EMOTION DETECTION WITH PRIVACY PRESERVATION USING ADVERSARIAL LEARNING

A Thesis

by

RAVIKIRAN RAMESH

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Dr. Theodora Chaspari
Committee Members,	Dr. Shuiwang Ji
	Dr. Tie Liu
Head of Department,	Dr. Scott Schaefer

December 2021

Major Subject: Computer Science

ABSTRACT

The continuous monitoring of one's emotional state can provide valuable insights about their psychological well-being and can be used as a foundation for diagnosis and treatment applications. Yet, due to privacy concerns, technologies that continuously monitor signals that reflect emotions, such as images, are met with strong skepticism. This thesis aims to design a privacy-preserving image generation algorithm that anonymizes the input image and at the same time maintains emotion-related information. To do so, we identify landmarks in human faces and quantify the amount of emotion and identity based information carried by each of the landmarks. We then propose a modification of a conditional generative adversarial network that can transform facial images in such a way that the identity based information is ignored while the emotion based information is retained. We then evaluate the degree of emotion and identity content in the transformed images by performing emotion and identity classification using these images. The proposed system is trained and evaluated on two publicly available datasets, namely the Yale Face Database and the Japanese Female Facial Expression dataset, and the generated images achieve moderate to high emotion classification accuracy and low identity classification accuracy.

DEDICATION

To my mother Radha, my father Ramesh and my sister Roshini.

ACKNOWLEDGMENTS

I would like to take this opportunity to sincerely thank to my advisor Dr. Theodora Chaspari for guiding me through each and every step of my research. Throughout the duration of my research, she was extremely involved in my work. She was really patient with me and allowed me to explore this domain at my own pace. I am extremely thankful for her guidance and support. I would also like to thank Dr. Shuiwang Ji and Dr. Tie Liu for serving on my committee and supporting my work.

Furthermore, I would like to thank Vansh Narula for doing a fine job with with his thesis, which served as a solid foundation for my work. Thank you for being an amazing senior, and helping me carry over your work in a smooth manner. I would further like to thank each and every member of the Human Bio-behavioral signals (HUBBS) lab for their support, feedback and suggestions. I would like to thank all the friends I made at Texas A&M for making the last 2 years one of the most memorable times of my life. Finally, I would like to thank my family for supporting me in more ways that I can express in words.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The thesis committee for this work include Dr. Theodora Chaspari (Chair) and Dr. Shuiwang Ji (Member) of the Department of Computer Science at Texas AM University and Dr. Tie Liu (Member) of the Department of Electrical and Computer Engineering. The ideation and implementation of this research was conducted with the help and guidance of Dr. Theodora Chaspari. The prior work of Vansh Narula, Dr. Zhangyang Wang and Dr. Theodora Chaspari served as the inspiration behind this work. All other work conducted as part of this thesis was completed by the author independently.

Funding Sources

This work has been supported by the Texas A&M University High Performance Research Computing (HPRC) facility and the National Science Foundation (IIS-2046118).

NOMENCLATURE

CNN	Convolutional Neural Networks
FDF	Flickr Diverse Faces dataset
JAFFE	Japanese Female Face Expressions database
YALE	Yale Faces database
AI	Artificial Intelligence
IoT	Internet of Things
MFCC	Mel Frequency Cepstral Coefficients
ASR	Automatic speech recognition
HIPAA	Health Insurance Portability and Accountability Act of 1996
HoG	Histogram of Oriented Gradients
RBF	Radial Basis Function
SIFT	Scale Invariant Feature Transformation
SVM	Support Vector Machine
TF - IDF	Term frequency - Inverse Document Frequency
CLI	Command Line Interface
PSNR	Peak signal to noise ratio
LPIPS	Learned perceptual image patch similarity
ResNet	Residual Network
API	Application Program Interface

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Importance of emotions in human-computer interaction

Emotional awareness is a vital part in interpersonal interactions. The ability to understand the emotional state of other people can largely determine the effectiveness of our communication. This is also applicable to human-computer interaction, since systems that can adapt their responses and behavioral patterns according to the emotions of humans can make the interaction more natural and organic [5]. The understanding of emotions can also enable us to design more effective intervention interfaces that track emotional states and provide appropriate suggestions in a personalized manner [6]. Therefore, an increasing number of intelligent systems is currently using emotion recognition models to improve human-computer interaction. Today, emotion recognition is a widely-researched topic, and it has been successfully implemented for speech [7, 8, 9, 10, 11, 12], face [13, 14, 15, 16, 17], text [18, 19, 20, 21], and even handwriting data [22, 23, 24, 25]. Detecting and understanding emotions can also contribute to critical life-saving tasks, such as the detection and early intervention upon people's mental health [26, 27] and even suicide prevention [28].

1.2 Privacy risks of emotion detection

Despite the benefits of these applications, the wide-spread adoption of emotion recognition algorithms is hindered by various privacy and confidentiality concerns. Emotion detection involves the collection of personal ly identifiable information, such as face, voice, and/or written text. Users are often skeptical of such technologies, since they are afraid that information relevant to their identity will be permanently stored in third-party servers or will be abused by hacker attacks [29]. This concern is further supported by studies that indicate that emotional information and user identity information often interact with each other, while it is not possible to completely decouple them [30]. Current methods used to extract emotion-based information from various human generated signals also preserve identity information. For instance, the Mel-Frequency Spectral Coefficients (MFCC) extracted from speech are commonly used for Automatic speech recognition (ASR) [31]

[32] [33], emotion recognition [34], and user identity recognition [34] [35] with a high degree of accuracy. The histogram of oriented gradients (HoG) is utilized for a wide variety of object detection tasks, including, but not limited to person detection [36], handwritten digit recognition [37], and scene text recognition [38]. These features also capture the rich textual information of the facial image which can be employed to perform user identification [39]. Eigen-faces are a set of eigen-vectors used in facial recognition [40]. They rely on the most significant eigen-vectors that preserve the most information about the face, which also includes uniquely identifiable personal information. Other approaches that leverage the frequency characteristics of the image using Gabor filters and wavelets [41] are sensitive to illumination and texture, which can be used to extract the identity of the face.

Current computer vision algorithms revolve around convolutional neural networks (CNN) [42], which provide state-of-the-art results on any data with strong spatial characteristics [43]. The reason for the widespread adoption of CNNs is due to its ability to capture general and highly reusable information in its convolutional kernels [44]. This makes it an excellent candidate for transfer learning, an approach where we utilize the knowledge gained while solving one problem and applying it to a different but related problem [45] [46]. Often, CNNs are trained on massive datasets that may contain sensitive user information [43], and therefore, these pre-trained networks are used to learn another task with increased performance and decreased training time [47] [45]. However, this results in both the desired utility-based information (e.g., emotional, mental, and psychological state) as well as the undesired privacy-sensitive information (e.g., user identity) being preserved in the convolutional base and the subsequent fully-connected layers [2]. This is further exacerbated by the fact that it is notoriously difficult to make these machine learning models "unlearn" or "forget" any information that has been learnt in the past [48].

This renders data privacy a major barrier for collecting and sharing human behavioral signals, stalling the research and development of emotion tracking systems that can potentially contribute to improving well-being and mental health outcomes. This privacy compromising landscape renders essential the design of novel machine learning systems that conceal one's identity, while at the

same time preserve useful information for emotion recognition [49].

1.3 Prior work in privacy preservation

While it is not possible to completely decouple the user identity information and emotional information from human signals [30], we can suppress the identity information to such an extent that it becomes infeasible to uniquely identify people based on their sensitive data. There are multiple prior works that focus on preserving privacy in two of the most sensitive human signals: human speech and face.

1.3.1 Speech anonymization

Speech is a very sensitive human signal, since it is possible to recognize the identity of the speaker with a very high degree of accuracy [50] [51]. Therefore it is important to promote the development of privacy preservation techniques for speech technology.

1.3.1.1 VoicePrivacy challenge

One of the best initiatives created in order to encourage research in the field of voice anonymization is the VoicePrivacy challenge [52]. The purpose of the challenge is to anonymize the user's speech and perform Automatic speech recognition (ASR) such that the user's identity is not compromised. The challenge consists of 2 baseline models with which the proposed models will be evaluated. The first baseline model follows an ensemble approach, involving the following features: X-vectors [53], fundamental frequency (F0) and bottleneck features (BN) [54]. The X-vector is anonymized by finding the N farthest x-vectors in the dataset. A neural source-filter [55] is used to resynthesize the speech using the anonymized x-vector and the F0+BN features. The second baseline is a much simpler model that performs speaker anonymization using McAdams co-efficient [56].

The challenge also consists of an attack model, where the attacker has access to the anonymized data and the list of original speakers. The anonymization system is evaluated on a series of objective and subjective metrics [57]. The objective metrics include log likelihood of speaker verifiability and ASR decoding error (ratio of original ASR accuracy and anonymized ASR accuracy). The

subjective metrics are calculated using listening tests carried out using volunteers. The volunteers will rate the subject speaker verifiability, speaker linkability, speech intelligibility, and speech naturalness. The challenge has yielded some interesting results, which anonymize the speakers while maintaining a high degree of ASR accuracy and speech naturalness [58]

1.3.1.2 Other Notable approaches

There are many other interesting works done in the field of speech anonymization. VoiceMask [59] is an application that anonymizes the speaker's identity by randomly modifying the speaker's voice. It also replaces the sensitive keywords in the speech in order to hide sensitive information about the speaker. This is really important, since even if the speaker's identity is masked, an attacker will be able to obtain sensitive information about the speaker by analysing the contents of the speech. Speech Sanitizer [60] is a very similar work, but is also performs generic keyword spotting using user feedback, and a personalized keyword spotting using TF - IDF scoring [61]. Voice anonymization in urban sound recordings [62] is another interesting work which identifies human voices in public recordings, removes the voices, blurs it and recombines it with the original recordings. This approach is aimed at performing speaker obfuscation (by eliminating both speaker and ASR information), while preserving the quality of the background noise.

1.3.2 Face anonymization

Facial recognition technology is currently being used for a wide variety of real-world applications, including smartphone unlocking, forensic investigations, surveillance, tagging people on social media, and so on. Due to such wide-spread usage of facial recognition technologies, it is extremely important to consider user privacy, now more than ever.

There are multiple promising prior works that perform user anonymization on the face, while preserving other useful information in the image. Ren *et al.* [63] proposed an approach that performs action detection while maintaining user anonymity. The proposed model uses an adversarial approach, which is aimed at maximizing the face detection loss and minimizing action detection loss as well as the L1 loss between the original and anonymized image (to preserve the basic facial

structure). Kim *et al.* [64] proposes a face anonymization technique that can be used in robot vision to protect individual privacy. The robot views the world in extremely low resolution and dynamically scales up the resolution of only the background in order to navigate properly. When scaling up the resolution, individual human faces are detected and those pixels are not scaled up, therefore the privacy-sensitive information will remain in extremely low resolution.

For this thesis, we will be outlining three particular privacy preserving approaches in more detail. These approaches are of significant interest and they contributed largely to the ideas proposed.

1.3.2.1 DeepPrivacy

DeepPrivacy [1] is a privacy preserving model that employs a conditional Generative Adversarial Network (GAN) to anonymize and generate realistic faces. DeepPrivacy uses the image background and the original image's pose as conditions to generate realistic faces with seamless transition between the face and background. The generated face will have the same facial structure as the original face, but it has different facial features which can effectively anonymize the original image. Figure 1.1 represents the data pipeline utilized by DeepPrivacy.

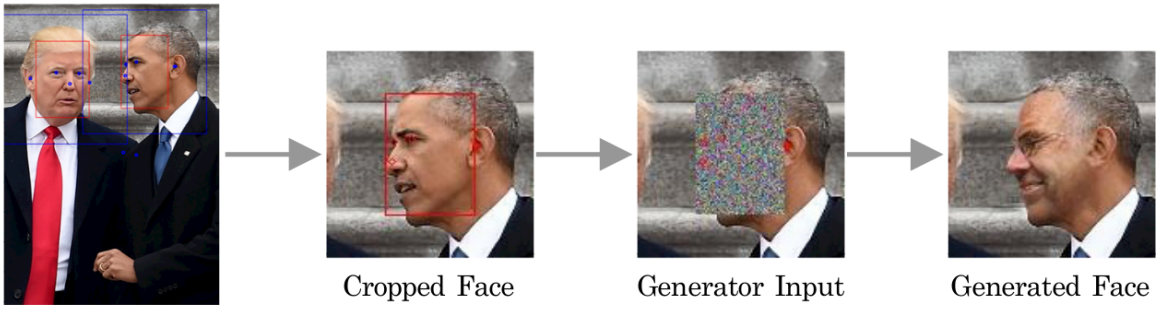


Figure 1.1: Input pipeline of DeepPrivacy [1]. Each detected face is cropped to a quadratic image, then the face pixels ae replaced with a constant value, and feed it to the generator. The keypoints represent the image pose and are represented as a one-hot encoded image.

The model is based on the one proposed by Karras *et al.* [65]. Their model is a non-conditional GAN, and several alterations were performed to include conditional information. Seven key land-

marks were used to describe the pose of the face: left/right eye, left/right ear, left/right shoulder, and nose. To reduce the number of parameters in the network, the pose information is preprocessed into a one-hot encoded image of size K x M x M , where K is the number of landmarks and M is the target resolution. Progressive growing is applied to both the generator and discriminator to grow the networks from a starting resolution of 8. The resolution is doubled each time the network is expanded until the final resolution of 128 x 128 is reached. The pose information is included for each resolution in the generator and discriminator, making the pose information finer for each increase in resolution. Using this progressive growth technique results in a higher quality of output images. Figure 1.2 represents one such image along with its anonymized version.

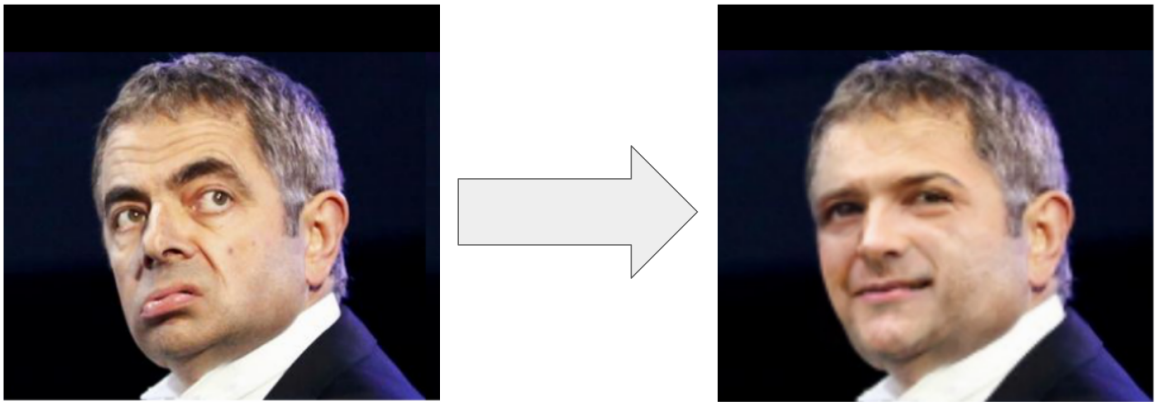
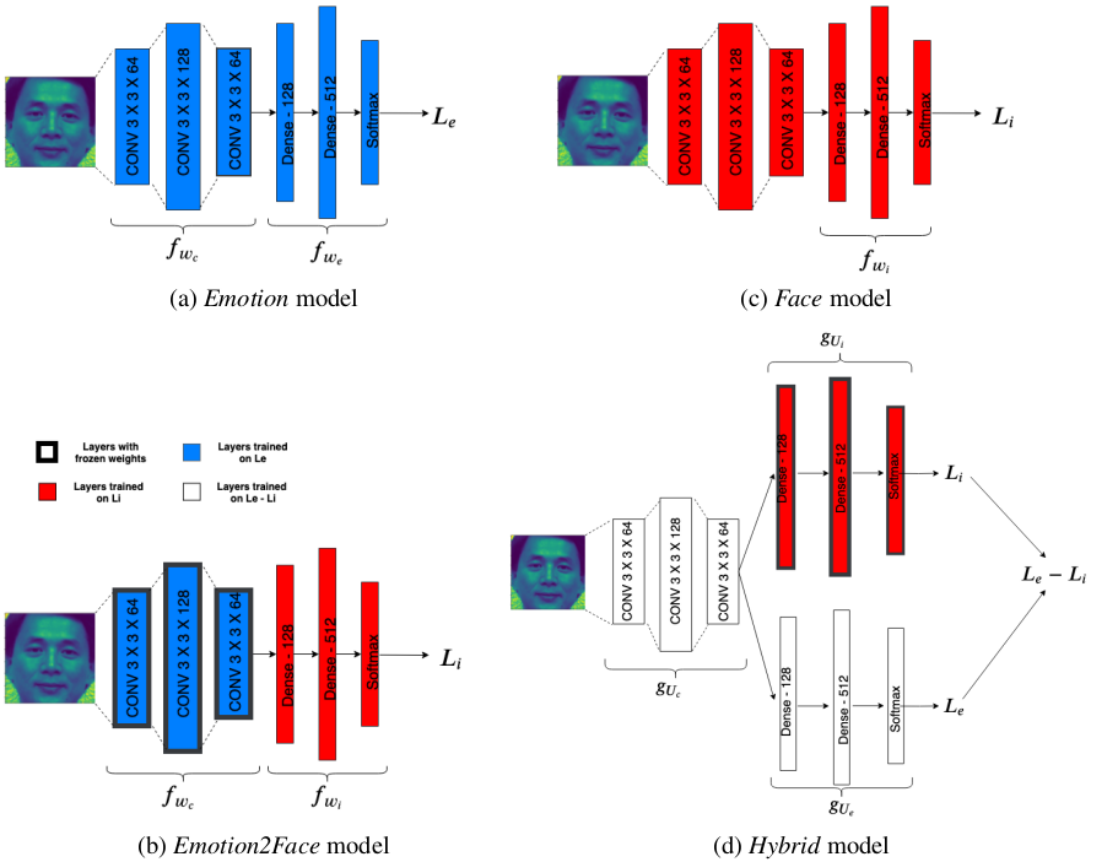


Figure 1.2: Example of an image along with its anonymized version transformed using DeepPrivacy [1].

1.3.2.2 An Adversarial learning framework for preserving users' anonymity in face based emotion recognition

Another work that has specifically focused on privacy-preserving emotion recognition [2] proposes the use of an adversarial framework that learns a transformation to maximize emotion specific information and minimize user information. First, two separate CNNs are trained that will each perform emotion and user identity classification. The CNNs are comprised on convolutional

layers followed by fully connected layers. Then, the approach uses a hybrid combination of the previously trained CNNs to iteratively unlearn the user ID information while preserving the maximum amount of emotional information. This approach is based on the hypothesis that the convolutional layers carry the most user-dependent information, and while it is not possible to completely unlearn that information, it is possible to reduce it to an acceptable degree. This approach is effective since it deals with images where the faces are in close proximity to the face, when compared to earlier approaches that utilizes distant surviellence cameras that captures the entire body [66, 67, 68]. This paper also describes clear methods for evaluating the trade-off between the degradation of utility-based information and preservation of user identity.



(a) *Emotion* model (c) *Face* model

(b) *Emotion2Face* model (d) *Hybrid* model

Figure 1.3: Architecture proposed by Narula *et al.* [2]

1.3.2.3 Privacy preservation through facial de-identification with simultaneous emotion preservation

This work [3], proposes an automated face de-identification algorithm that takes as input a facial image and generates a new face that preserves the emotion and non-biometric facial attributes of a target face. A proxy set of a large collection of artificial faces is generated by StyleGAN [69] and the most appropriate face from the proxy set are selected that has a facial expression and pose similar to that of the target face. The faces in the proxy set are artificially generated, and hence the face selected from this set is completely anonymous. To retain the non-biometric attributes of the target face, a generative adversarial network (GAN) [70] is employed with a suitable loss function that fuses the non-biometric attributes of the target face with the face selected from the proxy set to obtain the final de-identified face. Figure 1.4 describes the overall algorithm utilized by this approach. This work is fascinating as this is the only prior approach that realistically transforms the image into an anonymized version while preserving the emotional and non-biometric information.

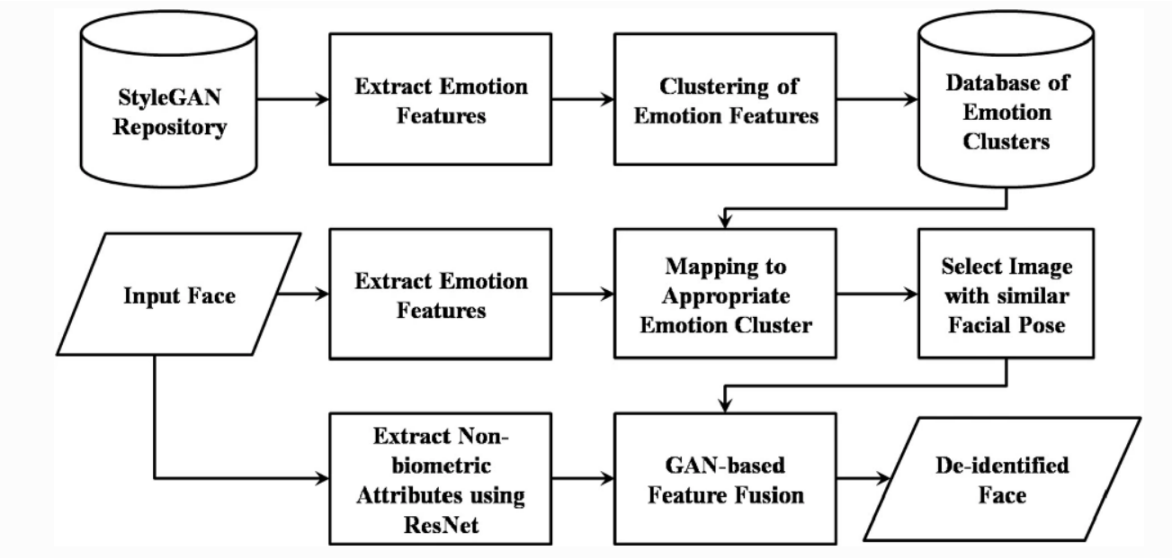


Figure 1.4: Block diagram of the approach proposed by Agarwal *et al.* [3]

1.4 Novelty of this thesis

The prior works described in section 1.3.2 can be subjected to the following critique:

1. The approach proposed by Narula *et al.*[2] aims at reducing the user-dependent identity information learnt by the convolutional layers, since it hypothesises that the convolutional base retains the largest degree of the spatial information of an image. While this approach is effective in reducing the user-dependent information being stored in the model parameters, it still leaves vulnerabilities in the client-side. Since this approach does not transform the original image, it is possible to compromise user identity by directly targeting the user's device. This is not outside the realm of possibility, since social hacking [71] and phishing [72] are two of the most common hacking techniques and both of them directly target the user's device. Also, this approach creates an intermediate image that has been pre-processed using the privacy-preserving CNN layers, which can be stored in the cloud for further analysis. While this intermediate image almost completely eliminates the user identity based information, it is not human readable, and hence cannot be used for non-ML based approaches.

2. The approach proposed by Agarwal *et al.*[3] alleviates this problem by transforming the user's face using an image-to-image translation approach. For every user's face, it identifies a proxy face that closely resembles the user's original face. This approach can be potentially problematic when applied to the real world, since there are infinitely many possibilities of facial configurations and poses, and the corpus of data required to effectively map the faces can quickly get arbitrarily large. If the set of proxy faces is too small with respect to the general population, then the anonymized images will start looking similar to each other, which reduces the generalizability of the model.

3. The DeepPrivacy approach, proposed by Hukkelas *et al.* [1], addresses this problem of infinite combinations, since it relies on a one-shot approach that directly performs image-to-image translation from the original image to an anonymized image. This model is trained using FDF dataset [73], which is an exhaustive dataset consisting of 1.3 million real-world

images. While this ensures generalizability, DeepPrivacy does not explicitly preserve the emotion based information in the original image, since it performs a complete anonymization of the original image, which also eliminates the emotion based information from the face.

This thesis advances the existing literature in the following ways:

1. We identified and conducted a detailed analysis on facial landmarks, and quantified the user identity and emotion based information conveyed by each landmark. We then identified the landmarks that contains the most emotion-based information and the least identity-based information so that we can maximize the emotional information that we extract from the faces while preserving user privacy.

2. We designed a conditional GAN similar to the one proposed by Hukkelas *et al.* [1] and we used the above selected landmarks as the conditions for both the generator and the discriminator. This enabled us to generate realistic faces that carry the least identity based information and the most emotion based information when compared to the original faces.

This thesis also delivers some key engineering contributions that can be useful for future research:

1. We created a modified version of the Flickr Diverse Faces (FDF) dataset by (1) filtering out the low quality images that does not have a clear face and (2) labelling the facial landmarks for each face present in the each of the image.

2. We developed a command line interface (CLI) that can anonymize multiple images in one shot while retaining the emotion-based information from each of the faces.

The code for this thesis will be made available online.

1.5 Research objectives and proposed approach

In the light of the above challenges, this thesis will focus on three main research questions:

1. Is it possible to quantify the amount of relevant emotion and identity related information carried by different facial landmarks?

2. Can we condition machine learning models on the aforementioned landmarks to conduct privacy-preserving emotion recognition?

3. To what extent are the identified landmarks generalizable across datasets?

First, we identify different facial landmarks that contribute to the overall structure of the face. The landmarks are extracted using ensemble regression trees [74] with the help of Dlib machine learning toolkit [75]. We locate 68 landmarks across the face that denotes the location the eyes, eyebrows, nose, mouth, ears, jawline and chin. We perform correlation analysis by using ficher scoring to quantify the landmarks based on the amount of emotion and identity based information carried by each landmark. We assign a composite score to each landmark which is proportional to the emotion related fisher score and inversely proportional to the user identity related fisher score. We rank the landmarks based on this score and select the top k landmarks for our further analysis.

After selecting the k landmarks, we create a modified version of FDF dataset [73] where we identify these k landmarks for each of the images in dataset. We then modify the DeepPrivacy network to utilize these k landmarks as the pose information for each image. After training DeepPrivacy till the generator and adversarial losses converge, we create a modified version of YALE [76] and JAFFE [77] where we identify the same k landmarks for each of the images along with the bounding box for each face inside the image. Then, we split this modified dataset into train and validation images. We then fine-tune the pre-trained network using the train imaged of YALE and JAFFE dataset seperately, till the losses converge to a stable value.

We use the fine-tuned models to perform image-to-image translation and anonymize the validation images of YALE and JAFFE dataset. We train a classifier using ResNet to perform emotion recognition and user identification using the train images of YALE and JAFFE dataset. We first use this trained ResNet model to classify the original validation images from YALE and JAFFE. This will serve as the baseline for our further analysis. We again this trained ResNet model to classify the anonymized validation images of YALE and JAFFE and compare this results with our baseline classification. We repeat this experiment with different values of k. Our qualitative and quantitative results obtained in the YALE and JAFFE dataset will be discussed and examined to check if they

demonstrate the feasibility of the proposed framework for user-privacy in image-based emotion recognition.

2. METHODOLOGY

In the following, we will describe our work on quantifying user anonymity and designing a privacy-aware machine learning system for emotion recognition.

2.1 Landmark detection in faces

The first step of our approach is landmark detection. We identify the location of the components of our face which are responsible for conveying the most information about a person, which includes, but not limited to the person's eyes, nose, mouth, eyebrows, jawline, chin and the ears. In order to perform this, we use a machine learning model that makes use of ensemble regression trees [74], which is implemented using dlib machine learning toolkit [75]. This keypoint detector model extract 68 landmarks across the face that denotes the location of the facial components mentioned above. Figure 2.1 represents the 68 landmarks and their corresponding locations in the face, and the same 68 landmarks identified and superimposed on a stock image.

2.2 Analysing the impact of different facial landmarks on the emotion and user identity

Once we identify these 68 landmarks for each of our faces, our next task is to quantify the amount of identity and emotion information conveyed by each of these points. To do this, we make use of fisher scoring. The following is the step by step method that we followed to quantify the amount of information conveyed by each of the landmarks:

1. Extract the coordinates of each of the landmark. For each landmark, we have 2 coordinates (x, y), and since we identify 68 landmarks per face, we would have $68 \times 2 = 136$ coordinate values in total. These coordinates will serve as our features in our analysis. We normalize these coordinate values between 0 and 1 for ease of computation. So, for each image we have a 136 length feature vector. Therefore, for n images, we will have a feature matrix of $n \times 136$.

2. For these n images, we obtain the emotion label y_e and identity label y_i.

(a) 68 facial landmarks identified by the face detector [78].

(b) 68 facial landmarks identified from a sample stock image .

Figure 2.1: The 68 facial landmarks identified by the face detector and the location of these landmarks on a sample stock image.

3. For each normalized coordinate p, the corresponding emotion-related fisher score S_{ep} is computed as:

$$S_{ep} = \frac{\sum_{j \in y_e} n_j * (\mu_{pj} - \mu_p)^2}{\sum_{j \in y_e} n_j * \rho_{pj}^2} \tag{2.1}$$

where μ_{pj} and ρ_{pj} are the mean and variance of the p^{th} feature in the j^{th} class, where j represents an emotion present in the y_e label vector. Similarly, n_j represents the number of instances of emotion j in the dataset, and μ_p is the overall, class-independent mean of the p^{th} feature.

4. Similarly, for each normalized coordinate p, the corresponsing identity-related fisher score S_{ip} is computed as:

$$S_{ip} = \frac{\sum_{j \in y_i} n_j * (\mu_{pj} - \mu_p)^2}{\sum_{j \in y_i} n_j * \rho_{pj}^2} \tag{2.2}$$

5. The most desirable landmarks are the ones that carry the most emotion based information and the least identity based information. So, the coordinate should have a high emotion

related fisher score S_{ep} and a low identity related fisher score S_{ip}. But keep in mind that p represents the coordinate, not the landmark, since the landmark comprises of both the x and y coordinates. Therefore, in order to rank the landmarks based on the desirability, we compute the composite score C_l as follow:

$$c_l = \frac{S_{ep_x} + S_{ep_y}}{S_{ip_x} + S_{ip_y}} \tag{2.3}$$

where,

S_{ep_x} represents the emotion based fisher score of the x coordinate of the landmark l

S_{ep_y} represents the emotion based fisher score of the y coordinate of the landmark l

S_{ip_x} represents the identity based fisher score of the x coordinate of the landmark l

S_{ip_y} represents the identity based fisher score of the y coordinate of the landmark l

6. We then sort the 68 landmarks based on their composite score C_l and select the top k landmarks that have the k highest composite scores.

2.3 Adversarial approach to anonymizing facial images

After selecting the k best landmarks with the highest composite score, we utilize DeepPrivacy [1] to perform image generation and anonymize the image while preserving a high degree of emotional information. We do that by using the k landmarks as the pose information for the network. The pose information is represented using one-hot encoding, which will help reduce the number of parameters required by the model.

2.3.1 Generator

Figure 2.2 represents the generator architecture used by DeepPrivacy. The input pipeline for the generator is the same as the one described in figure 1.1. The generator is fed with the images where the face is greyed out. At each stage of the upsampling layer, we concatenate this one-hot pose information. Therefore, the generator never observes the original image, therefore the privacy sensitive information is not translated onto the anonymized image. However, since we are using

the k landmarks as pose information, it will use these landmarks to reconstruct the face. Since these landmarks are optimized to carry the most emotion-based information and the least identity based information, the resulting image will retain the same emotional information as the input image.

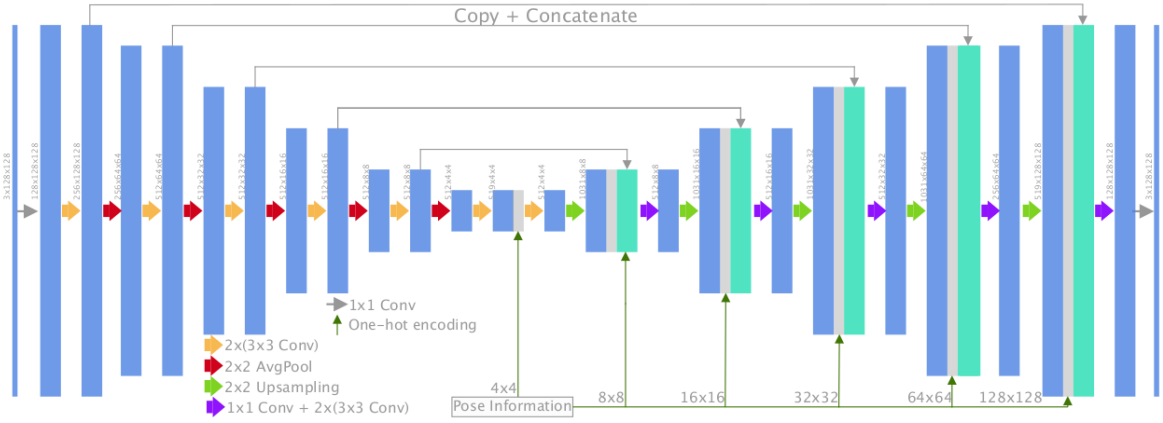


Figure 2.2: Generator architecture used by DeepPrivacy [1]. The k landmarks is used as the pose information for both the generator and discriminator network.

2.3.2 Discriminator

The discriminator architecture is similar to that of DeepPrivacy [1]. The discriminator's input includes the background information as conditional input to the start of the discriminator, making the input image have six channels instead of three. The discriminator's task is to identify whether the input image is an original image or an anonymized image. The background information ensures that the generated anonymized image has the same background information. In other words, only the facial region is anonymized, and the background is untouched. To prevent model collapse and training instabilities, we used the improved Wasserstein loss [79].

2.3.3 Training

The training is monitored by visualizing and measuring the quality of the output images after each epoch. After each epoch, 2 images from the validation data is anonymized and the output is stored and visualized. We also calculate multiple metrics to assess the quality of the generated

images at the end of each epoch. We calculate the average $l1$ norm, $l2$ norm, peak signal to noise ratio (PSNR) [80] and learned perceptual image patch similarity (LPIPS) [81] to qualtitatively measure the quality of the output images with respect to the input image. Once these values converge to an acceptable value and once the anonymized images are of sufficient quality, we halt the training process.

2.4 Evaluation of modified DeepPrivacy

Evaluation of the modified DeepPrivacy will be done on 2 grounds: (1) amount of emotion information retained by the anonymized image, (2) amount of user identity information retained by the image. Ideally, we would want the image to contain most of the emotion based information and the least user identity information. To evaluate this, we perform emotion and user classification on the original and anonymized image and compare the results.

To perform emotion and user classification, we make use of a type of convolutional neural networks (CNN) known as ResNets. ResNet stands for residual network [82], which makes use of residual blocks for learning complex non-linear patterns. ResNets were introduced in 2015 and they have been used in state of the art neural networks ever since. ResNets introduce the concept of "identity shortcut connection" that skips one or more layers. This reduces the risk of overfitting and allows us to stack more layers without degrading the network performance.

We made use of a specific architecture of ResNets known as ResNet9. This architecture has 3 residual blocks, with each block containing 3 convolutional layers followed by batch normalization and ReLU activation. Figure 2.3 represents the residual block utilized by our network.

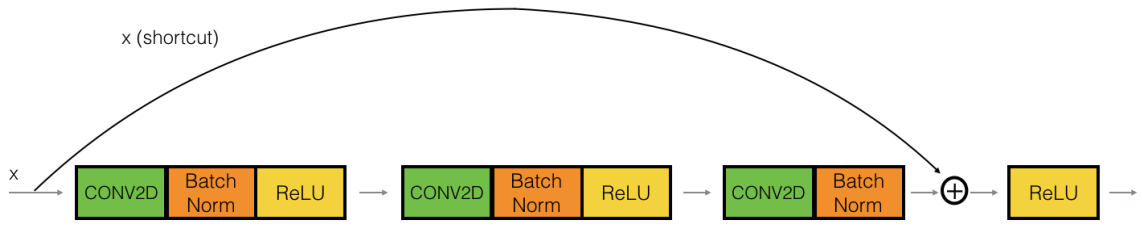


Figure 2.3: Residual block used for ResNet9 [4].

Using this network, we evaluate the DeepPrivacy network through the following steps:

1. Perform emotion classification on the train images of the dataset (lets call this classifier E). Perform user identity classification on the train images of the dataset (lets call this classifier I).

2. Evaluate classifier E and classifier I by classifying the validation images of the dataset and calculating the validation accuracies A_{Eb} and A_{Ib}. These 2 accuracy scores will serve as our baseline.

3. Anonymize the validation images by using the pre-trained DeepPrivacy model.

4. Classify the anonymized validation images using classifier E and calculate the accuracy A_E by using the emotion labels of the corresponding original validation images. Classify the anonymized validation images using classifier I and calculate the accuracy A_I by using the identity labels of the corresponding original validation images.

We compare A_E with A_{Eb} to understand the amount of emotion based information retained and compare A_I with A_{Ib} to understand the amount of user identity based information retained by the DeepPrivacy anonymizer. It is desirable to have the value of A_E as high as possible and A_I as low as possible.

3. EXPERIMENTATION

3.1 Datasets description

We used the following 3 datasets for our experiments.

3.1.1 FDF dataset

Flickr Diverse Faces (FDF) dataset [73] is a comprehensive dataset that contains images of faces in the wild. It consists of 1.47M human faces with a minimum resolution of 128 x 128, containing facial keypoints and a bounding box annotation for each face. The dataset has a vast diversity in terms of age, ethnicity, facial pose, image background, and face occlusion. Each face in the image is annotated with a tight bounding box and 7 keypoints representing the facial posture. Figure 3.1 contains some randomly picked examples from the dataset.

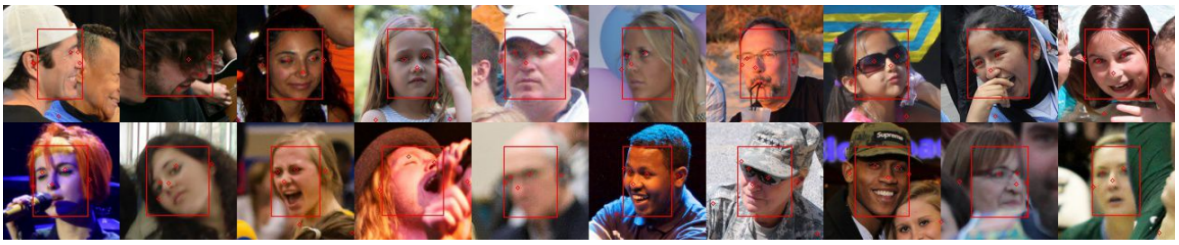


Figure 3.1: Example images in the FDF dataset [1]

We had to pre-process the FDF dataset with our selected landmarks before using it for training our model. We replaced the 7 landmarks with the k landmarks we selected for the particular task. We used images that contained only 1 face, for the sake of simplicity. We used dlib shape predictor to detect the 68 facial landmarks represented by figure **??**, and selected the k landmarks that are of interest for the particular task. We selected 50,000 images from the training dataset and 1000 images from the test dataset, which are of high quality where the face detector was able to identify

the landmarks with a confidence level of over 95%. Confidence level is calculated as follows:

$$\text{Confidence} = \frac{\text{number of landmarks located within the facial bounding box}}{\text{total number of landmarks detected for the face}} \tag{3.1}$$

For example, if we want to identify 68 facial landmarks for each image, then the selected image should have atleast 65 landmarks located within the facial bounding box.

3.1.2 YALE dataset

YALE dataset [76] is a dataset that contains images that are in close proximity to the face, and some of the faces are annotated with the information about the identity of the subject and the emotion represented by the face. There are 60 images with both the emotion and identity labels, which span across 15 male and female users and 4 emotion classes (happy, sad, normal and surprised). Since this is not adequate to train the adversarial model, the dataset was augmented using random rotation, horizontal flip, and random noise addition. This increased the size of the dataset to 2067 images, which are divided into train and validation datasets with a ratio of 4:1. The dataset is fairly balanced across the different emotion and user identity classes.

3.1.3 JAFFE dataset

JAFFE dataset [77] a dataset that contains images specifically from japanese female subjects. The faces are in close proximity to the camera and contain information about the subject's identity and the emotion information represented by each image. There are multiple images with different emotions for the same set of subjects. There are 10 subjects with 7 emotions (neutral, sadness, surprise, happiness, fear, anger, and disgust) with a total of 213 images. The images were augmented in the same way as YALE dataset, which resulted in a total of 3067 images which are divided into train and validation datasets with a ratio of 4:1. The dataset is also fairly balanced across the different emotion and user identity classes.

3.2 Landmarks selection

The following landmarks were selected to train the DeepPrivacy model:

1. All 68 landmarks. Lets call this k_{68}

2. The landmarks having above average composite score in the YALE. Lets call this k_{yale}^{mean}.

3. The landmarks having above average composite score in the JAFFE. Lets call this k_{jaffe}^{mean}.

4. The top 12 landmarks with the highest composite score from YALE. Lets call this k_{yale}^{12}.

5. The top 12 landmarks with the highest composite score from JAFFE. Lets call this k_{jaffe}^{12}.

3.3 Training DeepPrivacy model

After selecting the landmarks, we trained the following models using the DeepPrivacy network.

1. Pre-train using FDF dataset with k_{68} landmarks

   (a) Finetune the model with YALE dataset with the same k_{68} landmarks.

   (b) Finetune the model with JAFFE dataset with the same k_{68} landmarks.

2. Pre-train using FDF dataset with k_{yale}^{mean} landmarks

   (a) Finetune the model with YALE dataset with the same k_{yale}^{mean} landmarks.

   (b) Finetune the model with JAFFE dataset with the same k_{yale}^{mean} landmarks.

3. Pre-train using FDF dataset with k_{jaffe}^{mean} landmarks

   (a) Finetune the model with YALE dataset with the same k_{jaffe}^{mean} landmarks.

   (b) Finetune the model with JAFFE dataset with the same k_{jaffe}^{mean} landmarks.

4. Pre-train using FDF dataset with k_{yale}^{12} landmarks

   (a) Finetune the model with YALE dataset with the same k_{yale}^{12} landmarks.

   (b) Finetune the model with JAFFE dataset with the same k_{yale}^{12} landmarks.

5. Pre-train using FDF dataset with k_{jaffe}^{12} landmarks

(a) Finetune the model with YALE dataset with the same k^{12}_{jaffe} landmarks.

(b) Finetune the model with JAFFE dataset with the same k^{12}_{jaffe} landmarks.

For training DeepPrivacy, we used an NVIDIA RTX 2060 GPU with 6GB of memory. The batch size was used was 2. Training on FDF dataset took approximately 18 hours for 1 set of landmarks. Finetuning on YALE dataset took approximately 5 hours and finetuning on JAFFE took approximately 7 hours for the same set of landmarks. Automatic mixed precision implemented by NVIDIA apex was used during the training process in order to reduce the amount of GPU memory ulitized and boost the training speed.

3.4 Evaluating DeepPrivacy model

The fine-tuned DeepPrivacy models for various landmarks were evaluated using the method described in section 2.4. The ResNet classifier was trained using an NVIDIA RTX 2060 GPU with 6GB of memory. The batch size was used was 128. We used a learning rate scheduler, which will change the learning rate after every batch of training. The strategy for scheduling that we used is called the "One Cycle Learning Rate Policy", which involves starting with a low learning rate, gradually increasing it batch-by-batch to a high learning rate for about 30% of epochs, then gradually decreasing it to a very low value for the remaining epochs. Figure 3.2 represents the learning rate scheduler that was used for this purpose.

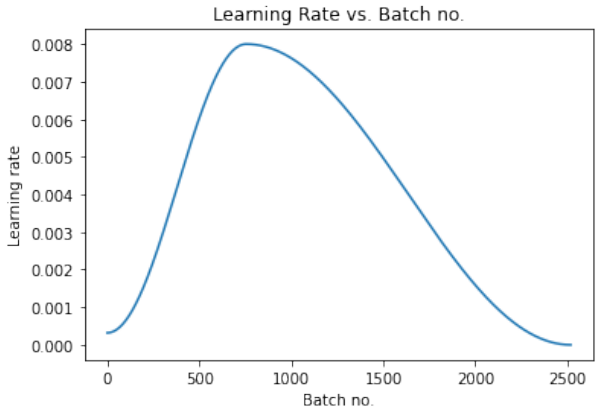


Figure 3.2: Learning rate scheduler used in training the ResNet classifier

We used weight decay in order to prevent the weights from becoming too large by adding the weight values to the loss function. We also utilized gradient clipping in order to prevent the gradients from jumping momentarily to a high value and ruining the model mid training. This limits the gradient values within a particular range and prevents it from increasing arbitrarily. The training was carried out for a maximum of 30 epochs, which took approximately 3 minutes for each individual classification task.

4.1 Objective 1: Quantifying the amount of emotion and identity information carried by facial landmarks

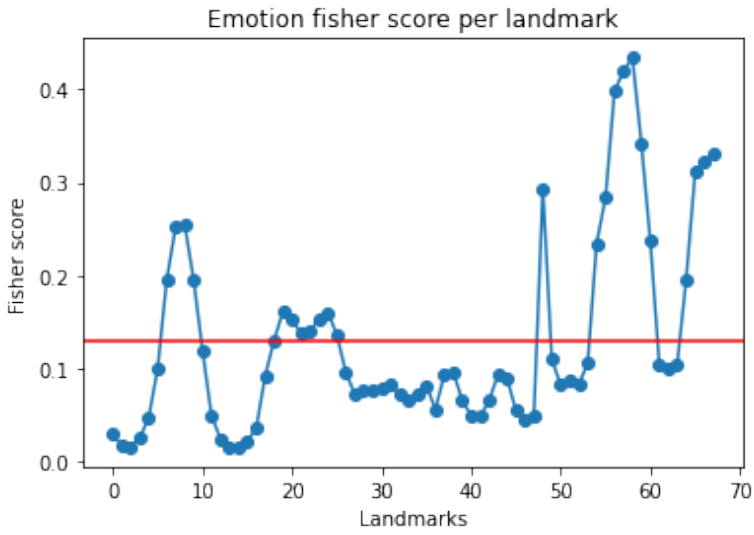
To quantify the emotion and identity information carried by each of the facial landmarks, we follow the steps mentioned in section 2.2. We find the composite scores for the landmarks separately for YALE and JAFFE dataset because both the datasets have different set of users and emotion classes.

4.1.1 Quantifying landmarks using YALE dataset

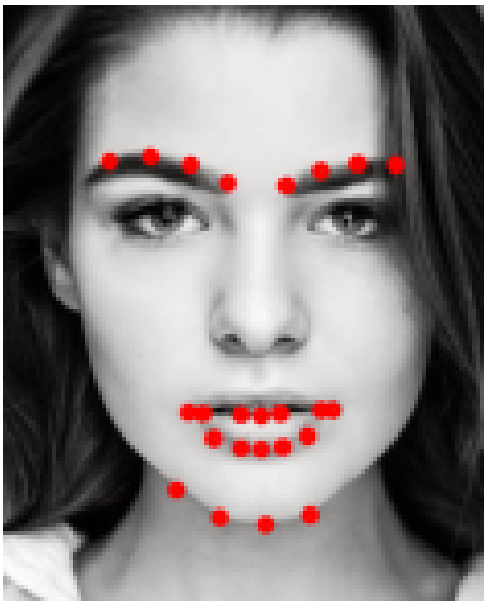
Figure 4.1a represent the fisher scores of the facial landmarks with respect to the emotion labels in the YALE dataset. As from the graph, it is evident that some of the landmarks carry more emotion information compared to the rest. The landmarks at the inner eyebrows and the lower lip have a score of above 0.12. This result makes sense, since our eyebrows and lips are sensitive to our emotional expression and hence they carry a lot of emotion based information about the face. Figure 4.1b represents these landmarks with these scores.

Figure 4.2a represent the fisher scores of the facial landmarks with respect to the identity labels in the YALE dataset. As from the graph, it is evident that some of the landmarks carry more identity-based information compared to the rest. The landmarks at the inner eyebrows and the lower lip have a score of above 1.1. This result makes sense, since our eyes, eyebrows and the shape of our nose can help us uniquely identify human faces and hence they carry a lot of identity based information about the face. Figure 4.2b represents these landmarks with these scores.

Figure 4.3a represent the overall composite scores of each landmark in the YALE dataset. Here, we have selected landmarks that have an overall score of greater than 0.15. Figure 4.3b represents these landmarks with these scores.

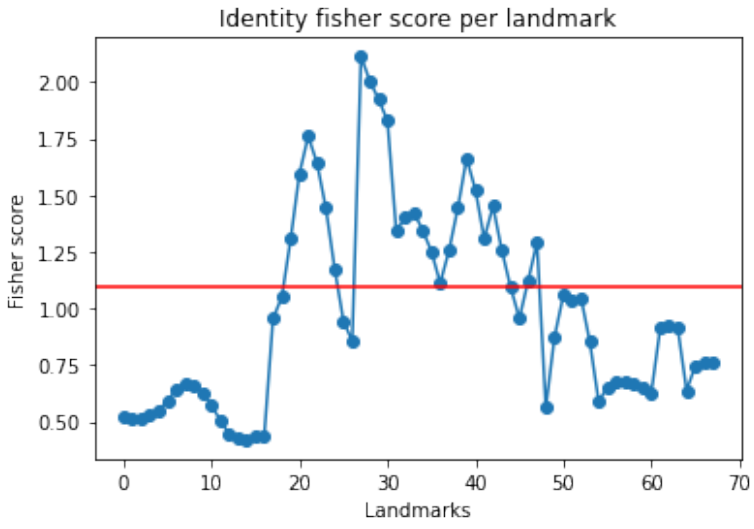


(a) Graph representing the fisher score of the facial landmarks with respect to the emotion labels in the YALE dataset

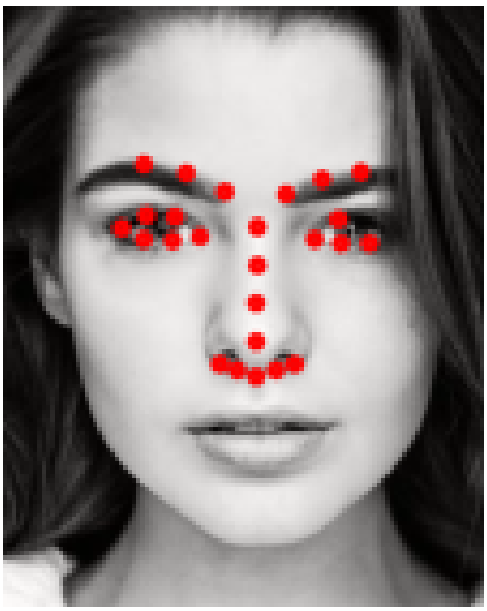


(b) Landmarks that have a high fisher score with respect to emotion.

Figure 4.1: Fisher scores of each facial landmark with respect to emotion and the equivalent landmarks that have a high fisher score.

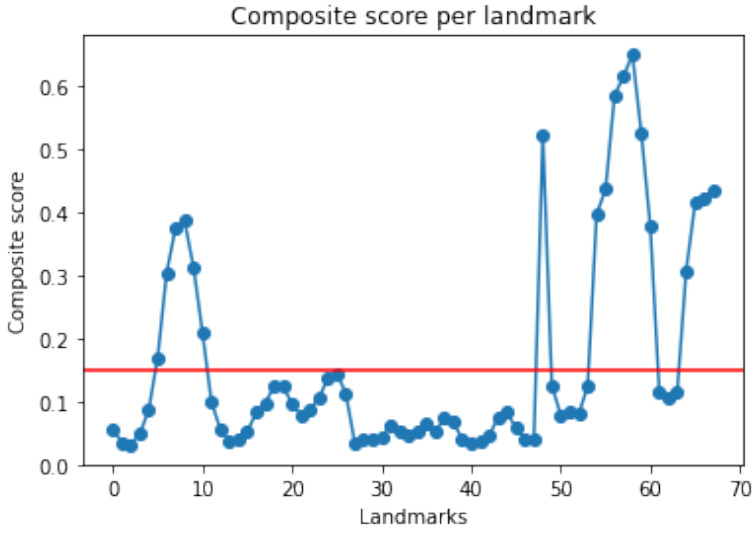


(a) Graph representing the fisher score of the facial landmarks with respect to the identity labels in the YALE dataset

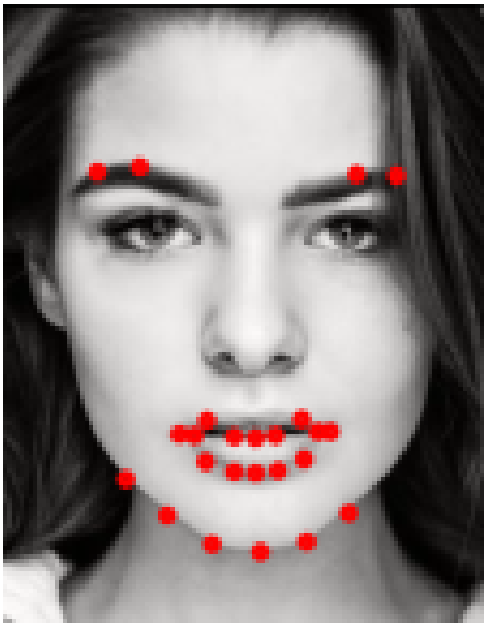


(b) Landmarks that have a high fisher score with respect to user identity.

Figure 4.2: Fisher scores of each facial landmark with respect to user identity and the equivalent landmarks that have a high fisher score.



(a) Graph representing the composite score of each landmarks in the YALE dataset



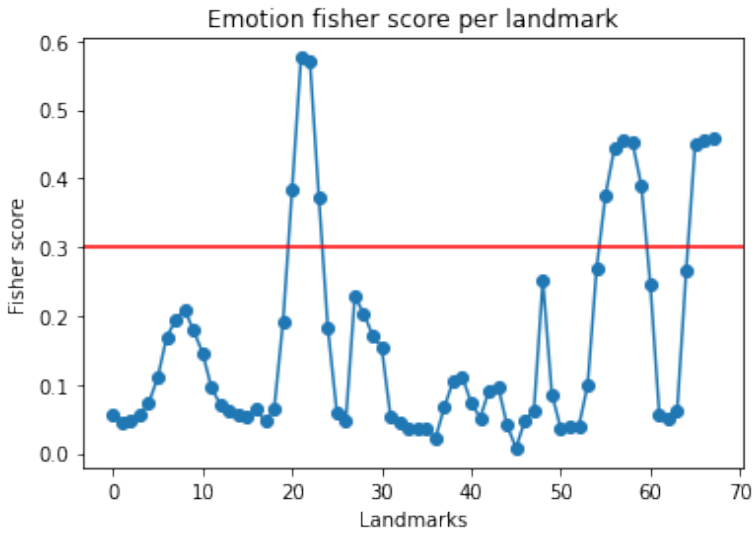
(b) Landmarks that have a high composite score.

Figure 4.3: Composite scores of each facial landmark and the equivalent landmarks that have a high composite score.

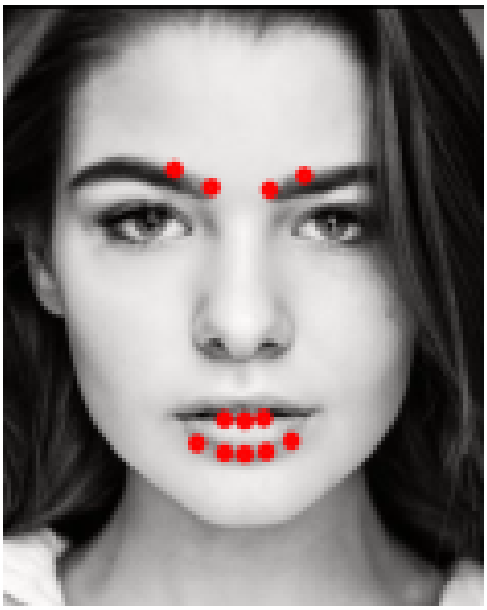
4.1.2 Quantifying landmarks using JAFFE dataset

Figure 4.4a represent the fisher scores of the facial landmarks with respect to the emotion labels in the JAFFE dataset. As from the graph, it is evident that some of the landmarks carry more emotion information compared to the rest. The landmarks at the inner eyebrows and the lower lip have a score of above 0.3. This result makes sense and it is similar to the results that we obtained from YALE dataset 4.1b, since our eyebrows and lips are sensitive to our emotional expression and hence they carry a lot of emotion based information about the face. Figure 4.4b represents these landmarks with these scores.

Figure 4.5a represent the fisher scores of the facial landmarks with respect to the identity labels in the JAFFE dataset. As from the graph, it is evident that some of the landmarks carry more identity-based information compared to the rest. The landmarks at the nose, upper lip and chin have a score of above 0.18. This is different from the results thaat we obtained from YALE dataset



(a) Graph representing the fisher score of the facial landmarks with respect to the emotion labels in the JAFFE dataset



(b) Landmarks that have a high fisher score with respect to emotion.

Figure 4.4: Fisher scores of each facial landmark with respect to emotion and the equivalent landmarks that have a high fisher score.

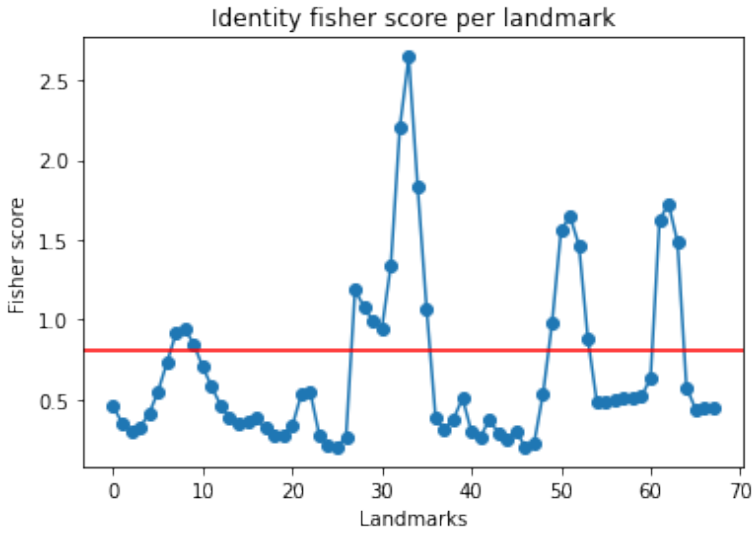
in figure 4.2b, which maybe because JAFFE dataset is not as diverse as YALE, and this may contribute to a skew in the set of landmarks that helps us uniquely identify faces in the dataset. Figure 4.5b represents these landmarks with these scores.

Figure 4.6a represent the overall composite scores of each landmark in the JAFFE dataset. Here, we have selected landmarks that have an overall score of greater than 0.6. Figure 4.6b represents these landmarks with these scores.

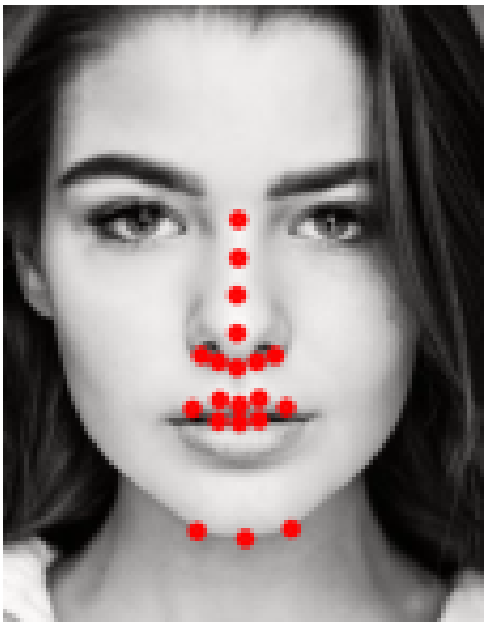
4.2 Objective 2: Employing machine learning models to perform privacy preserving emotion detection using selected landmarks

In order to answer this question, we followed the methodology described in section 2.3. We selected the landmarks mentioned in section 3.2. The first set of landmarks that we used it k_{68}. Let us call this case 1. k_{68} is represented by the figure 2.1. The next set of landmarks were k_{yale}^{mean} and k_{jaffe}^{mean}. Lets call this case 2. These two set of landmarks are represented by the figure 4.7

The next set of landmarks were k_{yale}^{12} and k_{jaffe}^{12}. Lets call this case 3. These two set of

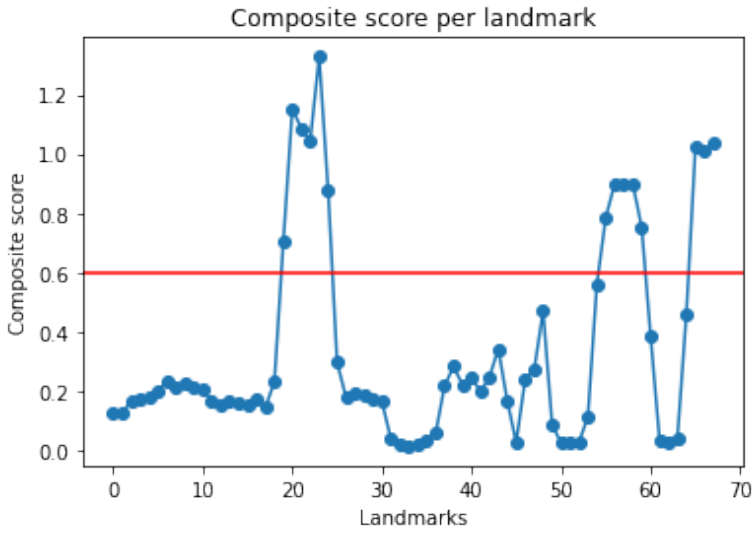


(a) Graph representing the fisher score of the facial landmarks with respect to the identity labels in the JAFFE dataset

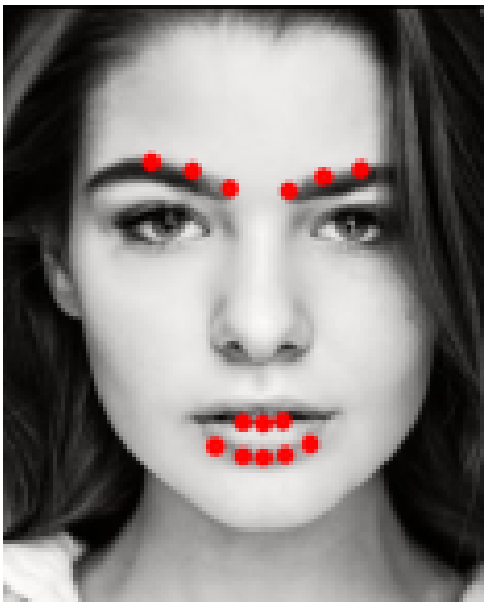


(b) Landmarks that have a high fisher score with respect to user identity.

Figure 4.5: Fisher scores of each facial landmark with respect to user identity and the equivalent landmarks that have a high fisher score.

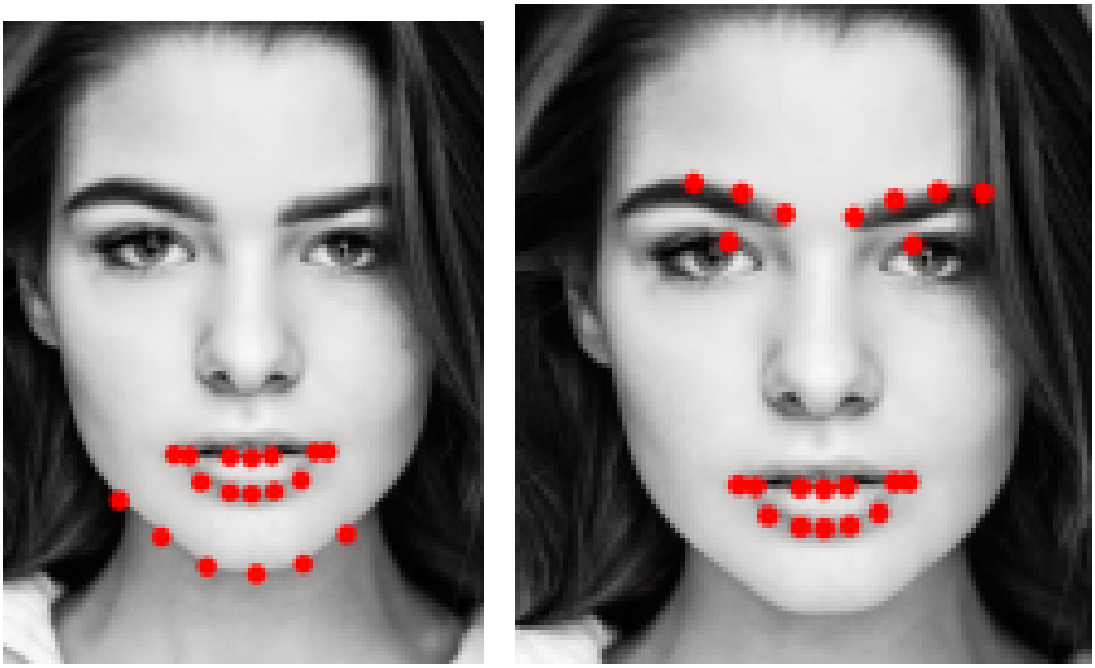


(a) Graph representing the composite score of each landmarks in the JAFFE dataset



(b) Landmarks that have a high composite score.

Figure 4.6: Composite scores of each facial landmark and the equivalent landmarks that have a high composite score.



(a) Case 2a: Landmarks with above average composite score in YALE dataset

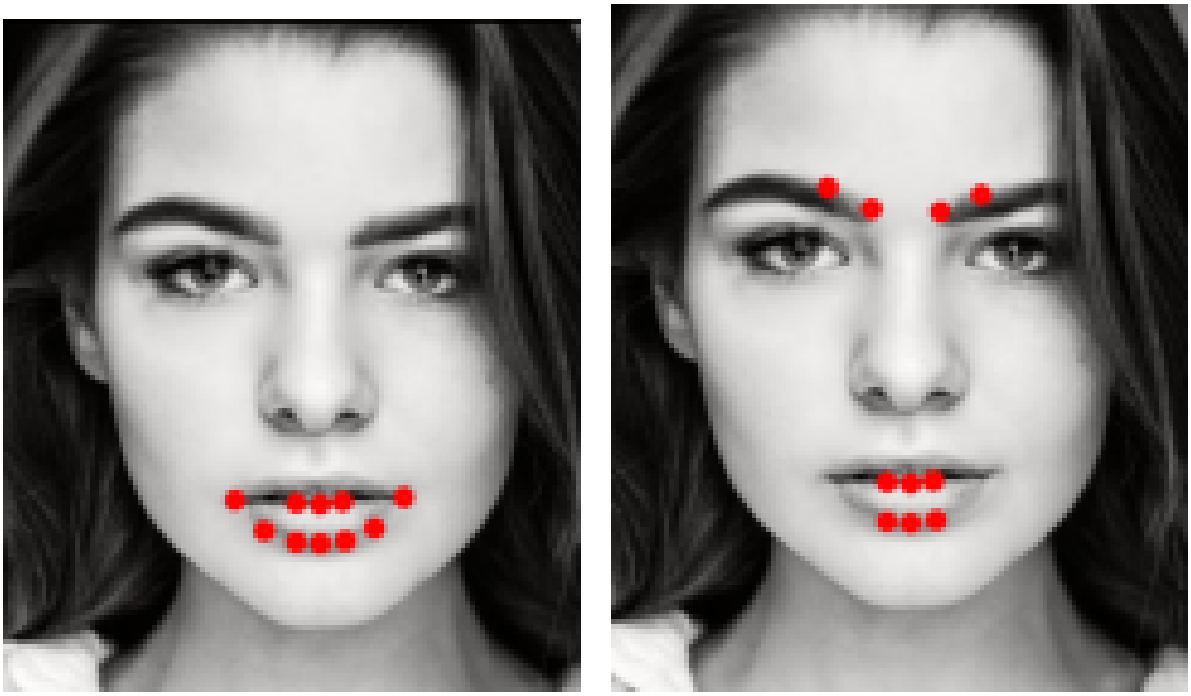
(b) Case 2b: Landmarks with above average composite score in JAFFE dataset.

Figure 4.7: Set of landmarks used for case 2 analysis.

landmarks are represented by the figure 4.8

After selecting the landmarks, we followed the evaluation approach described in section 2.4. Emotion recognition and user identity recognition tasks were performed using the anonymized images trained using the landmarks mentioned in case 1, case2 and case3 and the resulting unweighted accuracies were recorded. Table 4.1 represents the results obtained on YALE dataset and table 4.2 represents the results obtained in JAFFE dataset.

Looking at table 4.1, we can observe that the baseline model is pretty good at understanding the emotional information contained in the image, and shows decent performance when understanding the identity-based information. When we use all the 68 landmarks (Case 1), we are able to reduce the amount of identity information by a considerable margin, but we also lose quite a bit of emotion information as well. We hypothesize that this is because the modified DeepPrivacy network is struggling to optimize for all the 68 landmarks as constraints. And hence it is not able to retain emotion information effectively. When we select only the top landmarks (Case 2a and Case 3a),



(a) Case 3a: Landmarks with the top 12 composite score in YALE dataset

(b) Case 3b: Landmarks with the top 12 composite score in JAFFE dataset.

Figure 4.8: Set of landmarks used for case 3 analysis.

Set of landmarks	Emotion classification score	Identity classification score
Baseline: No landmarks/non-anonymized	83.71%	48.44%
Case 1: All 68 landmarks	38.84 %	30.36%
Case 2a: Landmarks with above average composite score (24)	71.88%	27.34%
Case 3a: Top 12 landmarks with highest composite score	55.47%	10.74%

Table 4.1: Unweighted accuracies of emotion and identity classification using the landmarks from YALE dataset.

Set of landmarks	Emotion classification score	Identity classification score
Baseline: No landmarks/non-anonymized	89.66%	99.55%
Case 1: All 68 landmarks	20.54 %	88.17%
Case 2b: Landmarks with above average composite score (21)	32.37%	79.32%
Case 3b: Top 12 landmarks with highest composite score	25.74%	47.77%

Table 4.2: Unweighted accuracies of emotion and identity classification using the landmarks from JAFFE dataset.

we are able to effectively suppress the identity information, while retaining a considerable chunk of the emotion information. This result is promising since it confirms our earlier assumption that utilizing landmarks with high composite score can help us retain high emotion information while leaving out identity based information.

Taking a look at table 4.1, we can observe that the baseline model is extremely good at understanding the emotional information contained in the image, and shows excellent performance when understanding the identity-based information. When we use all the 68 landmarks (Case 1), a considerable amount of identity information is retained, while we lose a lot of emotion information. This is similar to the results we saw with the YALE dataset, where the modified DeepPrivacy network is struggling to optimize for all the 68 landmarks as constraints. And hence it is not able to retain emotion information effectively, and is retaining too much identity information. When we select only the top landmarks (Case 2b and Case 3b), we are able to suppress the identity information to a considerable extent, while retaining a decent chunk of the emotion information. This result is not as promising as the one obtained using YALE dataset, which maybe because JAFFE dataset is not as ethnically diverse as YALE dataset. We previously saw that this may create a skew in the dataset resulting in landmarks having a different distribution of composite scores.

Figure 4.9 represents some sample visualizations of the best performing anonymizer model that was finetuned using YALE dataset. The first 3 images represent a successful anonymization, where the resulting image represent the same emotion as the original image, but does not resemble

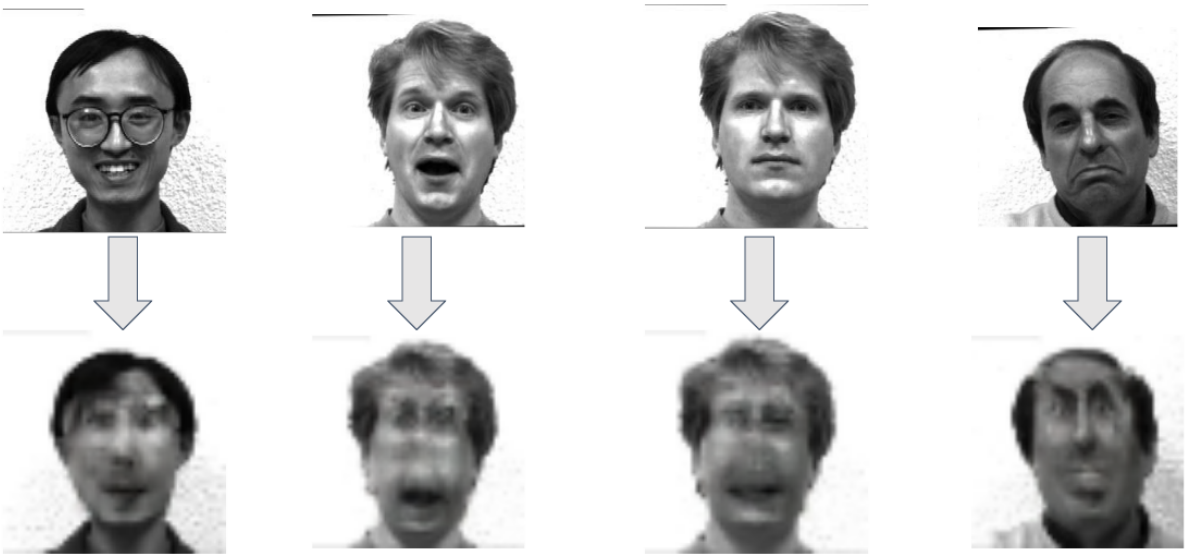


Figure 4.9: Sample results obtained from YALE dataset using the best performing model.

the original image in terms of user identity. The 4th image is an un-successful anonymization, where the resulting image fails to convey the same emotion as the original image. Even though emotional understanding can be subjective based on the user, it is clear from the first 3 images that the original image and the resulting image convey nearly the same emotion information.

Figure 4.10 represents some sample visualizations of the best performing anonymizer model finetuned using the JAFFE dataset. Similar to the earlier example, the first 3 images represent a successful anonymization, and the 4th image represents an unsuccessful anonymization. For JAFFE dataset, the emotion of the transformed images are not as clear as the images observed using YALE dataset. This is consistent with the results obtained in table 4.2.

4.3 Objective 3: Evaluating the generalizability of the selected landmarks across different datasets

In order to evaluate the generalizability of the selected landmarks, we anonymize the images in the YALE dataset with the landmarks chosen using JAFFE dataset and vice versa, and evaluate the images the same way we described in section 2.4. We use the landmarks k_{yale}^{mean} and k_{yale}^{12} on the JAFFE dataset and k_{jaffe}^{mean} and k_{jaffe}^{12} on the YALE dataset. If the landmarks that we chose

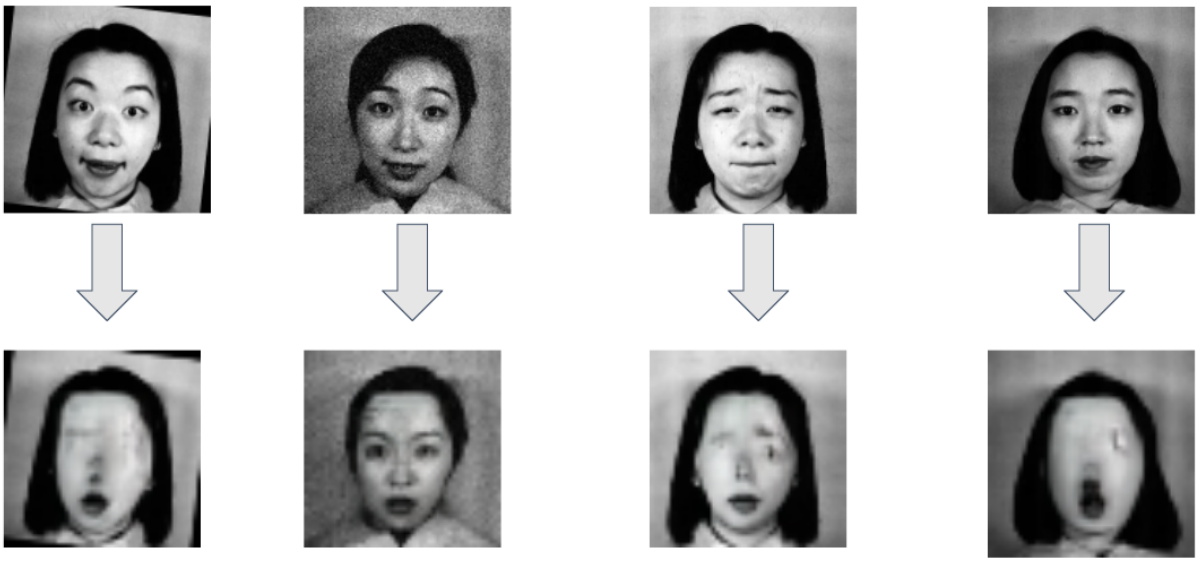


Figure 4.10: Sample results obtained from JAFFE dataset using the best performing model.

Set of landmarks	Emotion classification score	Identity classification score
Baseline: No landmarks/non-anonymized	83.71%	48.44%
Landmarks with above average composite score in JAFFE	51.54%	27.34%
Top 12 landmarks with highest composite score in JAFFE	62.48%	0.0%

Table 4.3: Unweighted accuracies of emotion and identity classification using the landmarks from JAFFE dataset evaluated on YALE dataset.

are generalizable, we would see a similar performance on both the datasets using the same set of landmarks.

Table 4.3 represents the results obtained using landmarks k_{jaffe}^{mean} and k_{jaffe}^{12} on the YALE dataset. We can observe that using JAFFE's landmarks on YALE dataset still yields good results, when compared to table 4.1. We are getting good scores on the emotion classification task and bad scores on identity classification task. This means these landmarks are also doing a good job in retaining emotion information and suppressing identity information.

Table 4.4 represents the results obtained using landmarks k_{yale}^{mean} and k_{yale}^{12} on the JAFFE dataset.

Set of landmarks	Emotion classification score	Identity classification score
Baseline: No landmarks/non-anonymized	89.66%	99.55%
Landmarks with above average composite score in YALE	26.34%	95.31%
Top 12 landmarks with highest composite score in YALE	30.12%	47.29%

Table 4.4: Unweighted accuracies of emotion and identity classification using the landmarks from YALE dataset evaluated on JAFFE dataset.

We can observe that using YALE's landmarks on JAFFE dataset yield similar results, when compared to table 4.1.

Therefore, after comparing the results from table 4.1 to 4.3 and comparing table 4.2 to 4.4, it is clear that the chosen landmarks can be generalizable across datasets.

5. DISCUSSION

In this theses, we proposed an approach for emotion-preserving image anonymization. We identified landmarks across the face and quantified the amount of emotion and identity information carried by each landmark. We then utilized these landmarks to modify the DeepPrivacy network architecture [1] that allows us to generate realistic looking anonymized images that preserve the emotion based information. We then evaluated the quality of the anonymized images by performing emotion and user classification, which proved the working of our approach. In answering our research questions, our results indicate the existence of key landmarks in the face that contribute most to emotion-related information. Using these landmarks, we can achieve moderate to high emotion recognition accuracy and reduced face recognition accuracy for the YALE dataset. However, the proposed approach did not perform that well on the JAFFE dataset, potentially due to the low diversity of data on the same dataset, which potentially prevented the conditional GAN from generalizing. Finally, the landmarks identified as the most emotion-relevant in JAFFE dataset were also capable of conducting privacy preserving emotion recognition in YALE>

Despite the promising results, our current study depicts the following limitations. We trained and evaluated our model using YALE and JAFFE dataset, which were collected in laboratory conditions where the user's emotion is clearly visible. Therefore the data is clean and does not represent the natural emotions that we encounter in real life. In the future, we aim to extend this study for datasets that more closely represent real world faces and emotions, such as the CAS-PEAL Faces database [83], and Indian movie faces database [84] which contain more images and represent a bigger pool of users and expressions. Also, the size of the dataset was a limitation, since the YALE and JAFFE dataset covers only a handful of emotions and test subjects. The lack of a large, diverse dataset with high quality images where each image is annotated with both the emotion and identity information was an impediment to our research.

There are many ways in which one can build upon the work presented in this thesis. In our approach, we consider emotions as static, well defined classes. But in real world, emotions are

much more subtle, and can be quantified via a continuous spectrum rather than discrete categories. As part of future work, one could model emotion recognition as a multi-variate regression problem where we measure the intensity of different emotions in the face. Also, we presented the emotion information to the generator in the form of facial landmarks that carry a high degree of emotion-based information. The landmarks were represented through a one-hot encoding. This means that each landmark is given the same weight while training the generator. However, each landmark carries a varying degree of information, as calculated using the composite score formula represented by equation 2.3. Future work can use a weighted one-hot encoding approach based on the composite score of each landmark so we can add an extra layer of knowledge to the adversarial network. Despite the encouraging result, the proposed approach would benefit from existing baselines that conduct image anonymizations, such as generic conditional GANs or adversarial learning. We can use the default DeepPrivacy to anonymize the YALE nad JAFFE dataset and use this as one such baseline to compare the results from our modified DeepPrivacy. We can also implement the method described by Agarwal *et al.* [3], which is another approach towards differential privacy, and compare it with our method for further analysis. In addition to this, as part of the future work, it would be beneficial to encode in the proposed approach potential interactions between landmarks. Emotional expression relies on the interplay between different facial points, therefore it would be reasonable to encode this interplay when conducting user anonymization.

Even though this approach is focused on the task of privacy preserving emotion recognition, it can be potentially generalized to other key behaviors relevant to mental health. Privacy and user anonymization are inherent issues in several behavioral studies that involve collection of human generated signals to predict psychological and cognitive outcomes. For example, leveraging publicly available data, we can develop privacy-aware systems for stress detection, cognitive demand recognition, and performance prediction [85, 86].

Privacy-preservation is particularly relevant to IoT devices. The proposed framework is computationally expensive and requires a lot of storage space and memory, since the generator network involved the learning of approximately 47.4 million parameters. Despite the high computational

capability and storage capacity of today's IoT devices, efficiency is a prominent issue. Designing compressed privacy- preserving behavior recognition systems remains an open problem, which can be an interesting topic for future work. Potential solutions towards this would be to develop a cloud based API with a pre-trained module which can perform image anonymization. This could alleviate a lot of computational load from the edge devices.

6. CONCLUSIONS

Privacy is of utmost importance when we are dealing with sensitive human generated signals, such as human faces. In this thesis, we first identified facial landmarks and quantified the amount of emotion and identity based information carried by each landmark. Then, we identified desirable landmarks that carry a high degree of emotion based information and a low degree of identity based information. We utilized these landmarks to modify DeepPrivacy to transform images by ignoring the identity based information and retaining emotion based information from the input images, while ensuring that the tranformed image is as realistic as possible. We then evaluated the quality of the anonymized images by performing emotion and user classification. Results obtained using YALE and JAFFE dataset indicate the feasibility of our proposed approach in anonymizing facial images while retaining sufficient amounts of emotion based information. Our approach can be used to implement safer, more privacy-aware emotion recognition and other behavior recognition systems.

REFERENCES

[1] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International Symposium on Visual Computing*, pp. 565–578, Springer, 2019.

[2] V. Narula, T. Chaspari, *et al.*, "An adversarial learning framework for preserving users' anonymity in face-based emotion recognition," *arXiv preprint arXiv:2001.06103*, 2020.

[3] A. Agarwal, P. Chattopadhyay, and L. Wang, "Privacy preservation through facial de-identification with simultaneous emotion preservation," *Signal, Image and Video Processing*, vol. 15, no. 5, pp. 951–958, 2021.

[4] R. Sharma, "Natural image classification using resnet9 model," Jul 2020.

[5] P. Tzirakis, S. Zafeiriou, and B. Schuller, "Real-world automatic continuous affect recognition from audiovisual signals," in *Multimodal Behavior Analysis in the Wild*, pp. 387–406, Elsevier, 2019.

[6] S. Brave and C. Nass, "Emotion in human-computer interaction," in *The human-computer interaction handbook*, pp. 103–118, CRC Press, 2007.

[7] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[8] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European conference on speech communication and technology*, 2003.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000.

[10] V. A. Petrushin, "Emotion recognition in speech signal: experimental study, development, and application," in *Sixth international conference on spoken language processing*, 2000.

[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[12] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiaion.," in *LREC*, pp. 1123–1126, 2006.

[13] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[14] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face.," *Journal of personality and social psychology*, vol. 37, no. 11, p. 2049, 1979.

[15] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 878–883, IEEE, 2011.

[16] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[17] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PloS one*, vol. 12, no. 5, p. e0177239, 2017.

[18] S. N. Shivhare and S. Khethawat, "Emotion detection from text," *arXiv preprint arXiv:1205.4944*, 2012.

[19] H. Binali, C. Wu, and V. Potdar, "Computational approaches for emotion detection in text," in *4th IEEE International Conference on Digital Ecosystems and Technologies*, pp. 172–177, IEEE, 2010.

[20] L. Canales and P. Martínez-Barco, "Emotion detection from text: A survey," in *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pp. 37–43, 2014.

[21] A. Agrawal and A. An, "Unsupervised emotion detection from text using semantic and syntactic relations," in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 346–353, IEEE, 2012.

[22] S. Kedar, D. Bormane, A. Dhadwal, S. Alone, and R. Agarwal, "Automatic emotion recognition through handwriting analysis: a review," in *2015 International Conference on Computing Communication Control and Automation*, pp. 811–816, IEEE, 2015.

[23] L. Likforman-Sulem, A. Esposito, M. Faundez-Zanuy, S. Clémençon, and G. Cordasco, "Emothaw: A novel database for emotional state recognition from handwriting and drawing," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 2, pp. 273–284, 2017.

[24] K. K. Amend and M. S. Ruiz, *Handwriting analysis: The complete basic book*. Red Wheel/Weiser, 2000.

[25] S. Mutalib, R. Ramli, S. A. Rahman, M. Yusoff, and A. Mohamed, "Towards emotional control recognition through handwriting using fuzzy inference," in *2008 International Symposium on Information Technology*, vol. 2, pp. 1–5, IEEE, 2008.

[26] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[27] G. Simcock, L. T. McLoughlin, T. De Regt, K. M. Broadhouse, D. Beaudequin, J. Lagopoulos, and D. F. Hermens, "Associations between facial emotion recognition and mental health in early adolescence," *International journal of environmental research and public health*, vol. 17, no. 1, p. 330, 2020.

[28] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351–6358, 2013.

[29] N. Kshetri and J. Voas, "Cyberthreats under the bed," *Computer*, vol. 51, no. 5, pp. 92–95, 2018.

[30] A. Yankouskaya, G. W. Humphreys, and P. Rotshtein, "The processing of facial identity and expression is interactive, but dependent on task and experience," *Frontiers in human neuroscience*, vol. 8, p. 920, 2014.

[31] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International conference on computer graphics, simulation and modeling*, pp. 135–138, 2012.

[32] V. Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[33] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *2006 IEEE international symposium on circuits and systems*, pp. 4–pp, IEEE, 2006.

[34] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion detection using mfcc and cepstrum features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015.

[35] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in *Interspeech*, pp. 2803–2807, 2019.

[36] T. Surasak, I. Takahiro, C.-h. Cheng, C.-e. Wang, and P.-y. Sheng, "Histogram of oriented gradients for human detection in video," in *2018 5th International conference on business and industrial research (ICBIR)*, pp. 172–176, IEEE, 2018.

[37] R. Ebrahimzadeh and M. Jampour, "Efficient handwritten digit recognition based on histogram of oriented gradients and svm," *International Journal of Computer Applications*, vol. 104, no. 9, 2014.

[38] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C. L. Tan, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognition*, vol. 51, pp. 125–134, 2016.

[39] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based hog features," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 884–888, IEEE, 2011.

[40] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, pp. 586–587, IEEE Computer Society, 1991.

[41] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 80–80, IEEE, 2004.

[42] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[43] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1215–1223, IEEE, 2017.

[44] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6, Ieee, 2017.

[45] J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan, *et al.*, "On robustness and transferability of convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021.

[46] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?," *arXiv preprint arXiv:1608.08614*, 2016.

[47] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[48] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, IEEE, 2021.

[49] V. Narula, *A Computational Framework for Exploring and Mitigating Privacy Risks in Image-Based Emotion Recognition*. PhD thesis, 2020.

[50] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[51] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[52] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, *et al.*, "Introducing the voiceprivacy initiative," *arXiv preprint arXiv:2005.01387*, 2020.

[53] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[54] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[55] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," *arXiv preprint arXiv:1908.10256*, 2019.

[56] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[57] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," 2020.

[58] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *arXiv preprint arXiv:2109.00648*, 2021.

[59] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460*, 2017.

[60] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X. Li, "Speech sanitizer: Speech content desensitization and voice anonymization," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[61] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 29–48, Citeseer, 2003.

[62] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. P. Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2019.

[63] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the european conference on computer vision (ECCV)*, pp. 620–636, 2018.

[64] M. U. Kim, H. Lee, H. J. Yang, and M. S. Ryoo, "Privacy-preserving robot vision with anonymized faces by extreme low resolution," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 462–467, IEEE, 2019.

[65] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[66] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 606–624, 2018.

[67] J. Hamm, "Minimax filter: Learning to preserve privacy from inference attacks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4704–4734, 2017.

[68] N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1329–1332, IEEE, 2017.

[69] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[70] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[71] C. Hadnagy, *Social engineering: The art of human hacking*. John Wiley & Sons, 2010.

[72] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 581–590, 2006.

[73] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Advances in Visual Computing*, pp. 565–578, Springer International Publishing, 2019.

[74] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874, 2014.

[75] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[76] "Ucsd computer vision."

[77] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset," Apr. 1998. The images are provided at no cost for non-commercial scientific research only. If you agree to the conditions listed below, you may request access to download.

[78] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

[79] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio, "Learning with a wasserstein loss," *arXiv preprint arXiv:1506.05439*, 2015.

[80] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[83] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2007.

[84] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. Karure, R. Raju, B. Rajan, *et al.*, "Indian movie face database: a benchmark for face recognition under wide variations," in *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pp. 1–5, IEEE, 2013.

[85] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[86] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.