on RAF-DBraf-db benchmark. https://paperswithcode.com/sota/facial-expression-recognition-on-raf-db. Accessed: 22-04-15.

Agarwal, A., Chattopadhyay, P., & Wang, L. (2021). Privacy preservation through facial de-identification with simultaneous emotion preservation. *Signal, Image and Video Processing*, 15(5):951–958.

and European Data Protection Supervisor, Vemou, K., & Horvath, A. (2021). *EDPS TechDispatch : facial emotion recognition. Issue 1, 2021*. Publications Office.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., & Grundmann, M. (2020). Attention mesh: High-fidelity face mesh prediction in real-time. *CoRR*, abs/2006.10962.

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks.

Khanan, A., Abdullah, S., Mohamed, A. H. H., Mehmood, A., & Ariffin, K. A. Z. (2019). Big data security and privacy concerns: a review. *smart technologies and innovation for a sustainable future*, 55–61.

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

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Maximov, M., Elezi, I., & Leal-Taixe, L. (2020). Ciagan: Conditional identity anonymization generative adversarial networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

McCarthy, J. (2007). What is artificial intelligence?

Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Narula, V., Feng, K., & Chaspari, T. (2020). *Preserving Privacy in Image-Based Emotion Recognition through User Anonymization*, 452–460. Association for Computing Machinery, New York, NY, USA.

Newton, E., Sweeney, L., & Malin, B. (2003). Preserving privacy by de-identifying facial images.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1):55–64. Applications of Affective Computing in Human-Computer Interaction.

Polonetsky, J., Tene, O., & Finch, K. (2016). Shades of gray: Seeing the full spectrum of practical data de-intentification. *Santa Clara L. Rev.*, 56:593.

Raghunathan, B. (2013). *The Complete Book of Data Anonymization: From Planning to Implementation*. Auerbach Publications, USA.

Sim, T. & Zhang, L. (2015). Controllable face privacy. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 04:1–8.

Wen, Z., Lin, W., Wang, T., & Xu, G. (2021). Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*.