CHALMERS
UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Image Quality Assessment Pipeline and Semi-Automated Annotation method for Synthetic Data

Validation and Annotation Techniques for Performance Enhancement of a Classification Model

Master's thesis in Computer Science and Engineering

LIAM LE TRAN and EDINA DEDOVIC

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Image Quality Assessment Pipeline and Semi-Automated Annotation method for Synthetic Data

Validation and Annotation Techniques for Performance Enhancement of a Classification Model

EDINA DEDOVIC AND LIAM LE TRAN

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Image Quality Assessment Pipeline and Semi-Automated Annotation method for Synthetic data

Validation and Annotation Techniques for Performance Enhancement of a Classification Model

Supervisor: Yinan Yu, Department of Computer Science and Engineering
Examiner: Krasimir Angelov, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Gothenburg, Sweden 2023

Image Quality Assessment Pipeline and Semi-Automated Annotation method for Synthetic Data
Validation and Annotation Techniques for Performance Enhancement of a Classification Model
Edina Dedovic and Liam Le Tran
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Predicting human emotions through facial expression, particularly in relation to medication field such as clinical trial settings, has garnered scientific interest in recent years due to significant understanding of the impact of treatment on emotions and social functioning. This thesis aims to improve performance of a FER model using large scale of synthetic data. A FER classification neutral network's performance is validated to accurately detect Action Units (AUs) in human facial images. To select the high-quality images among a pool of synthetic data, a Training Data Selection (TDS) pipeline is utilized, incorporating both no-reference and reference Image Quality Assessment (IQA) metrics. Furthermore, this thesis contributes through the development of a semi-automated annotation method, which offers an efficient approach to leverage an minimal amount of human annotation for labeling of a large number of images depicting various AUs. The proposed methodology incorporates seed tracking embedded in image names as a means to annotate the images. By integrating this annotation method with synthetic data generation, it minimizes the need for labor-intensive manual efforts and enables streamlined synthetic data annotation. Increasing the number of synthetic images to over 40,000, the model's classification performance shows moderate improvement. Namely, the enhanced FER model performance outperforms or show the same result compared to the baseline result for the majority of the classes. This outcome highlights the efficacy of utilizing the TDS pipeline using IQA in conjunction with the semi-automated annotation method in improving the overall performance of the classification model. The model achieves a range of ROC AUCs that vary between 0.80 and 0.92 over six AUs for cross validation.

These findings shed light on the challenges and limitations associated with using synthetic data for FER models. The findings also emphasize the need for further research to enhance the accuracy and reliability of synthetic data in this domain and the need for more accurate annotation of data with minimal interventions of human annotators.

Keywords: Facial Expression Recognition, FACS, Action Units, styleGAN2-ada, synthetic data, Image Quality Assessment, Multi-stage Pre-training, Pipeline Processing, Semi-automated Human Annotation.

Acknowledgements

First and foremost, we would like to express our thanks to our supervisor Yinan Yu for her support, guidance, and encouragement throughout this project. Her insights, feedback, and valuable suggestions were instrumental in shaping our work and we are grateful for her mentorship.

We would also like to extend my thanks to our examiner, Krasimir Angelov, who provided us with valuable feedback and insights during the final processes.

Furthermore, we would like to thank our friends for their support, specially in proof reading. Their contributions have been instrumental in shaping this report.

Edina Dedovic and Liam Le Tran, Gothenburg, June 2023

Contents

List of Figures

List of Tables

List of Algorithms

1
Introduction

The following sections provides a background and context for the research conducted in this master thesis. Problem statement, research questions, and objectives of the research are presented in this chapter. The introduction will briefly discuss the significant concepts in the related studies, some of the research methodology employed, and the structure of the thesis. The goal of thesis research is to investigate the potential of automated emotion detection based on facial expressions detection. The focus also lies on bettering the its applicability in clinical trials based on the existing related works. Lastly, the introduction section aims to provide a clear and concise overview of the research conducted in this study, which will set the stage for the chapters that follow.

1.1 Background

Emotions are vital to human communication and facial expressions are naturally used as signals to convey emotions and intentions. Consequently, the prediction of human emotions has gained increased scientific interest in recent decades due to the growing number of industrial applications and research that suggests a link between one's emotions and well-being.

Through continuous research and innovative solutions, the pharmaceutical industry is constantly evolving and always strives towards advancing its capabilities in addressing and resolving various health-related issues and challenges. One problem with medications is that they can have unpleasant flavors or textures that make them difficult for patients to take. In particular, this relationship is crucial in clinical settings where the ultimate objective of any treatment is to enhance patient's health. Therefore, understanding of emotional responses to a given medication or treatment is a significant question within the medical sector.

Facial expressions play a significant role in non-verbal communication, which can be used to identify the emotional state of humans. They account for 55 percents of non-verbal communication, which is a significant proportion, along with other non-verbal cues such as body language, tone of voice, and gestures [1]. Studies suggest a connection between a clinical trial subject's facial expressions and their self-reported quality of life. Thus, automated facial expression recognition (FER) tools hold potential in improving our understanding of how emotions and social functioning are impacted by treatment. This master's thesis delves into the domain of automated emotion detection, utilizing publicly available facial expression datasets, and explores the potential for introducing such methods in clinical trials.

1.2 Related Works

The development of accurate FER tools that can be applied to diverse populations/etcnicities has been limited by the scarcity of human facial datasets with comprehensive labeling. In recent years, deep learning (DL) has been widely applied in the medical field due to its ability to improve healthcare services. Medical and pharmaceutical researchers have developed DL models for supervised and unsupervised algorithms that can be used in healthcare radiology and clinical trials. Hence, an exploration of studies employing state-of-the-art techniques can help to gain a comprehensive outlook on the current research topic.

The following subsections will provide a summary of two research studies that are closely connected to the present thesis work. These summaries will assist in providing the underlying motivations for the current master's thesis project and introduce fundamental concepts relating to automated FER system using neural network. However, this section only briefly introduces relevant concepts. Detailed definitions of these concepts will be provided in Chapter 2.

1.2.1 Image-based Automated Emotion Prediction Research with Scarce Data

FER for Clinical Trial Self-recordings is a thesis paper authored by [2]. The paper suggests that there is potential for using a FER neural network in a classification task by utilizing Action Units (AUs) and a limited dataset [2]. A classification model for emotion detection was developed to classify facial expressions using the Facial Action Coding System (FACS). The FACS system separates facial muscle movements into AUs and mutiple AUs can be aggregated into emotions such as sadness and happiness.

The primary objective of the previous project was to gain a better understanding of emotional responses to medication or treatment, specifically in the context of clinical trials for treating conditions such as depression. In such a scenario, a video recording of the subject will be obtained during which they would articulate their emotional state, either in a general sense or in relation to a particular topic of interest. These AUs were aggregated into emotions such as happiness or sadness, which could be used to evaluate the effectiveness of a treatment. Figure 1.2 displays six AUs that were of interest in the predecessor project as well as this project. According to Von Numer, the mentioned AUs can be aggregated into emotions such as sadness and happiness and these emotions convey state of condition such as depression.

Figure 1.1 shows a subset of frames extracted from the FACS AU-encoded Denver Intensity of Spontaneous Facial Action Database (DISFA) dataset, which was the chosen dataset for training and validation of the final model. This set consists of

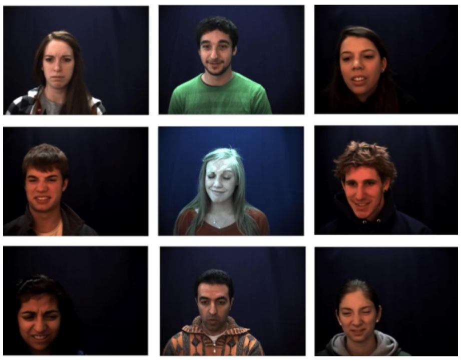


Figure 1.1: *A subsample of frames in the DISFA dataset*

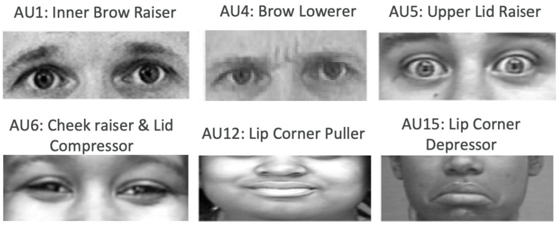


Figure 1.2: *6 AUs of interest in Von Numer's project*

a relatively small number of samples, namely 27 subjects, leading to a restricted distribution of data. All images in this dataset were captured with near-frontal view and under similar lighting conditions. Additionally, the recordings were made with a constant dark blue background.

According to the thesis, achieving accurate predictions of FACS encodings is challenging due to the limited number of subjects and the resulting sparsity of data distribution. Therefore, successful FACS encoding prediction strategies may require the use of appropriate pre-training of feature extractors, efficient data augmentation, and synthetic data generation techniques.

Although the deployment of a multi-layers pre-trained neural network in this study showed moderate success for the classification task, the author noted that limited training data, particularly the imbalance of AUs in the DISFA dataset, could have hindered the model's performance. In terms of training dataset, the DISFA dataset included a limited number of images from a small subset of individuals and exhibited a certain degree of imbalance with regards to the representation of AUs.

Given the challenges highlighted earlier, a promising avenue for future research was suggested by the author, which involved utilizing Generative Adversarial Networks (GANs) to generate synthetic face images that could supplement the existing training data. Furthermore, the author proposed the possibility to expand the work for automating predictions of six relevant AUs mentioned in the paper.

1.2.2 FER using Deep Neural Networks - A successor project

FER using deep neural networks, [3], is an extended project inspired by the projected in the previous section, 1.2.1. As a successor, the project aimed to explore the potential of supplementing the real dataset DISFA with synthetic data generated by generative models during the training stage. The aims of the successor project was to enhance the baseline model's predicting performance by using synthetic data together with real data. The authors suggest that synthetic data generated by Generative Adversarial Networks (GANs) such as styleGAN2-ada hold potential to enhance classification performance for of the existing model in [2].

In order to generate synthetic images that express specific AUs, this study incorporated a method that involves exploring the latent space of a GAN model, namely styleGAN2-ADA. The concept of the latent space is denoted as z in styleGAN. It serves as an initial space that can be sampled from a uniform or normal distribution. Furthermore, a mapping network is employed in styleGAN to enhance the control over the latent space z. This network consists of eight fully-connected layers and is responsible for transforming the original latent space, z, into an intermediate latent space known as w. The resulting w can be seen as a modified version of z, represented as z. Manipulation of the involved parameters w and z were done in order to control the direction of image generation. Specifically, this approach seeks to generate synthetic images that reflect the desired AU expressions. To streamline data generation process and facilitate subsequent analysis, the use of pre-trained models containing images of 256x256 pixels resolution was explored.

K-fold Cross-validation was a method used for validation on both the curated DISFA dataset and generated synthetic dataset, Eigenfaces. The DISFA dataset is split into subsets of thirteen folds, with two subjects per validation dataset for each fold. After each split, the Eigenfaces dataset is appended to the remaining subsets of the DISFA dataset, and each of these curated sets is used as the training dataset for each fold training. This results in a total of 15,000 images in the training set for each fold. This procedure is performed on all folds to train and evaluate the facial expression recognition system.

The results of this project was closely similar to that of the original model in term of classification performance. In this paper its acknowledged that the approach pursued in this study shows promise and has the potential for advancements with additional time and computational resources.

In light of the project's findings, the authors recommend several avenues for further investigation. Firstly, they suggest exploring the center-cropping technique for synthetic images, which could potentially enhance the model's performance. Additionally, the authors propose engaging experts in the field of annotation for facial AUs to refine the annotation process. Moreover, they advocate for exploring alternative methods to navigate the latent space, which can offer opportunities for enhanced model performance, as well as exploring more in depth the different validation techniques for quality of a generation model using Image Quality Assessment (IQA). Lastly, the authors highlight the potential of investigating an inductive semi-supervised learning approach, particularly in scenarios where a small labeled dataset is available alongside a larger unlabeled dataset.

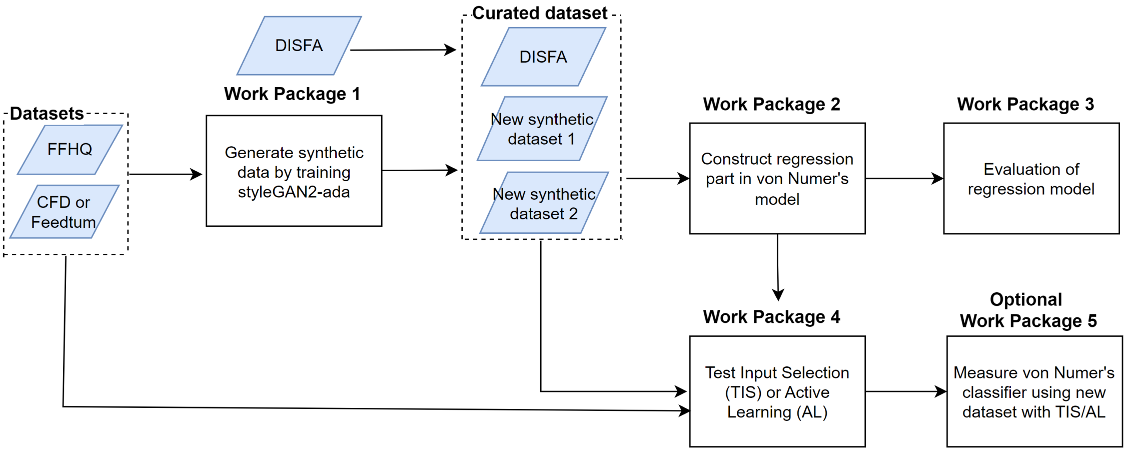
1.3 Problem Definition

Upon reviewing of the aforementioned research studies, it has become clear that the successor project has shown improvement potential for the original project's limitations by using generated synthetic data. The successor project successfully utilized the classification model from the baseline project to implement two benchmarks. By comparing the results of these benchmarks, the project gained valuable insights into the effects of Semi-Automated annotation, Training Data Selection and Latent Space Exploration. These approaches proved instrumental in addressing the challenges posed by the thesis and significantly contributed to enhancing the accuracy of AU predictions. It is worth mentioning that that the limited amount of data as well as the data imbalance are significant issues that requires solutions, but that they have been solved in this thesis by focusing on balancing both the quantity of the data but also the quality of the data. In this regard, the successor project has introduced potential methods such as using Training Data Selection, annotation in a more robust way and Latent Space Exploration to tackle the scarcity and data imbalance, which can potentially be a direct link to why the baseline model was not able to outperform the successor results.

Additionally, it is noteworthy that while the successor project is a stepping stone in improving the data scarcity and imbalance, it lacks a comprehensive, structured pipeline to support and facilitate the data flow. Hence, by incorporating supplementary constituent elements, it is feasible to create a more cohesive system, which could potentially boost the performance of the baseline model. Accordingly, a proposed approach can be, for instance to investigate how the Training Data Selection (TDS) of the input data influences the performance of the baseline model. A thorough examination would provide valuable insights into how a comprehensive pipeline can impact a neural network's performance positively or negatively. As the predecessor projects have made moderate progress in implementing the mentioned components, the current study's primary contribution is to develop a comprehensive structure of the pipeline that integrates these components effectively.

1.4 Project Aim

The following figure presents a high-level flowchart of various main work packages in this initial stage of the project. The blue parallelogram symbols indicate datasets, and the DISFA dataset will be utilized in the same manner as it was in the original project by von Numer.


Datasets
FFHQ
CFD or Feedtum
DISFA
Work Package 1
Generate synthetic data by training styleGAN2-ada
Curated dataset
DISFA
New synthetic dataset 1
New synthetic dataset 2
Work Package 2
Construct regression part in von Numer's model
Work Package 3
Evaluation of regression model
Work Package 4
Test Input Selection (TIS) or Active Learning (AL)
Optional Work Package 5
Measure von Numer's classifier using new dataset with TIS/AL

Given the various components involved in this project, it is crucial to establish clear objectives for each stage, as depicted in the flowchart, this way it ensures the coherence of the project's end result. Based on the mentioned aims in the previous section and the presented work packages, the preliminary objectives of this thesis are articulated below.

- **Obtaining a curated dataset using generated synthetic images**. The process involves the selection of appropriate datasets, the training of the selected generative model, and the generation of synthetic data that can be utilized as input for subsequent stages/objectives of the project. It is important to note that the chosen datasets in the initial stage should have the potential to support the other objectives of the project as well.

- **Obtaining functional neural network for the classification task of predicting AUs in the curated dataset**. The primary goal here is to use the curated synthetic dataset of images into the framework of the baseline model, resulting in a model that effectively captures the complexity of the feature maps and provides a representations of the presence of each AU.

- **Obtaining the performance/evaluation of the classification model**. This objective seeks to address the effectiveness of utilizing classification models for predicting the AUs using curated dataset of images via a selection methods such as Image Quality Assessment. The evaluation will be conducted using the curated dataset from the previous stage of the project as the input and evaluation metrics as the assessment tool.

- **Obtaining an understanding of the influence of Training Data Selection on the performance of the established classification model**. At this initial stage, the aim for this objective is to gain a comprehensive understanding of the impact of the use of TDS on the performance of the established classification model. This aims to address, on a high-level, the effect of modifying the input of the classification model through TDS.

The proposed solution is a simple pipeline in the TDS which can sort out the defects of all the synthetic images to the highest image quality degree. In other words this pipeline should work as a systematic approach to finding images that do not correspond to the need and criteria of the type of pictures that are desired. Instead this pipeline should implement sub-image processes that can easily detect faults within the images, for example image color, image noise et cetera, in order to discard of those images with the lowest reported image quality.

In order to construct the pipeline one needs to set requirements of what these sub-processes within the pipeline should do. Thus, upon doing systematic analysis of the images, it is found that the best results are given when one is provided with a formulation. The list below provides readers with a couple of requirements:

- Image should contain a complete face - By using landmarks the process of detecting face can be done very easily. If a face does not appear we simply discard this image and do not use the defect as an input to the classification model.
- Images should not contain any facial features that are not proportional or appear to be unrealistic to the human eye.
- Images should not appear to be in any color other than the original images; i.e color should only be appear in RGB-standard not grayscaled or any other spectrum.

One commonly used metric for evaluating synthetic face images is the Fréchet Inception Distance (FID), which measures the similarity between the distribution of feature vectors for real and synthetic images. The lower the FID score, the higher the quality of the generated images [4].

Notice that image alignment was not present in the aforementioned list. Generated image of a face should not usually appear to be tilted, or in any way aligned, such that it is not aligned perfectly to the image format. I.e people should appear to be somewhat aligned in center of image. However, to make a robust model we will need to focus on implementing a system that is not affected by the changes in an image, such as the alignment.

The main objective of this thesis is to compare the baseline, which was run without considering image quality, with the curated dataset of training data samples from this project. The aim is to assess the effectiveness of the suggested pipeline implementation by comparing the end results of the two benchmarks that will be conducted. The first benchmark will involve the baseline benchmarks reported in the original project, which utilized a dataset of approximately 40,000 images without considering their image quality. The second benchmark will use the same classification model but with the curated set of training data selection samples that take into account image quality. To concretize these goals, three research questions are formulated as follows:

Question 1 – *How to effectively evaluate synthetic image quality and opti-*

> *mize model performance using Training Data Selection (TDS)?*
>
> Question 2 – *How does integrating TDS pipeline, based on IQA metrics, impact emotion prediction using synthetic images? Does TDS improve prediction accuracy?*
>
> Question 3 – *What is the effectiveness of integrating more synthetic images in the baseline model for predicting AUs?*

1.5 Challenges and Limitations

Every research project has its own set of challenges and limitations that must be acknowledged and addressed. This section will discuss the challenges and limitations that are likely to be encountered during the course of this thesis research.

1.5.1 Challenges

To achieve the objectives mentioned in previous section, the following challenges should be taken into account. Such challenges will be delineated from the project by setting clear limitations, in order to keep track of the planed timeline. This will also shed light on how complex such machine learning task can be, especially because this project involves many different methods combined into one complete pipeline.

- The absence of images of human faces exhibiting negative emotions or those displaying emotions such as sadness, fear or disgust in the FFHQ dataset may present a challenge in the generation of images depicting such emotions.

- The prior study [3] has indicated the potential of the Eigen vector solution, which has resulted in partial reliance on this result for the investigation of the latent space in this thesis. This approach offers considerable benefits and insights. However, it also carries the potential risk of time-intensive analyses and technical difficulties in demonstrating the optimality of the Eigen vector solution. Therefore, alternative solutions will also be subject to examination in light of this consideration.

1.5.2 Limitations

In regards to the aforementioned challenges, it is urgent to establish clear and specific limitations for the scope of this master's thesis as articulated below. This will ensure that the project remains within its intended parameters and remains on track throughout the process.

- The generated synthetic images may possess a diminished quality when compared to their real counterparts, potentially leading to an adverse impact on the performance of the models. Although various strategies for optimizing the latent vector will be analyzed, the time constraint will be taken into account. Thus the goal is achieving reasonably satisfactory, rather than overly realistic, synthetic images. The assessment regarding the quality of the of the synthetic images will include both objective and subjective evaluations, including visual inspection.

- Over-fitting can occur as a result of excessive model complexity arising from an excessive number of parameters in the training process. Nonetheless, the present project will not focus on this matter and thus will not take actions to reduce the over-fitting.

- Given that this project builds upon existing elements and models of two preceding projects, the related works and theory sections will cover fundamental theories that are essential to the understanding of the project. However, due to limited time and some elements being irrelevant to the present project, detailed steps of the previous projects will be omitted.

- Furthermore, since the primary objective with this project is to implement a comprehensive pipeline and improving the data scarcity and class imbalance, prioritizing any individual component will not occur. In instances where an element in the pipeline is deemed to have potential for improving the final performance but time constraints do not allow for its complete implementation, it will be either suggested as future work or describe shortly as an alternative.

- The results of the preceding projects have been deemed successful, and therefore, any enhancements to the baseline model utilizing newly generated synthetic data in this project should be viewed as a supplementary accomplishment. Considering the presence of multiple constituents in this project and the construction of a pipeline being a contribution in itself, if both individual components and the entire pipeline do not produce an improvement in performance, we will not exceed the pre-determined time frame allocated for each phase exclusively for the purpose of enhancing performance.

1.6 Thesis Outline

This section provides a comprehensive outline of this thesis report. Given the multifaceted nature of the topics covered in this study, the thesis outline serves as a guide to facilitate the reading and comprehension of the intricate subject matter across different domains.

Firstly, Chapter 2 provides an overview of the FER domain, covering the theory behind styleGAN2-ADA, Latent Space exploration, Image Quality Assessment, and Classification models. It also includes a section on the evaluation method, relevant research, new directions, and challenges.

Chapter 3 presents the methodology of the thesis, starting with the data acquisition process. It discusses the statistics of the used datasets, preprocessing steps, and the methodology for feature extraction. It also describes the evaluation of good and poor quality pictures, along with the algorithms used.

Note that some of the sections mentioned in Chapters 2 and 3 will explore the technical details and offer a comprehensive explanation of the algorithms utilized in this research. These sections may contain a significant amount of technical content, as they delve into the intricacies of the algorithms. However, this level of detail is crucial for readers with an interest in the technical aspects of computer science and related disciplines. By discussing the algorithms and techniques employed, these sections will provide valuable insights into the underlying mechanisms and processes employed in generating synthetic facial images using GANs.

Chapter 4 focuses on the results, beginning with the outcomes of the feature extraction. It then presents the results from different phases in the Test Input Selection pipeline. A comparison is made between two benchmarks, one using all generated images and the other using only the synthetic images with the best reported image quality score. The chapter concludes with the testing of the classification model.

Chapter 5 draws conclusions, discussing potential weaknesses in implementation, future work, remaining challenges, and recommendations. It also provides an assessment of the performance that can still be improved.

2

Theory

The following Chapters 2 provides an overview of the FER domain with a fundamental understanding of the research, techniques, and challenges involved in FER, as well as the current state-of-the-art FER methods. Furthermore, techniques to conduct the pipeline for the entirety of the project is also introduced, i.e ll the processes that build up the complete pipeline will be introduced.. The insights gained from this chapter will serve as a foundation for the subsequent analysis and evaluation of the FER model.

2.1 Overview of FER Domain

FER is a field that involves recognizing human emotions from facial expressions. This task requires accurate feature extraction and analysis from static images or videos [5]. FER has many applications in various fields, such as psychology, marketing, and robotics. Nevertheless, the domain of FER is a complex and rapidly evolving field that remains a multifaceted domain characterized by constant evolution and inherent challenges. These challenges can have significant impacts on the development and progress of the domain. In the following subsection, a comprehensive explanation of the challenges that the FER domain commonly faces will be presented in detail.

2.1.1 Challenges in FER

One of the significant challenges in this domain is the lack of large datasets [5] [6]. Data scarcity makes it difficult to train and test FER algorithms effectively. Additionally, another prominent challenge in FER is the high variability of facial expressions. People express emotions differently, which makes it difficult to generalize FER models across different subjects. Moreover, the same person can express the same emotion differently at different times. Thus, FER models should be robust enough to account for the high variability of facial expressions.

2.1.2 Techniques used in FER

To overcome the challenges and enhance the performance of FER, researchers employ different techniques, such as convolutional neural networks (CNNs) and domain adaptation. CNNs have been widely used in FER research due to their ability to extract features automatically from images and videos. CNNs can learn a hierarchical representation of facial features that can discriminate between different emotions effectively.

Domain adaptation is another technique used in FER to overcome the problem of data scarcity. Domain adaptation aims to transfer the knowledge learned from a source domain to a target domain with different distribution. In FER, domain adaptation can be used to transfer the knowledge learned from a large dataset to a small dataset to improve the performance of FER models.

2.2 Facial Action Coding System (FACS) and Action Units (AUs)

One gold-standard tool that is widely used in psychology and other fields for research on emotion and nonverbal behavior is the Facial Action Coding System (FACS). It is a system for objectively measuring and describing facial expressions. FACS decomposes facial expressions into individual muscle movements, known as Action Units (AUs). Developed by psychologist Paul Ekman and his colleagues in the 1970s, FACS is based on the observation that facial expressions are composed of a set of discrete muscle movements in the face. These muscle movements can be observed and measured, allowing researchers to objectively describe and analyze facial expressions [7].

In behaviour science, the Facial Action Coding System (FACS) has become the gold-standard used in behavioral science to decode and study facial expressions. It was originally developed by Ekman and Friesen in the 1970s and has become a widely used tool to analyze facial expressions and this is by decomposing facial expressions into muscle movements in the face for analysis . FACS utilizes AUs to further break down facial movements, whereby a single or combination of AUs represent a facial expression or emotion, see figure below.

FACS consists of 46 different AUs that correspond to individual movements of the face, which are objectively measured by trained annotators. In this project, the key interest is in generating a subset of AUs classified in the original project, including Inner brow raiser (AU1), Brow lowerer (AU4), Upper lid raiser (AU5), Cheek raiser and Lid tightener (AU6 and AU7), Lip corner puller (AU12), and Lip corner depressor (AU15), as shown in Figure 2.1. The subset of AUs is a relatively objective method to measure facial movements, which is critical for examining facial expressions and emotions. Figure 2.1 displays a selection of the 46 AUs initially specified by Ekman and Friesen.

2.3 Generative Adversarial Networks (GANs)

Vanilla GANs, short for "Generative Adversarial Networks," were initially proposed by Goodfellow et al [8]. Their name "vanilla" simply refers to the basic or funda-

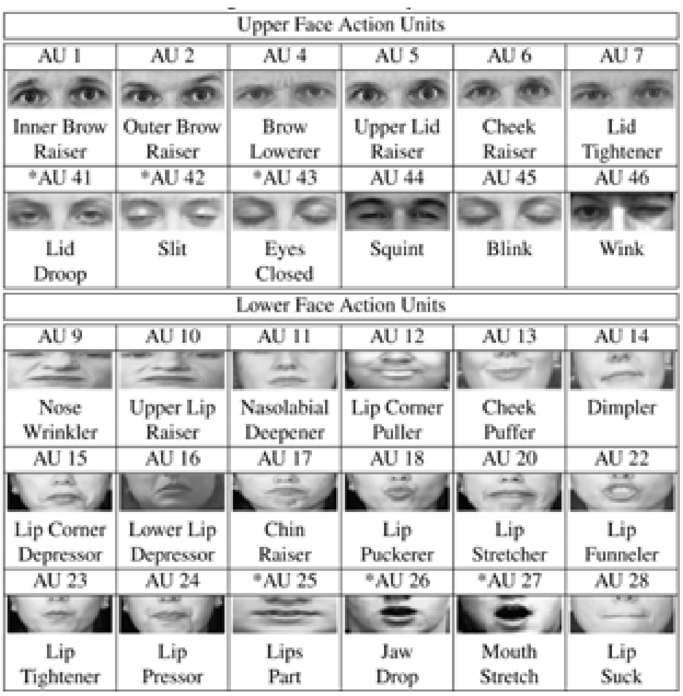


Figure 2.1: *A subset of FACS AUs with descriptive text*

mental version of GANs. While Vanilla GANs are commonly associated with image generation, there are also variations of GANs that can generate diverse forms of data such as text, video, and audio.

They are a type of generative models that can produce new data instances through an adversarial process. The basic architecture of the system consists of two models, a discriminator (D) and a generator (G). This process involves two models, G is a generative model and D is a discriminative model, that work against each other in a mini-max game with an objective function that can be formulated as follows.

$$min_G max_D V(D,G) = E_{x \sim p_d ata(x)}[logD(x)] + E_{z \sim p_z(z)}[log(1 - D(G(z))] \qquad (2.1)$$

Both of the models start out as untrained and their output will be improved over the training process. While not having access to the real training data, G aims to generate instances that resemble the real training samples and fool D. This makes possible by G receiving feedback from D on how realistic its synthetic samples are. Meanwhile, D attempts to correctly classify whether the instances are real or synthetic, given both real training samples and fake instances produced by G. After an instance is classified by D, the discriminative model undergoes improvement through reviewing of actual "real" or "fake" labels. Thus it learns by means of back-propagation in a manner similar to that of a standard classifier [8].

In the realm of deep learning, loss function is a method used to evaluate how well an algorithm models a dataset. The measurement yielded with a loss function is the difference between the predicted output and the actual output. Thus, the objective is to minimize the value of the loss function. A decreasing result of a loss function implies that the model performance is improved. Various types of loss functions can be employed depended on the type of problem being solved. [9]. GANs employ a Binary Cross-Entropy (BCE) loss function [8]. In general, BCE loss can formulated as follows.

$$L(y, \hat{y}) = -(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y})) \tag{2.2}$$

The training procedure for G is to maximize the probability of D making a mistake by classifying generated images as the true training data. G captures the data distribution as a vector of random noise and generates instances that resemble the target output, i.e., the input data. D learns to distinguish the true training data from the output of G by estimating the probability that the generated instances come from the training data or from G. Thus, D is used as a tool to train G. Furthermore, vanilla GANs do not require class labels [8].

As D improves its ability to distinguish synthetic images from real ones, G must likewise advance its ability to generate data instances that more closely resemble those in the training data, in order to continue deceiving D. Eventually, an equilibrium is reached in which further iterations do not increase the chances of success for either G or D, and the network subsequently converges [10]. Upon completion of training, the discriminative model may be discarded.

In comparison, Markov chains, which are applied to similar problems, represent a different approach. Markov chains are stochastic models that transition between states based on probabilistic rules. They are commonly used for modeling sequential data. In the context of generative models, Markov chains rely on the current state to determine the probability distribution of the next state. However, Markov chains often struggle to model sharp distributions accurately.

Thus, GANS offer several advantages. Firstly, GANs demonstrate computational efficiency since the generative component is updated solely based on the gradients provided by the discriminative model, eliminating the need for actual training examples. On the other hand, Markov chains typically require explicit training using observed data sequences.

Furthermore, GANs exhibit a high level of flexibility. A well-trained generator in GANs is capable of approximating a wide range of distributions across different domains. This means that GANs can effectively capture and generate diverse data patterns. In contrast, Markov chains often struggle to model sharp distributions, making them less effective in representing complex and intricate data structures accurately.

2.4 StyleGAN

The styleGAN network was proposed by researchers at NVIDIA, and is a novel approach for generating realistic artificial human faces. Its key feature is the adoption of a progressive growth mechanism, inspired by Progressive GAN [11]. It has several major improvements over other GANs, including fidelity, diversity, and feature control, achieved through a more involved architecture based on the idea of the Progressively Growing GAN (PGGAN) [12]. PGGAN trains the generator with a low resolution at 4x4 pixels in each block, gradually increasing the resolution for each block through nearest-neighbour upsampling until it reaches the desired resolution. During training, the upsampling is gradually replaced with transposed convolutions, resulting in a more stable training process and reduced training time. This approach ensures that higher-resolution transposed convolutions are only used once the model has enough capability to handle them.

The discriminator in styleGAN has a similar structure but in a reversed fashion. The z vector is fed through a feedforward network to obtain a learned noise vector w, making the input vector more directly connected to semantic features in the output instances. In styleGAN, "v" and "z" are important vectors in the generative process. The z represents the latent space, a random noise vector that influences the generator's output. The v is derived from z using a feedforward network, connecting the input vector to the semantic features of the generated instances. With v, the generator controls the style and appearance of the images. It is progressively introduced during transpose convolutions, allowing for multiple versions of v at different stages. This enables style mixing, blending different styles in the generated images. Note that the usage of v and z in styleGAN may differ from general GANs, but both vectors impact the generation process.

This procedure affects the feedforward network producing w in backpropagation operations [12]. The first layer feeds a constant value into the generator, and w is gradually fed into the network at different stages of the transpose convolutions. This enables the creation of multiple versions of the vector w and allows for style mixing possibility in images [12].

2.4.1 StyleGAN2

StyleGAN2, an improved version of styleGAN, removes residual effects on the generated images by modifying the instance normalization. It also makes interpolation among different w vectors smoother, further advancing feature control. styleGAN2 also improves upon the progressive growing by simultaneously training all resolution blocks from the start, while focusing more on low-resolution blocks in the beginning and gradually shifting to later blocks as the model improves [13]. This approach maintains the progressive feature inherited from PGGAN while training all blocks simultaneously.

2.4.2 Latent Space in styleGAN2-ADA

StyleGAN2-ADA is a modification of the original styleGAN2 model that incorporates adaptive discriminators to improve its performance. This model has been shown to generate highly realistic images, and the quality of the generated images is highly dependent on the quality of the latent space representation [14].

Latent space is a fundamental concept in generative deep learning models, particularly in the recent state-of-the-art StyleGAN2-ADA model. Latent space is a high-dimensional space in which the generative model G learns to represent the essential features of the data. In the general styleGAN architecture, G selects latent vectors from an intermediate space called w using eight fully-connected layers and applies various transformations to encode style, such as normalizing feature maps and injecting noise. Thereafter, the synthesized images will undergo an inversion process to map them back into the latent space. The reason for the intermediate latent space is to enforce disentanglement of the feature mappings and to enable fine-grained control over image synthesis in styleGAN. In StyleGAN2-ADA, the latent space is a 512-dimensional vector space that encodes the key features of an image [14]. This vector is then transformed into a feature map that is fed into the synthesis network to generate an image.

Visualizations and manipulations of the latent space provide insights into how different regions correspond to specific visual features in generated images. One can explore and navigate this space to discover meaningful directions and make targeted modifications, enhancing interpretability and control of the generative model. This capability allows for creative exploration and fine-tuning of image synthesis.

2.4.2.1 Latent Space Exploration with Synthetic CFD

In this project, the model was trained on dataset CFD in order to learn the structural similarities between images. The model is able to classify the images in a way it learns the features and structures from the images in a classifier fashion. At a frist galnce, the generation of synthetic images may appear to be entirely random. However, this procedure is not completely random since it is inherently a latent process. The term "latent" refers to its hidden nature and lack of immediate visibility.

The concept of latent space is important because it’s a utility connected to deep learning. In other words, the latent space explorations is the task of leaning the features of data and simplifying data representations or the purpose of finding patterns.

In order to understand latent space one needs to understand why data compression is one of the processes in the latent space exploration. Data compression is a common process used in machine learning in order to compress data points into a more human readable space, for example 2D or 3D space. Data compression, is the process of encoding information using fewer bits that the original representation. For example, taking a multiple dimensional like 10D (10 values needed to define a unique output) and then squishing all into for example lesser dimensions, 2D data point. This practice of compressing data is largely used in the field of machine learning. It involves acquiring crucial details concerning data points, which facilitates comprehension and manipulation of input and output, as well as various machine learning models. Article [15] discusses unsupervised compression algorithms applied to gene expression data and how they extract latent or hidden signals representing technical and biological sources of variation. Another article [16], presents a method that allows for efficient compression of data by structuring the latent space in a way that makes it easier to control the reconstruction process.

Typically, when referring to latent space in spatial terms, it refers a space consisting of 4D or n-dimensional points. But latent space is usually more complex to imagine. Thus, a space that is higher than n>3 is nothing we can demonstrate in this paper. Nonetheless, utilizing techniques such as t-SNE (t-distributed stochastic neighbor embedding), it is possible to convert the latent space representations from a higher dimension into a more comprehensive and visualizable form, such as 2D or 3D. The generator model in this paper takes data points from the latent space as input and generates a new image. The latent space itself is a abstract representation, but it hyper sphere represents a large amount of untested ground that can be explored by generating and traversing the space.

The processing in this paper takes into regard the generation of images but via a latent vector 'z' from a Gaussian distribution, then mapping 'z' to the w space using the mapping network and finally producing the image from the w space using a synthesis network. This results in tensor of images that have the dimension of (n_images, 3, image_size, image_size) where the images are the number of images and 'image_size' is the size if the images (256 in this case). The size of the tensor depends on the number of input of images that the functions takes.

Images are saved as tensor because they can easily be processed and manipulated using mathematical operations. Tensors are multi-dimensional arrays that can represent complex data structures, such as the images we are saving. In this case, dimensions for width, height and colour channels. So by saving images as tensors, they can be resized, normalized, transformed, and manipulate for various machine learning tasks without a lot of programming changes or re-works. Another advantage of saving images as tensor is the amount of space one saves by effectively compressing and storing into smaller file sizes without losing important information.

Coming back to t-SNE, the reasons of using this reduction technique is that it is a reduction technique design which enables the preservation of the local structure of the data. Practically this means that data points that are close together in the original high-dimensional space will be mapped to nearby points in the lower-dimensional space, and thus t-SNE is particularly interesting for identifying data points in clusters or k-groups of similar data points.

To visualise the latent space using t-SNE the data first needs to be encoded into the latent space using deep learning network. Once you have defined the latent vectors, the t-SNE is applied to reduce the dimensionality of the vectors and visualise them in 2D or 3D space. Therefore, using this technique one can visualise the CFD dataset and see if there are any apparent patterns and relationships in the data that may have been hard to test or analyse in the original high-dimensional space. It is crucial to note that, while t-SNE is an tool that provides insights into the underlying structure, it is a non-linear technique and may not preserve all of the information from the previous high-dimensional space. Upon implementing this technique on CFD one can try to analyse the latent space and explore it before digging in directions and the process of picking seeds. Below shows how the CFD's latent vectors are scattered in a 2D space; see Figure 2.2.

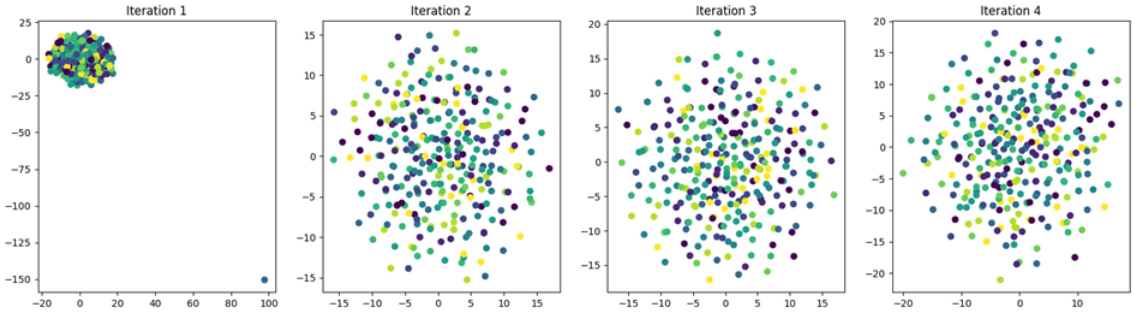


Figure 2.2: t-SNE method for the latent space of the CFD model.

The plot generated shows the t-SNE embeddings for each iteration and the Mean Squared Error (MSE) between the t-SNE embeddings for each iteration. The plots visualised are 2D representation of the high-dimensional data. Each point in the scatter plot represents an instance in the dataset and the colour of the point represents its class label. By comparing the scatter plots for each iteration, you can observe how the t-SNE algorithms is grouping the instances of similar classes together.

Through visualization, one can observe clusters, groupings, and separations of the samples, enabling the identification of similarities or dissimilarities between them. This process potentially uncovers hidden patterns or structures within the data. By visually analyzing the results, it is apparent that the first iteration is clustered in the upper left corner, while the subsequent iterations display very similar feature distributions. The correlation coefficients between the t-SNE embeddings of each iteration offer insights into the stability and consistency of the results. Higher correlation coefficients indicate similar mappings, thus reinforcing the reliability of the visualization, this can be observed in the iterations 2 to 4. Consequently, it is evident that the latent space is not entirely randomized. By mapping the feature samples onto the latent space, one can also understand how to navigate it and generate desired images with corresponding facial features and characteristics.

The MSE heatmap, on the other hand, shows the pairwise mean squared error between t-SNE embeddings for each iteration. The darker the colour, the higher the MSE between the embeddings. Heatmaps are used to determine if the t-SNE algorithm is converging to a stable solution or if encountering stagnation in local optima. Hence, a low ME between t-SNE embeddings indicates that the algorithm is converging to a stable solution. The MSE heatmap is shown in the figure below:

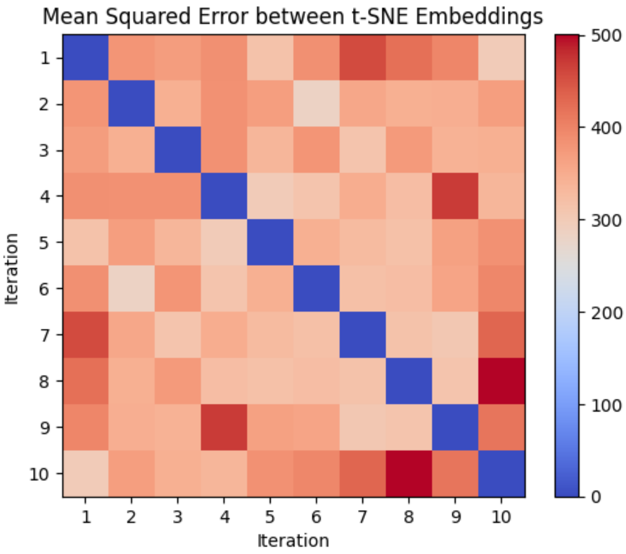


Figure 2.3: MSE heatmap method for the latent space embeddings model.

2.4.2.2 Latent space exploration with interpolation

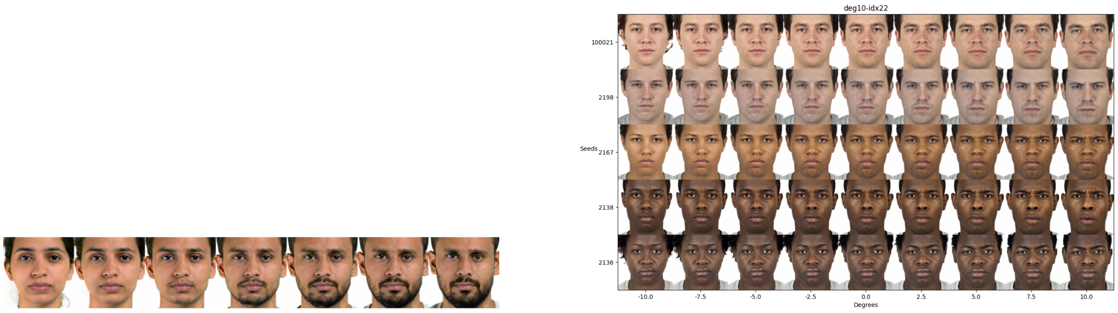
Interpolation is a mathematical technique used to estimate data points within a set of known values. In this case interpolation will be done by exploring the latent space within the border of analysing synthetic images. In the context of exploring the latent space of generative models, interpolation is used to generate intermediate samples between two given points in the latent space. It will allow for a more clear picture and allow one to study the properties of the model's internal representation. By using repetitive interpolation, we can observe how the model generates images as we move along a path in the latent space. The most common way to interpolate between two images in the latent space is to linearly interpolate their corresponding latent vectors, which are their numerical representations. Suppose there are two images, A and B, with corresponding vectors z_A and z_B. To interpolate between z_A and z_B, one can sample a set of intermediate latent vectors z_i by linearly interpolating between z_A and z_B using the formula:

$$z_i = (1 - \alpha) \cdot z_A + \alpha \cdot z_B \tag{2.3}$$

where α_i is a scalar that controls the degree of interpolation, with α_i=0 representing image A and α_i=1 representing image B. The scalar αi controls the degree of interpolation, allowing for the generation of intermediate latent vectors and corresponding images. By varying the values of α_i in a ranges from 0 to 1, one can generate a sequence of images that smoothly transition from A to B.

Recent state-of-the-art for exploring latent space of generative models often use interpolation as tool for studying the properties of the model's internal representations. For example, GANSpace in [17] uses interpolation to explore the semantic attributes of generative models. Other examples of interpolation uses Autoencoders to explore latent space [18]. Autoencoders shows the abilities to embed data manifold into low-dimensional latent space, making the data construct very usable in learning methods of representation and space. Additionally, there are more advance ways of constructing interpolation that can be used to generate more realistic and diverse images. For example, Riemannian Walk for Incremental Learning [19] which introduces a path-based interpolation method that follows a curved path in the latent space. This results in more natural-looking transitions between different images.

As previously mentioned, several methods exist for interpolation; however, the diversity that each new technique brings makes it challenging to obtain a complete image or overview. Instead, one needs o focus on generating and being able to analyse more directions, degrees, and seeds in several interpolation in single iterations. An individual interpolation can be represented as shown in Figure 2.4 (a). However, when comparing a grid of multiple iterations of interpolations, to single interpolation makes the choice of latent space method more evident, i.e choosing a grid of multiple iterations. This facilitates the exploration of the vector space and the identification of the appropriate vectors indexes (idx) to achieve desired changes. These changes could include neutral or intense facial expressions, morphing between male and female, or alterations in various ethnicities along a single vector direction. Refer to the figure below for clarification, see Figure 2.4.



(a) Single iteration of a interpolation of two images (morphing).

(b) Multiple iterations of interpolation, repeating the process of morphing.

Figure 2.4: Side by side comparison of the two different techniques, single iterations (a) and multiple grid like iterations (b). The grid allows to compare multiple interpolated images side by side. This allows for quick visual recognition when it comes to any consistent changes or transitions in the images. Which, may provide insights into the underlying data distribution or relationships.

Notice that the interpolation has a degree axis depending on the vector number of choice. The methods of exploring the feature space of a pre-trained GAN model are done by visualizing what images could be generated depending on the vectors directions. The vectors represent the directions in the feature space in which changes in the input vector will produce a specific change in the output image. To be more specific, the vectors represent the principal components of the weight matrix that maps the input vector to the output image. By using the vectors, one can perform operations such as interpolation between images and thus controlling specific features of the generated images. The degree of change in the output image is determined by the magnitude of the change in the input vector along the vector direction.

2.5 Stability Issues of styleGAN2-ADA Training

styleGAN2-ADA can create high-quality synthetic images by using adaptive discriminator augmentation (ADA) to improve the training stability and performance [20]. However, if the training time is too long, the synthetic images may become extreme or unrealistic. This could be due to overfitting, mode collapse, or insufficient data augmentation [20][21].

2.5.1 Overfitting and Underfitting

Overfitting is a well-known obstacle in machine learning, which often has to do with a model's tendency to learn the noise instead of the salient features. This phenomenon arises when a model excessively learns the training data, leading to a high level of accuracy in classifying the training dataset. However, such a model may fail to generalize well to novel or unseen data, resulting in suboptimal performance and inaccurate predictions.

The primary cause of overfitting can be two primary reasons, one is the presence of an excessive number of parameters. The other is an overly complex model, which leads to the capture of noise or random fluctuations in the training data. In order to address this issue, several techniques have been proposed, including regularization, cross-validation, and early stopping. These techniques are often used to prevent overfitting. The idea behind is to constrain the model's capacity, ensuring that it generalizes well to unseen data, and thus improves its overall performance.
In order to solve overfitting techniques such as increasing the number of parameters, introducing additional features, or adopting more advanced algorithms can help prevent underfitting. By incorporating these strategies, a model can more effectively capture the underlying patterns in the data, resulting in improved performance.

Low regularization rate and learning rate yields overfit data, high regularization rate yields partially fit data, and high learning rate yields unfit data [22]. In terms of machine learning as a domain, the regularization rate and learning rate are key hyperparameters that influence the performance of a model. In general, a low regularization rate and learning rate can lead to overfit data. The opposite is true for the regularization rate, where a higher rate tends to result in partially fit data. Additionally, high learning rates can cause the model to underfit, resulting in a failure to accurately capture the underlying patterns in the data.

For a even more detailed explanation, the regularization rate refers to the level of penalty imposed on the model for complex functions. Here, a lower regularization rate results in a decreased penalty. This can lead to a higher degree of overfitting, in which the model becomes too specific to the training data. Consequently, it results in poor generalization to new, unseen data. In opposite, a high regularization rate can result in underfitting, where the model is too generalized and unable to effectively capture the patterns in the data.

Similarly, the learning rate refers to the step size taken by the model in updating its parameters during training. A high learning rate can result in the model overshooting optimal parameters and failing to converge effectively. This can lead to poor performance on both the training data and new, unseen data. In contrast, a low learning rate can result in slower convergence, but can also lead to better generalization to new data.

2.5.2 Mode collapse and Ending Training

The problem of vanishing gradients is a significant challenge when training GANs. It causes the network to stop learning and is often accompanied by mode collapse, which is when G is only capable of producing one type of image. Mode collapse is a result of G being stuck in generating a single optimal solution. The reason for mode collapse is due to a weakness in D network, which causes G to produce one type of image repeatedly. This limits the generator's ability to spread to other modes and thus remains at the one it had generated already [23].

The conventional approach for implementing D involves employing the sigmoid activation function and Binary Cross Entropy (BCE) loss function. However, the task assigned to D is often deemed less challenging than the task assigned to G. Specifically, assessing whether a painting is a genuine piece or a replica is generally easier than generating a replica that closely resembles an original artwork. Consequently, D may learn rapidly, causing the BCE loss to converge towards either zero or one. This leads to increasingly small gradients being passed on to G, slowing down learning and, in the worst case, causing it to stop completely [24].

2.6 Transfer Learning

Transfer learning is a ML technique that use knowledge gained from a previously learned tasks and aims to enhance the performance of a target task by leveraging those knowledge acquired from previously learned tasks, capitalizing on the underlying similarities between them. The technique is particularly useful in situations where there is limited labeled data available for the task at hand [25].

Applications of transfer learning spans various domains in ML, including computer vision, natural language processing, and speech recognition. Among these mentioned domains, the most common field that utilizes a lot of transfer learning is computer vision. In essence, pre-trained models can used to improve the performance of other models. For instance, a pre-trained model trained on a large dataset of images can be employed as a foundational component to enhance the performance of a novel model designed for classifying a smaller dataset of images [25].

2.6.1 Challenges of Transfer Learning

The methods has several benefits, however, it does not come without challenges. One of the main challenges of transfer learning is that it can be difficult to find a suitable pre-trained model that is relevant to the task at hand [25]. In addition, the pre-trained model may not always be transferable to the new task, which can result in poor performance [25].

Another challenge of transfer learning is that it can be difficult to fine-tune the pre-trained model to a new task. Fine-tuning would mean that one has to be involved with adjusting the pre-trained model to better fit the new task. This can be challenging due to the complex nature of machine learning models [25].

2.6.2 Transfer Learning of GANs

GANs training are known to be difficult. During the training process, G and D employ and maintain several parameters through forward-propagation. Additionally, any errors that arise are back-propagated through all the layers of the two models [26]. As a result, these models are computationally demanding and time-consuming to train from scratch. Using pre-trained GAN models would be ideal to facilitate the training process while still obtaining results on smaller training datasets. Moreover, this also introduces challenges as one needs to find suitable pre-trained model for their right task.

To address this issue, the research paper by [27] suggests that freezing the lower layers of the discriminator could assist a pre-trained GANs network in outperforming previous methods despite its simplicity when retrained on various datasets [27]. Additionally, the authors argue that their baseline model, FreezeD, yielded the best performance when fine-tuned for both the generator and discriminator [27]. However, the styleGAN2-ADA network has a highly complex architecture, and the input dataset is typically in tfrecords file format for the network to be able to use it. Tfrecord format is used for storing a sequence of binary records [28].

2.7 Using Pickle files to maintain and resume training stages

The use of Pickle files has become a popular approach utilized in training deep neural networks. By using Pickle files, trained models can be saved and restored with ease, and the need for re-training from scratch can be eliminated [29]. It serves as the preservation of trained models, thereby minimizing the need for lengthy retraining. Pickling is a technique to serialize Python objects such as complex deep learning models into a binary format that can be stored and reloaded at a later time. This is especially useful when working with large datasets or computationally expensive models, where the re-training process can be time-consuming and resource-intensive. Consequently, this enables training of models to be started and stopped more flexibly, thus saving time and computational resources. This also provides the additional benefit by making trained models portable and shareable across platforms and users [29]. Furthermore, henceforth, the utilization of pickle files for the purpose of resuming or conducting training will be denoted as a pre-trained or trained model.

2.8 Training data Selection

Training Data Selection is a machine learning technique used to choose a set or subset of representative inputs for evaluating the performance of a model. The process involves selecting training data that meets specific requirements, such as selecting synthetic images that exhibit the highest image quality based on Image Quality Assessment metrics. The goal of the training data selection is to identify that subset or subsets of inputs that are diverse enough to thoroughly test the model's capabilities and expose any weaknesses or limitations. The inputs should have high coverage of any specified requirements, and the requirements need to be specific enough to reflect the processes within the Training Data Selection. Considering that, on might think, what should then be interrelated and tested in the TDS process.

In deep learning methods such as Training Data Selection in this thesis and especially GANs, are usually evaluated on generated images. The point is to reduce the overall overfitting and increase generalization since training selection is quite complex in many ways. Even though it is challenging, one can, through a set of training selection metrics for deep learning systems, make training selection practical.

The problem with testing in deep learning networks, for example DNN, is that it is very costly to validate the correctness of a model's predictions. Which largely affects the efficiency of the model testing and also affects the whole process of development. To relieve this problem of labeling test inputs to check correctness, the authors propose a novel test input prioritization approach. Thus, it facilitates improving the efficiency of model's testing and consequently selecting images that are the most relevant to the requirements described in previous chapter, Chapter 1. Similar approach is used in the domain of Quality Assurance and Assurance where testers first want to prioritize tests that are more bug-revealing and cover crucial parts of code instead of having tests that would not cover new areas (code coverage) or that do not cover any new specific requirements. Other articles, [30], discuss diverse test input generators (TIGs) that have been proposed to produce artificial inputs that expose issues of deep learning systems by triggering misbehaviors.

Regarding the present thesis pipeline, the research focus will shift from the perspective of Quality Assurance and Testing. Instead, the pipeline shall concentrate predominantly on the implementation of metrics. These metrics will enable the quantification of image quality and are related to feature extraction analysis and other relevant metric analyses. Prior to the presentation of the proposed pipeline, a systematic exploration of the literature in this field is conducted. Note that at the initial phase of this exploration systematic search was done for research that involved generating training data selections, methods like IQA were systematically reviewed. In the initial phase, the identified papers were skimmed down to extract relevant information. However, searching for papers with similar approach as the proposed TDS Pipeline in this thesis did not yield any results. Therefore, the proposed pipeline below is a novel method based on the conventional TDS methods used in deep learning. It is a hybrid solution that incorporates selection methods in computer science such as algorithms, via Image Quality Assessment (IQA) metrics to convey how the selection of synthetic images is chosen.

In other words, the proposed pipeline suggests that by implementing IQA as an indicator of good and bad quality, one can filter out the good quality synthetic images by analyzing the IQA quality number metric. This approach ensures that images that closely reflect the original CFD dataset's characteristics will be selected as input for the last layer output in the pipeline. Figure 2.5 is a visual representation of the proposed pipeline.

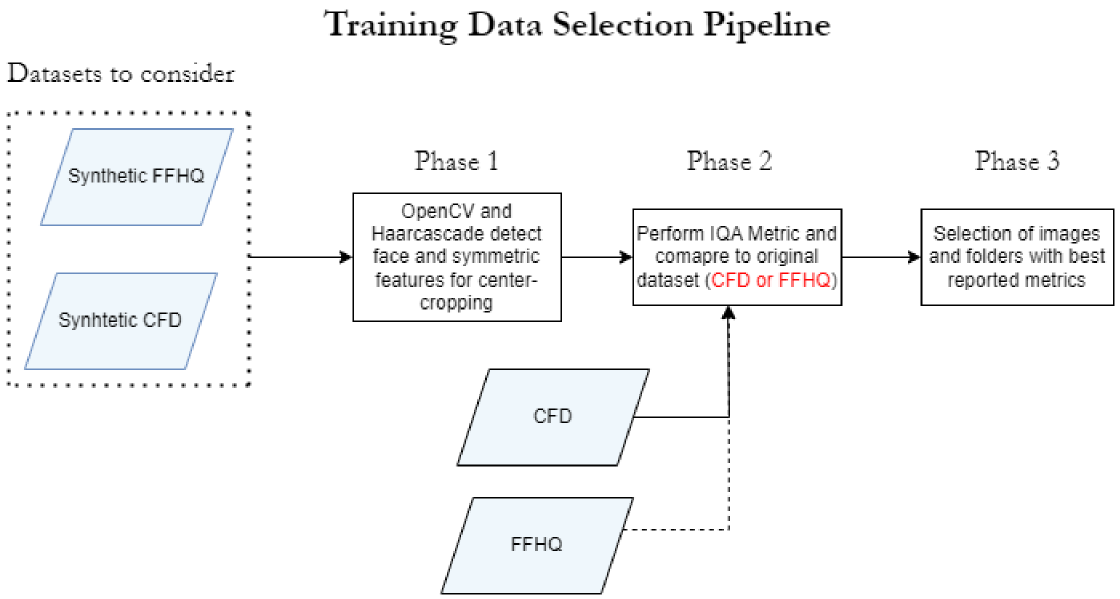

Training Data Selection Pipeline
Datasets to consider
Synthetic FFHQ
Synhtetic CFD
Phase 1
OpenCV and Haarcascade detect face and symmetric features for center-cropping
Phase 2
Perform IQA Metric and comapre to original dataset (CFD or FFHQ)
Phase 3
Selection of images and folders with best reported metrics
CFD
FFHQ

Figure 2.5: The pipeline consist of 3 phases where each steps includes a careful analysis of the data it holds within that phase. Phase 1 includes selection of datasets to consider. Phase 2 includes pre-processing images with OpenCV library. Phase 3 performs IQA analysis on the synthetic sets and compares them to their original datasets.

Figure 2.5 displays the pipeline proposed in this project and comprises four distinct phases, each of which include a careful examination of the data involved within that particular phase. In the first phase, various methods of dataset selection are evaluated (evaluation of distribution and evalution of how each datsets fits into data related requirements). The second phase involves preprocessing of the images utilizing the OpenCV library. In the third phase, IQA analysis is performed on each of the synthetic set, and the results are compared with its corresponding original dataset.

2.9 Data Pipeline

Test image quality assessment can be incorporated as a step in the Training Data Selection (TDS) pipeline by using objective metrics to quantify the quality of generated images. These metrics can be used to rank the generated images based on their perceived quality, and the highest-ranking images can be selected for further processing or testing.

In this thesis, the approach involves integrating image quality assessment into the TDS pipeline. It utilizes Image Quality Assessment (IQA) techniques to compare and measure the perceived quality of the generated images against the reference images.

2.9.1 Image Quality Assessment (IQA)

Image Quality Assessment (IQA) is a process of evaluating the visual quality of images. Typically this is measured by analysing the degree of distortion or degradation that occurs in image capture, compression, transmission, resizing or rendering. IQA is an important area in the research of computer vision and image processing. It is a applicable process in a wide range of fields such as photography, multimedia, medical imaging and surveillance.

IQA has also been a topic of significant research interest in the field of computer vision and image processing. As imaging technology continues to advance, the demand for high-quality images has increased. Super resolution has been part of the imaging processing as well as other techniques. However, IQA has become a essential tool in the comparison of different datasets and allows for a objective analysis. Tools like this offer the performance of various image processing algorithms, such as compression, denoising, and super-resolution.

There are several approaches to IQA but the most common ones are Full-Reference (FR) IQA and No-Reference (NR) and Reduced-reference (RR) IQA [31]. A Full-reference IQA compares the quality of the distorted image with the quality of the original image (reference image) using a metric such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), or structural similarity (SSIM). The Full-reference IQA is a technique that is widely used and an example of this is the Siamese network-based IQA. It is a state-of-the-art technique and conveys image quality comparison [32]. Siamese networks are a type of neural network architecture that can learn similarity between two inputs [33]. In this particular case, the two inputs are two images that are being compared. Additionally, the neural network architecture allows for a faster comparison of two images since it learns the to compare them using a loss function. The loss function measures the difference between the predicted similarity score and the true similarity score. The network then adjusts its parameters to minimize the loss. By minimizing the loss, the network learns to output the similarity scores that are the closes or closer to the true similarity scores.

In term of popularity, Full-Reference is the most popular approaches in the common techniques of IQA. However, NR-IQA has gained increased attention in recent years. No-Reference IQA can evaluate image without a reference image. NR-IQA does assess images against the equivalent synthetic or the generated image. It assesses images based on statistical or perceptual models of quality such as blind image quality assessment (BIQA) or natural scene statistics (NSS). NR-IQA is often more complex than FR-IQA but is also more often the most practical choice in real-world scenarios where a reference image may not be available, as in many cases of GAN related image generations.

Recent works include novel model that addresses the NR-IQA task but by leveraging a hybrid approach that combines a transformer-based feature extractor with a multi-layer perceptron [34]. Other works include the use of GAN to predict the primary content of a distorted image and then measures different degradations simultaneously with a multi-stream convolutional neural network (CNN) for NR-IQA [35]. The proposed solution is called Active Interference of GAN for No-Reference Image Quality Assessment. It approaches the NR-IQA in a hybrid approach where it combines GAN with active interference to predict primary content of a distorted image.

Lastly, the Reduced-Reference (RR) IQA is an approach that is gaining popularity as well. RR-IQA is neither FR or NR-IQA. It's a compromise between them and uses partial information from the reference image to estimate image quality. The techniques rely on the characteristic information about the pixels, coefficients of certain transformation and/or other predominant features of the original image. There is recent work which concentrates on creating new ways of assessing videos and image in RR-fashion. In [36] the authors propose a novel deep learning-based method. It presents a classification of RR methods and discusses their advantages and limitations. It emphasizes in the need if objective methods of quality assessment as subjective assessment is time-consuming and expensive and usually not applicable in real-time scenarios.

2.9.2 Evaluation of the IQA techniques

All of these techniques have been used in great extent to assess images in multiple of ways and using both very popular model such as the Siamese model but there are uses with novel models that require more customization. In order to define the Image Quality Assessment best suited for the current model in this paper one needs to consider the relationships between the input images that have been used to training the model (both the CFD and the FFHQ) and what their equivalent if there are any equivalent, in the synthetic images pool. As stated before there are different techniques depending on image relationship in two of the pools. However, considering that the generated pool has little to nothing to do with the original dataset one needs to rethink and approach the comparison in a different way. If a synthetic dataset is generated using a pre-trained model, then it can be considered a Reduce-Reference (RR) IQA problem. The reference images for the RR-IQA task are the original images that the pre-trained model was trained on, and the disorted images as the generated images that were created using the same pre-trained model.

To assess the visual quality of the generated images, the RR-IQA algorithm will use partial information from the reference images (original set of pictures) to estimate the image quality. This can be achieved by extracting features from both the reference images and the distorted (generated) images. Lastly compare them to measure the degree of distortion or degradation. The feature extraction from the reference images will provide some context for the type of images that the pre-trained model was designed to generate. Therefore, the RR-IQA algorithm can evaluate how well the pre-trained model is able to generate images that are similar to the original images it was trained on.

It is important to consider that the effectiveness of the RR-IQA approach is dependent on the pre-trained model's ability to generate high-quality images. If the model fails to generate high-quality images, the RR-IQA algorithm may struggle to accurately assess their visual quality. Furthermore, if there are significant differences between the generated and reference images, the RR-IQA algorithm may face challenges in accurately evaluating their visual quality. While RR-IQA provides a more reliable method to test quality differences in feature-extraction-based solutions, it does have limitations that can restrict the assessment process.

The NR-IQA approach provides a more suitable method to assess the quality of generated images. It evaluates the statistical and perceptual properties of the generated images, comparing them to reference images. The algorithm considers features like blur, noise, contrast, and sharpness to determine image quality. One advantage of this approach is that it doesn't require reference images. Additionally, since the original data is consistent in terms of environment and settings, with only variations in objects, the task of finding similarities focuses on a specific region of pixels. However, the NR-IQA approach has limitations as it relies on statistical or perceptual models of image quality, which may not precisely reflect human perception of image quality.

Moreover, the generated images are very similar in terms of the generated image's background (environment) but also the properties they possess (ethnicity, ages and expressions). Here, the choice between the RR-IQA and NR-IQA has made it simple to choose. From now on and forward in this paper, there will be more discussion on the appropriate techniques using the NR-IQA approach.

2.9.3 Fréchet Inception Distance (FID)

Another way of processing and analysing different dataset, in order for a full objective comparison is with the Fréchet Inception Distance (FID) score. The Inception score is one of the first GAN-evaluation methods to become widely adopted [37]. It is based on the Inception-v3 model that is trained on the ImageNet dataset [38].

To calculate the Inception score, two probability distributions are calculated: p(y|x) and p(y). The first distribution measures the fidelity of the generated images to real images in the dataset, by calculating the probability of a certain class given a certain input. The second distribution measures the diversity of the generated images, by calculating the probability of each class being represented in the generated images. A high Inception score is achieved by a generator that produces high-quality images that are diverse and well-distributed across different classes [37].

In other words, the purpose of the Inception score is to give a high score to a generator that manages to both create high quality (high fidelity in the literature) and high diversity images. High fidelity is measured by calculating p(y|x), i.e. how probable a certain class is given a certain input. Given x, we want to have as high probability for y as possible.

The main objective is to achieve a precise distribution, aiming for a high certainty level of 100% for a majority of the images. This implies that the generated images should be accurate and correctly classified. Furthermore, we calculate p(y) to assess the diversity of the images and ensure that the generator produces a variety of image types, rather than just one. The goal is to evaluate the evenness of class distribution among the generated images [39]. Here we want a distribution that is as flat as possible, since we want to have a diverse output. Our ideal here is thus a uniform distribution. The Inception Score is then given by comparing these two probability distributions using the Kullback–Leibler divergence.

Using the Inception score [40] poses several challenges. First, it compares synthetic images exclusively and does not include real images in the evaluation, focusing solely on probabilities derived from the Inception classifier. Second, its applicability is limited to the ImageNet dataset and struggles when assessing generated images from different domains. Lastly, the score can give a perfect diversity rating even with just one instance of each class, as it measures probability rather than absolute numbers. However, it tends to produce better scores with larger sample sizes, indicating a bias towards large datasets [40].

The FID has proved to be a good way of measuring both fidelity and diversity, but it still has drawbacks. One is that, just like the Inception score, the embeddings it uses are trained on ImageNet. This makes it quite dependent on the pictures the Inception model is trained on. Another, is that we expect humans, and not feature extractors, to be looking at the generated images in their final use-case. So it is quite hard to get around the fact that humans still need to be involved in the evaluation process if we want the best possible benchmark. Therefore, there is still a need for frameworks that use human annotators. One such framework is HYPE (Human Eye Perceptual Evaluation), which relies on crowdsourcing human evaluators from platforms such as Mechanical Turk [41].

Comparing the performance of different generative adversarial network (GAN) models can be challenging due to the absence of an explicit objective function. While human annotation has been commonly employed to assess visual quality, it introduces potential variability and bias. In [39], the use of automated processes like Amazon Mechanical Turk (MTurk) for human annotation is explored. However, this approach still faces challenges, as assessment outcomes can vary based on task setup and annotator feedback. Incorporating feedback can lead to a more critical evaluation, highlighting flaws in generated images. To address these limitations, automatic methods like the Inception model and Fréchet Inception Distance (FID) are used for more objective and automated evaluation of generated images in the context of GANs [42]. Note that in this project the NPYViewer tool [43], is used in order to error check and validate the annotation process. For more details read section A.3 in the Appendix.

To measure the quality of the generated images some definitions need to be made. The idea is that there is some real data distribution (2.4) and a generating model data distribution (2.5).

$$p_r(.) \rightarrow \text{real world data} \tag{2.4}$$

$$p_g(.) \rightarrow \text{generating model data} \tag{2.5}$$

Ideally, one wants to know if the two distribution of real data and generated data are equal following the domain of partial differential equations and finite elements methods, such that:

$$p_r(.) = p_g(.) \text{ iff } \int p_r(x)f(x)dx = \int p_g(x)f(x)dx \tag{2.6}$$

Here, you multiply the distributions by some test functions *f(x)*, then compute the integral. Now similarly in real life scenario, if the two integrals are equal for all of the *f(x)* that are spanning the feature space, then $p_r(.) = p_g(.)$ will be equal as well. Here *f(x)* is/are the basis function that span the feature space.

$$f(x) \rightarrow \text{basis spanning the function space in which } p_r(.) \text{ \& } p_g(.) \text{ live} \tag{2.7}$$

For example, one type of basis functions are polynomials. However, one needs to limit themselves since you cannot compute for all polynomials, so let's say that one looks for polynomials of *zero* and *1*. That gives you the first and second moment of the distribution. One is the expected value and the other one is the second moment.

$$f(x) \rightarrow \text{ polynomials, first \& second moments, Gaussian} \tag{2.8}$$

Here we are looking at *f(x)* and $f(x)^2$. One results in the first moment and the other one give the second moment. Additionally the only distribution that one knows the entire properties, by knowing only the mean and the variance, is a Gaussian distribution. There by the equation above states that one needs to work with the Gaussian distribution.

Next, we take the images (real or fake) and probed them through the neural network that is pre-trained (in this instance the Inception Model based on Fréchet). It results in two codes, for the images.

$$x \rightarrow \text{coding layer of an Inception Model} \tag{2.9}$$

Second step is to look at the statistics of those images. In other words, you now first featureize the images using the Inception model, you compute the mean of the generated images and compute the variance of the generated images. Then same goes for real images, you compute the mean of the real images and the variance of the real images. Lastly you compute the distance (Euclidean Distance) between the two Gaussians. This is how you compute the Fréchet Distance, between two Gaussians:

$$d^2((m_g, C_g, m_g, C_r)) = \|m_g - m_r\|_2^2 + Tr(C_g + C_r) - 2(C_g C_r)^{1/2} \tag{2.10}$$

The FID scores are reported in Chapter 4 of this paper. The chapter includes results from both the baseline comparison of real and fake images, as well as the evaluation of different unique batches. The setup of the training selection will be described in more detail following Chapter 4. Moreover, FID algorithm is included in the Chapter 3 as it is a relevant metric with valuable insight when it comes to discussing the differences between the real and generated images.

2.9.4 Support Vector Regression (SVR)

In most linear regression models, the objective is to maximize the sum of squared errors, as demonstrated by Ordinary Least Squares (OLS) [44]. OLS is utilized to identify the best-fit line that minimizes the sum of squared errors between the predicted and actual values. However, Support Vector Regression (SVR) follows a distinct approach for regression problems.

SVR is a popular regression model that surpasses linear regression in its ability to handle non-i.i.d. data and non-linear relationships. It enables the construction of non-linear models and the definition of acceptable error levels by minimizing coefficients and maximizing the margin. By identifying a hyperplane within a high-dimensional feature space, SVR can separate data into two classes and minimize the disparity between predicted and actual quality scores [45]. This approach ensures robust performance even when dealing with noisy and complex data.

The SVR algorithm transforms input features into a higher-dimensional space using a kernel function, which facilitates the definition of a hyperplane. The radial basis function (RBF) and polynomial kernel are commonly employed in SVR for Image Quality Assessment (IQA) [46][47]. The SVR model predicts the quality score of a new image based on its feature vector and the defined hyperplane.

SVR demonstrates inherent flexibility, allowing for a range of predicted values within a specified error range. In the context of Image Quality Assessment (IQA), SVR serves as a regression analysis algorithm that predicts continuous variables based on input data, such as generated images. Trained on these images, particularly their feature maps, SVR aims to predict the feature maps of unseen data. It finds applications in image quality assessment by training models to predict quality perception features associated with visual quality. SVR has been employed in various IQA schemes, including assessments for image retargeting and blurred images, enabling the prediction and comparison of quality between generated and original images.

The final implementation of SVR occurs on the selected set of test and training data. In other words, the subsets of test and training data that have been evaluated and reported as the highest quality in the preceding non-reference image quality assessment (NR-IQA) step serve as the input for SVR in the final stage of the pipeline.

In terms of restrictions, one can additionally change the restrictions of SVR to output only images that are within certain score. The specific SVR implementation will use BRSIQUE as the evaluation score. Results from the run is reported in the Chapter 4. Lastly, the idea is to split the SVR prediction training into two phases. The first phase focuses on training on the original data of the CFD. However, one should not solely train on all the data. Focus lies on validating the images, so it is essential to train on partial original dataset and validation on the rest of the original dataset. In this way prediction of image quality can be validated and supervised based on the validation score output.

Next, the last phase of the SVR is to analyse the synthetic dataset and validate whether the metric outputs any logical score in terms of IQA score. This approach is like the Fréchet Inception Distance algorithm which similarly looks to validate the synthetic datasets and its image quality scores based on the original data scores. Below, are the steps that will define the work process of implementing SVR. The calculation of image quality prediction using SVR involves several steps:

1. Feature extraction: Extract relevant features from the image that are thought to be correlated with human quality perception. Examples of such features include colour histograms, texture features, and structural information.
2. Training: Train an SVR model using a set of training images and corresponding human quality ratings. The SVR model learns to map the feature vectors to the corresponding quality ratings.
3. Testing: For each test image, extract the same set of features and use the SVR model to predict its quality rating.
4. Evaluation: Compare the predicted quality rating with the actual quality rating assigned by human observers using a metric such as mean squared error (MSE) or Pearson correlation coefficient.

2.10 Evaluation Metrics

The evaluation of model performance is a critical component in determining the suitability of a machine learning model for real-world decision-making and prediction tasks.

2.10.1 Receiver operating characteristic curve (ROC)

A Receiver Operating Characteristic (ROC) curve serves as a graphical representation that illustrates a binary classifier system's classification ability as the discrimination threshold is varied. The construction of the ROC curve consists of plotting the true positive rate (TPR) against the false positive rate (FPR) at diverse threshold settings. TPR corresponds to the proportion of actual positive cases accurately identified as positive by the classifier. Conversely, FPR referred to as the probability of false alarm, denotes the proportion of actual negative cases miss-classified as positive by the model

The ROC is a widely recognized and employed metric for evaluating the performance of binary classification models. By plotting the True Positive rate against the False Positive rate across a range of classification thresholds, the ROC curve provides a visual representation of the inherent trade-off between a model's sensitivity and specificity [48]. This trade-off is observed by adjusting the decision boundary of the model, in order to generate binary predictions denoted as "Yes" or "No" for each instance. A decision boundary of a model is determined by setting thresholds of the model's outputs such as probabilities. By adjusting the threshold, variations of how much conservative the model will be are allows, thereby influencing both the False Positive and True Positive rates. Thus, the ROC curve serves as an evaluation tool for the performance of a model across a wide range of decision boundaries. This offers insights into its capacity to correctly identify positive instances while minimizing false classifications [48].

2.10.2 Area Under the ROC Curve (AUC)

The Area Under the ROC Curve (AUC) is an commonly used metric for evaluating the performance of a binary classification model, serving as a measurement of the distinguishability degree between the two classes [48]. As a summary metric, it captures the overall performance of a classifier by quantifying the model's ability to discriminate between positive and negative samples, where the area under the curve always presents a value between 0 and 1. A higher ROC AUC value that is closer to 1 indicates better model performance in terms of separating positive and negative samples. In contrast, a random classifier is expected to yield an AUC value of 0.5. Therefore, the higher the AUC is, the better the classifier's ability to distinguish between positive and negative instances. Maximizing AUC is desirable since it corresponds to the highest True Positive rate and lowest False Positive rate, achieved at some threshold [48].

2.11 K-fold Cross-Validation

K-fold cross-validation (KCV) is a technique used to evaluate the performance of machine learning models on a dataset, according to sources [49] and [50]. The process involves dividing the dataset into k equally sized subsets or folds and training the model k times, each time using a different fold as the validation set, while the remaining k-1 folds are used as the training set. This procedure is repeated k times, with each of the k subsamples used exactly once as the validation set. To evaluate the model's performance, the results of the k training runs are averaged. It is important to note that the partitioning into k equal sized subsamples is randomized.

The choice of k depends on the size of the dataset and the desired level of accuracy. A higher value of k provides a more accurate estimate of the model's performance, but it also increases computational costs. Conversely, a lower value of k reduces computation costs but may result in a less accurate estimate of the model's performance.

Studies on k-fold cross-validation in machine learning environments are limited, however, some research has shown that KCV can be used to evaluate the performance of machine learning models on datasets, as discussed in a study by [51]. The study explores the relationship between the choice of k in KCV and its impact on the size of the dataset. The author suggests that the optimal value of k decreases as the dataset size increases. Specifically, for small datasets, k=10 is commonly used, while for larger datasets, k=5 or k=3 can be used without sacrificing too much accuracy. The study found that a value of k=10 provided a good balance between computational cost and accuracy.

In another study by [52], the authors demonstrate that the choice of k in KCV can significantly impact the accuracy of performance estimates, and there is no one-size-fits-all approach. They suggest that the optimal value of k depends on the relative size of the training and test sets. For larger datasets, smaller values of k can be used to reduce the computational cost of cross-validation.

Finally, [53] provides recommendations for choosing the value of k in k-fold cross-validation. They suggest using k=5 or k=10 for small to medium-sized datasets and smaller values of k for larger datasets.

3

Methods

The methods section provides a detailed account of the techniques and procedures used in the research project. It includes a description of the study design, data collection and analysis methods, and any other relevant experimental techniques employed in the investigation. The aim of this section is to provide a comprehensive explanation of the methods used to collect and analyze data, ensuring that the reader has a clear understanding of the research process. In this master thesis, the methods section will be presented with a focus on the data collection process, data preprocessing techniques, and the machine learning models used to analyze the data.

3.1 Outline for the Thesis's Roadmap and Processes

This section includes some explanation on how the each of the process in the methodology looks like and it provides the user with a full overview of each of the steps take in this project. It will serves as a pin pointer in later chapters to easily redirect readers to specific areas and not introduce misunderstandings. For example, from data acquisition and dataset research to final output of the Training Data Selection (TDS) for the classification for this project to looks like. The roadmap in Appendix A.1, includes the following presented steps.

1. Data Acquisition: Selection of datasets to be used.
2. Distribution Analysis: Performing static analysis on gender, age, and male/female subjects within the main dataset (Chicago Faces Dataset - CFD).
3. Pre-processing: Face detection and cropping.
4. Curation: Manual annotation of the original CFD dataset.
5. GAN Model Training: Training a GAN model for generating synthetic images based on the original dataset.
6. Latent Space Exploration: Exploring the process of image generation using latent space.
7. Seed List Generation: Creating a list of seed values by leveraging the knowledge acquired through exploration of the latent space for the purpose of generation.
8. Synthetic Image Generation: Generating approximately 100,000 synthetic images based on the seed list and using combinations of index and degree.
9. TDS Pipeline: Utilizing the newly generated images in the TDS pipeline.
10. Classification Model Input: Using the output of the TDS pipeline (approximately 40,000 images) as input for the classification model.

11. Classification Model Training: Training the classification model on original images and validating it on a curated set of synthetic images.
12. Benchmark Results: Comparing the results of two benchmarks: preliminary results (DISFA + randomly chosen synthetic images without IQA-based selection) and the benchmark with synthetic images selected using the TDS pipeline.

In regard to the exploration realm, this research aims to make two key contributions regarding Development of Image Quality Assessment pipeline and a Semi-Automatic Annotation System that is integrated with synthetic data generation process. Upon generating the synthetic data using CFD, the name Synthetic CFD or SCFD is given the resulting dataset.

One aspect of this thesis research focuses on conducting experiments using Image Quality Assessment (IQA) techniques with different evaluation metrics and models. This approach aims to curate a dataset of the highest quality images, thereby minimizing the impact of lower quality data on model performance. By leveraging IQA techniques, one can identify and select images that exhibit the best image quality. Which ensures the reliability and accuracy of the synthetic dataset.

Knowing this, the IQA methods selected were chosen to such degree that they needed guarantee an accurate analysis of the synthetic images generated by the GAN model. Continuing on, a second goal was to do further analysis with IQA in order to differentiate the CFD and SCFD. Finally understanding to what degree they were dissimilar. Running indicated that a significant proportion of synthetic images generated from original SCFD exhibited minimal deviations from the original CFD dataset.

Given these findings, there is no need to prioritize one requirement over the other, as the generation process offers the advantage of producing larger datasets while maintaining promising image quality. Motivated by this finding, a deliberate decision was made to selectively choose only 180 out of the available 512 images in each batch of data. This deliberate selection aimed to build upon the insights gained from the aforementioned metric analysis.

3.2 Data Request Process and Outcomes

The process of requesting data involves multiple stages to facilitate the accuracy and relevance of the acquired datasets. Firstly, it is necessary to establish the data requirements. Data requirements have the purpose of guiding a team of developers, companies et cetera to established a firm ground of the involvement of data within a certain project. In other words, data requirement is need in order companies to track of how closely they are in acquiring data but within the requirement specification space. In this thesis, the requirements are there in order to guide this project forward and focus on attaining relevant data and understanding if that acquired data really is relevant to the project.

The second stage involves collating the data, which may necessitate accessing data from a range of sources. Then, validating and cleaning the data, and preparing it for analysis. It is also crucial for the people that handle the data to ensure that the data is secure and meets any applicable regulatory requirements.

The third stage involves analyzing the data and preparing it for use. It may also which be necessary to involve generating reports, visualizations, or other tools to help to understand whether that data is qualified and follows the requirements.

3.2.1 The Denver Intensity of Spontaneous Facial Action Database (DISFA)

The Denver Intensity of Spontaneous Facial Action Database (DISFA) is a publicly available dataset of spontaneous facial expressions, which was developed at the University of Colorado [54]. At time the original project was carried out, DISFA consisted of 26 subjects (comprising both males and females, with a range of ages and ethnicities) and contained 300 images per subject, resulting in a total of 7,800 images.

The videos capture facial expressions in a neutral state, as well as spontaneous expressions of six commonly occurring facial AUs, i.e. AU1, AU2, AU4, AU6, AU12, and AU15. The DISFA dataset is intended for researchers to study facial expression recognition, analysis and synthesis, as well as the physiological and psychological mechanisms underlying facial expressions. It has been widely used in the field of computer vision, facial expression recognition and affective computing, as well as in various applications in psychology, neuroscience, and social sciences [Disfa][54]. The author of the original project, von Numer, granted access to the DISFA dataset used in the present study.

3.2.2 The Chicago Faces Database (CFD)

The Chicago Face Dataset (CFD) is a collection of over 3,000 high-resolution images of human faces. They as well have been annotated with a range of attributes, including race, gender, age, and facial expression. The dataset was developed by researchers Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink at the University of Chicago. The CFD is a popular dataset in academic research on face perception, social psychology, and computer vision. The CFD provides high-resolution, standardized photographs of male and female faces of varying ethnicity between the ages of seventeen and sixty-five. Extensive norming data are available for each individual model.

To access the CFD, interested parties must submit a request form on the dataset's official website at [55], providing information about their research project and intended use of the dataset. Upon approval, users can download the dataset, which is provided as a set of image files that also come with annotation files. The CFD is intended for scientific research use only and is free for academic use, but users must sign an agreement outlining the terms and conditions of use. These terms include restrictions on redistribution, commercial use, and modification of the data. It also mentions requirement to acknowledge the source of the dataset in any resulting publications or presentations.

At the outset of the project, the CFD was identified as an appropriate dataset for use in this thesis research and access was granted. However, due to a lack of necessary information such as specific AUs present in the images, the provided annotations were not utilized in the research. Additionally, the CFD is a small dataset, containing approximately 1,200 images. Thus, making it more suitable for generating additional synthetic data. Through this process, the annotations become meaningless since the newly generated data's facial expressions depend on the parameters of the styleGAN2-ADA rather than the original annotations in the CFD. Consequently, annotation of this new data needs to be done.

3.3 Data Requirements

In the context of emotion prediction with limited access to datasets, striking a balance between data quality and label quality poses a common challenge. Data quality refers to the characteristics and properties of the dataset that contribute to its usefulness and reliability. In this project, data quality encompasses two main aspects: density and relevance.

In certain situations, accurate labels hold more significance than high-resolution data. Consider the example of emotion prediction, where having labels that precisely reflect the expressed emotions is crucial, even if the overall data quality is low. On the other hand, there are cases where high-resolution data takes precedence over accurate labels. For instance, when training a model to recognize emotions from facial expressions, it becomes more important to have high-resolution images rather than precise labels.

Thus, it is a common challenge to balance data quality and label quality in FER domain, specially when access to datasets is limited. To tackle this issue, it may not be possible to optimize both data quality and label quality. Therefore, one of them may need to be prioritized over the other. The choice of which to prioritize will depend on the specific problem and the available resources.

In situations where access to datasets is limited, combining datasets with different strengths can help create a more robust emotion prediction pipeline. By using a diverse range of datasets, it may be possible to address biases and achieve a more accurate and generalizable model. However, combining datasets may also introduce new challenges, such as ensuring compatibility between the datasets and addressing differences in labelling conventions.

Overall, the choice between prioritizing data quality or label quality and the use of combined datasets will depend on the specific problem at hand and the available resources. The following are the requirements to be considered when tackling this problem.

1. Should data quality or label quality be prioritized when dataset access is limited and one cannot simultaneously optimize for both?
2. Can datasets with different strengths be combined to create a robust emotion prediction pipeline?

It is worth noting that datasets characterized by both high-quality labels and high-resolution data are uncommon in practice.

3.3.1 Data Quality Definition

In this thesis, the term "high data distribution quality" is used to describe a dataset that has a high density and spans relevant dimensions. The relevance of the dimensions is domain-specific, and in the context of emotion prediction for clinical self-recordings, a dataset is considered to span relevant dimensions if it meets two requirements. Having large and varying dataset with different subjects is crucial to avoid bias and any discrimination that might occur towards certain groups; studies such as [56] and [57] provide deeper look int the biases present in facial expression recognition algorithms. They show that there is a cross-cultural difference in facial behaviour. Thus, having the two different dataset not only provides a large-scale dataset all in all but also a larger diversity of subjects and thus facial expressions.

The first (1) requirement is that the dataset must have a large demographic distribution, meaning that it should include subjects of different ages, genders, and ethnicities. It is also beneficial for the dataset to include common feature variations of the human face, such as beards, and make-up. The second (2) requirement is that the dataset should represent the human emotion range in a naturalistic fashion. This means that the emotion expression should occur spontaneously without too much involvement from an instructor. Unfortunately, many datasets are lab-recorded with varying elicitation methods, which can affect how natural the emotion expression is. Elicitation methods range from watching funny, sad, or scary video clips to performing expressions on demand. Overall, a high-quality dataset for emotion prediction in clinical self-recordings should have a high density and span relevant dimensions, including a large demographic distribution and naturalistic emotion expression.

3.3.2 Requirements

For this thesis project, a total of four datasets were requested and described in Table 3.1 below. However, only three of these datasets finally used since the fourth had restricted access. Among the accessed datasets, the DISFA dataset [58] was found to be the most suitable as it is FACS encoded, as well as the FFHQ and the CFD. These dataset are selected for deeper analysis due to its close resemblance to the trial self-recording setting in several ways. Such are, recordings and images taken are being fairly frontal and the elicitation not being instructed. Additionally, both CFD and DISFA are both unique were each subject either is instructed to express 5 different expression, and DISFA dataset subject has four minutes of video material at

20 frames per second. Both the datasets have the required demographic distribution, with subjects from diverse ethnicities, ages, and genders, although white young adults are in majority. For further details on the DISFA dataset, please refer to the Table A.1. For any background information regrading the DISFA dataset and previous project refer to the Chapter 1 and section 1.2.1. The dataset contains description of relevance between 1 and 5 for the 4 different datasets, their dataset size and how they have been accessed, as shown in Table 3.1.

Database	Num. of Images	Relevance (1-5)	Form of granted access
CFD	1,217	5 (critical relevance)	Via communication with authors
FFHQ	∼ 70,000	2 (low relevance)	Via public source
DISFA	7,800	4 (high relevance)	Via communication with authors
FEEDTUM	399	NaN	Via communication with authors

Table 3.1: The datasets considered in this project.

The table above shows fields *Database* and *Relevance*. Here the scale of relevance depends on the facial images within each dataset. The highest relevance, *Critical)*, was given to CFD since the outline for this thesis project is based on the usage of CFD within the model processes. FFHQ does not have the same relevance as it lacks expression that are required for this thesis. However it still needs to be incorporated in order to understand what the outcome of the results of the cross-validation and AU prediction model is. The third database, DISFA, is also needed in order to perform the process of AU prediction and cross-validation. The dataset is essential for this project as it enables comparison of results between von Numer's project and the current one. It is crucial for conducting fare comparisons and evaluating outcomes from both perspectives, making it a relevant component for analysis. Lastly, FEEDTUM [59] or Facial Expressions and Emotion Database is a collection of facial images that captures various individuals displaying the six primary emotions identified by Eckman & Friesen. The expressions range from Happiness, disgust, anger, fear, sadness, and surprise, all the way to neutral expression. The database was created as part of a project conducted at the Technical University Munich to facilitate research on the impacts of different facial expressions.

Image acquisition involved using a Sony XC-999P camera with an 8mm COSMICAR 1:1.4 television lens. Images were captured at a resolution of 640x480 pixels, with a color depth of 24 bits and a frame rate of 25 frames per second, using a BTTV 878 frame grabber card. Subsequently, the images were converted into 8-bit JPEG-compressed format with a reduced size of 320x240 pixels due to storage limitations.

The database comprises data obtained from 18 unique individuals, each performing the six target emotions three times, resulting in a total of 399 image sequences. The images are organized into subdirectories based on the type of emotion, and metadata is provided to indicate the start, apex, and hold frames of each sequence.

Access to the database is granted through a password-protected ZIP archive and a collection of MPEG compressed videos, which can be requested by contacting some of the researchers and completing a designated form attached to the email. Given these factors, it was ultimately determined that the utilization of FEEDTUM was unnecessary, as it was deemed to be a smaller-scale dataset in comparison to the already limited dataset of CFD. Moreover, challenges were encountered in comprehending the usage of FEEDTUM, as it relied on multiple file formats to generate images in a class-controllable manner.

3.4 Descriptive Statistics, Distribution Analysis and Data Visualization

3.4.1 CFD

The CFD dataset has been used along with an annotation CSV file, which contains general information such as age and ethnicities. While there are general emotion predictions included in the annotations, this study will not rely on them as they are based on predictions.

The subjects are divided into categories based on their etnicity (in CFD, CFD-MR and CFD-INDIA), and the ranging of age is between 16-65 years of age. Certain subjects are recorded with various facial expressions, while the most other subjects have only neutral facial expression. In CFD, the distribution of ethnicities was analyzed and the results are depicted in Figure 3.1. The total number of subjects in CFD is 827. The dataset consists of self-identified Asian, African-American, Caucasian, Latin and Indian descent who were recruited within the United States. All subjects are presented with neutral facial expressions, while a subset of the dataset is also inclusive of subjects with varying facial expressions. The dataset includes norming data for all neutral expression images, which were rated based on subjective rating norms derived from a sample of raters within the United States.

Due to the copyright policy entails with the acquisition of CFD, the content of the official dataset cannot be displayed in this thesis report, as the terms of use suggests that the content of CFD cannot be re-distributed. It is only allowed to be used for non-commercial research purpose only. To ensure that the thesis does not violate the copyright and thus acting on a copyright infringement, the thesis will not show any images related to the CFD original dataset, only the regenerated images. Thus, this report will only include images for already generated CFD subjects and will not display original CFD images. Figure 3.4 demonstrates that the ethnicities follow a roughly normal distribution. The groups of African-American and Caucasian descent are the most frequent, comprising 24 percents and 22 percents of the total number of subjects, respectively. Subsequently, individuals of Indian origin represent the third most frequent ethnic group with a proportion of 17 percent. Among the total number of subjects in the dataset, individuals of Asian and Latino and Mixed Race descent represent less common ethnicities, comprising of 13 percents, 13 percent and 11 percent respectively. Figure 3.2 displays the distributions of the genders in the

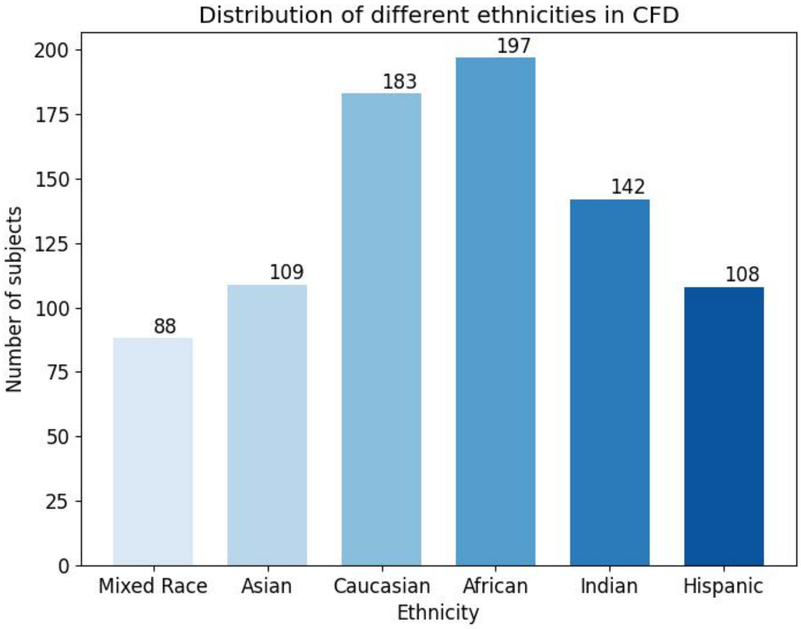


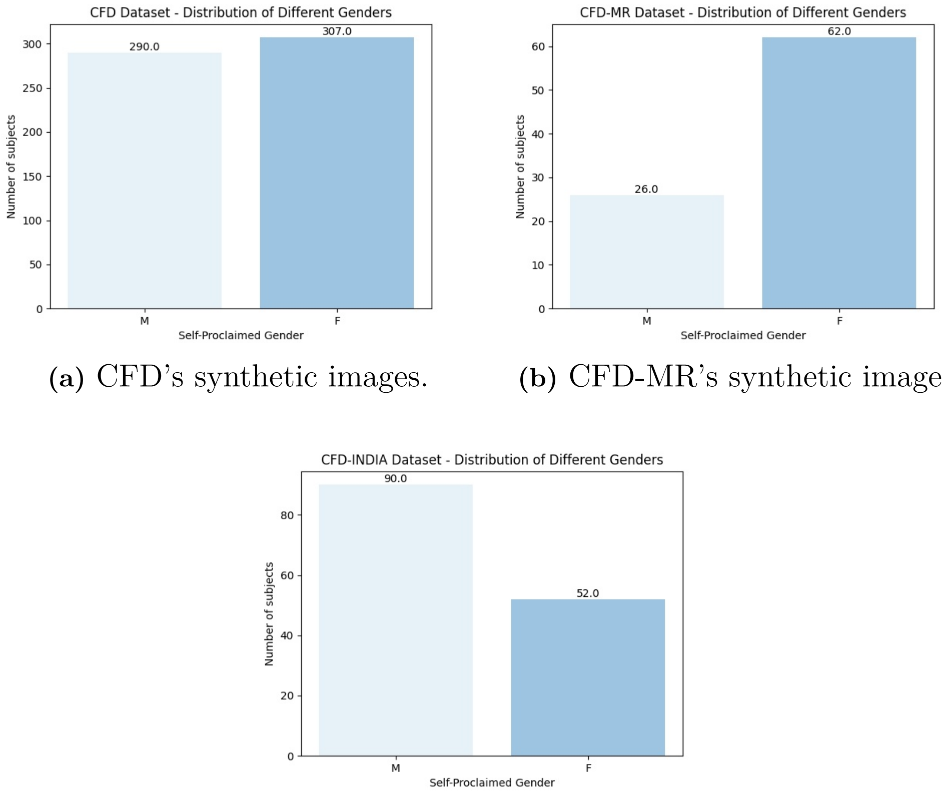
Figure 3.1: The ethnic distribution within CFD is illustrated in this figure, including African American, Asian, Caucasian, Latin, and Indian.

CFD, CFD-MR and the CFD-INDIA.

Finally, Figure 3.3 shows the gender distribution in CFD. Although there appears to be a modest difference in quantity between the two genders, it is negligible. Therefore, it is can be concluded that the gender distribution is quite equal for both gender, indicating that CFD does not have bias in this category.

Visual assessment reveals significant variations in the activation and intensity of specific AUs during facial expressions across different ethnicities. In the CFD dataset, a prominent trend emerges where certain ethnic groups exhibit higher AU intensities during expressive emotions, while other groups tend to utilize a lower number of AUs. Consequently, the intensity of AUs observed in the facial expressions of the latter groups is comparatively less pronounced.

Figure 3.5 displays a bar plot with a trend line for the frequency distribution of different ages in CFD. The x-axis represents different age values, while the y-axis shows the frequencies of each age value. The blue bars represent the frequency of each age value, while the red trend line represents the interpolated line between the tops of each bar.



(a) CFD's synthetic images.

(b) CFD-MR's synthetic images.

(c) CFD-INDIA's synthetic images.

Figure 3.2: All of the figures, (a), (b) and (c) display individual distribution if the genders.

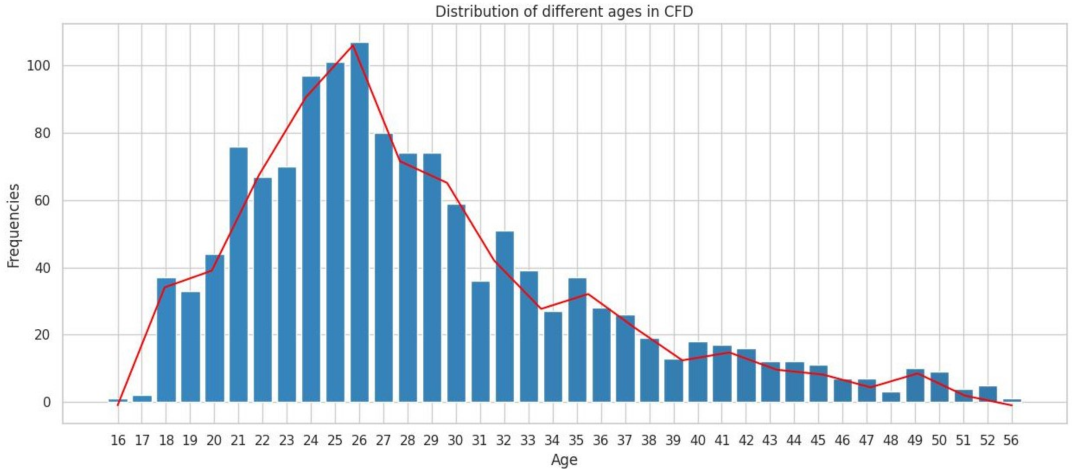


Figure 3.5: Distribution of different ages in original CFD.

Additionally, descriptive analysis of the age category in CFD was conducted and a range of statistical measures was derived and is presented in the following bullet points.

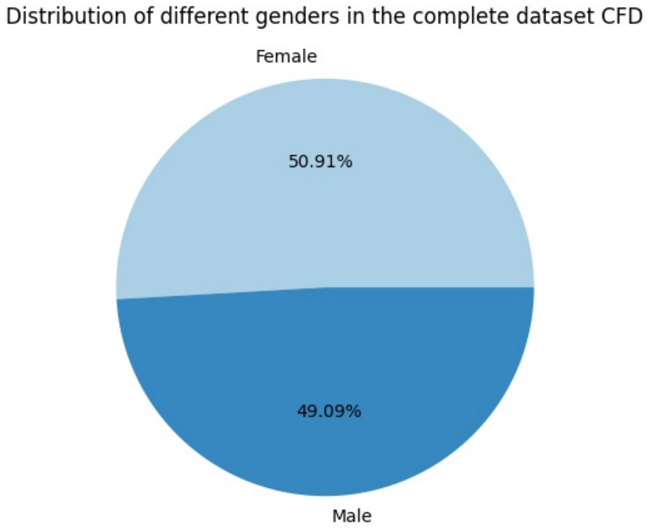


Figure 3.3: Distribution of genders in the complete dataset of CFD.

- Mean = 34.58
- Median = 34.5
- Range = 40
- Standard Deviation = 11.11
- Outliers = 0

The age distribution in the CFD dataset reveals several insights. The mean age of the subjects is 34.58 years old, which indicates the average age of the participants. The median age is determined to be 34.5 years old, representing the central value of the dataset. Thus the distribution in the CFD exhibits a slight left skewness, suggesting a prevalence of younger individuals within the dataset. Additionally, the age range spans 40 years, with the minimum age recorded at 16 and the maximum age at 56. The magnitude of this range, measuring 40 years, is accompanied by a standard deviation of approximately 11 years. Notably, the distribution displays a flatter tail, indicating a relatively lower representation of middle-aged and elderly individuals compared to young adults in the dataset.

Based on the provided data, zero outliers are observed in the age distribution. Outliers refer to data points that deviate significantly from the rest of the dataset in the current considering category, which is age. It is worth noting that outliers can have a notable impact on the mean, but the median remains unaffected by their presence. Thus, it is crucial to identify and handle outliers appropriately, although in this case, their absence implies a relatively homogeneous dataset.

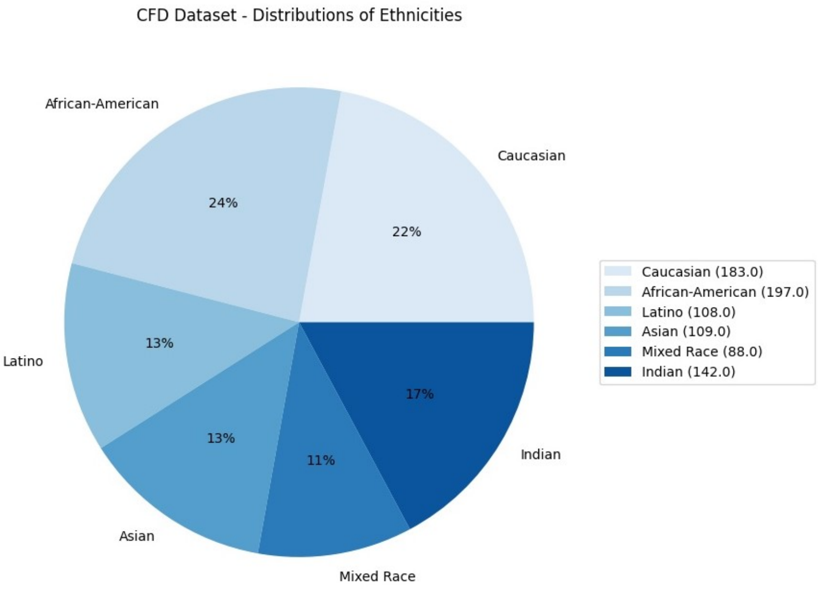


Figure 3.4: The ethnic distribution within CFD is illustrated in this figure. It comprises images of individuals with varying ethnicities, including African American, Asian, Caucasian, Latin, Indian descent and Mixed Race.

3.4.2 FFHQ compared to CFD

As per previous sections, relevance of the dimensions is domain-specific, and in the context of emotion prediction for clinical self-recordings as has been described in [2]. Consequently, the selection of a dataset is deemed to encompass pertinent dimensions if it satisfies two stipulations outlined in the aforementioned Section 3.3.

The first requirement is that the dataset must have a large demographic distribution, meaning that it should include subjects of different ages, genders, and ethnicities. It is also beneficial for the dataset to include common feature variations of the human face, such as glasses, beards, and make-up.

The second requirement is that the dataset should represent the human emotion range in a naturalistic fashion. This means that the emotion expression should occur spontaneously without too much involvement from an instructor. Unfortunately, many datasets are lab-recorded with varying elicitation methods, which can affect how natural the emotion expression is. Elicitation methods range from watching funny, sad, or scary video clips to performing expressions on demand.

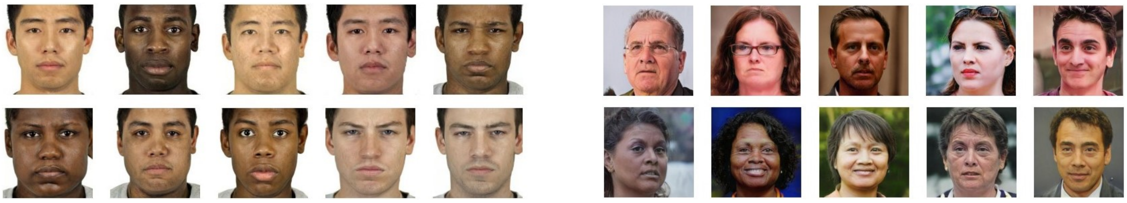
Overall, a high-quality dataset for emotion prediction in clinical self-recordings should have a high density and span relevant dimensions, including a large demographic distribution and naturalistic emotion expression. Flickr-Faces-HQ (FFHQ) is a large-scale face dataset that contains 70,000 high-quality images of human faces, with resolution up to 1024x1024 pixels. The images that were collected from Flickr are diverse in terms of age, gender, and ethnicity. Additionally, the dataset was created by Nvidia and is often used for training generative models such as GANs and VAEs (Variational Autoencoders).

In contrast, CFD is a comprehensive collection of high-resolution facial images that, although cannot be classified as a large-scale dataset, still offers considerable diversity as well as three sub-datasets, which comprises 827 subjects.

The difference between the datasets is that CFD offers ethnic diversity, including Asian, African, Latino, Indian and Caucasian subjects. All subjects have diverse range of facial expressions, including neutral, angry, fearful, but also a subset of subjects with happy expression with either open or closed mouth. One of th subsets is the CFD-MR, which includes the mentioned expressions and features but with individuals who self-reported multiracial descent. Further, CFD-INDIA includes subjects with Indian descent.

The main differences in the dataset's appearances are the varieties in lighting and the environmental factors. FFHQ, while producing high-quality synthetic images has many downsides. The focus of the environment causes the synthetic images to focus on the surrounding not just the faces in the images; whereas the CFD's original dataset was obtained in an adequately illuminated studio environment, and every subject is standing in the centre of the frame without any head tilting. The background for all subjects in CFD are uniform with a white adequately illuminated background.

Furthermore, the differences in lighting and environmental factors is what separates the dataset from each other. To some extent that can affect the performance of the GAN model trained on them. It is still important to note that even then GAN models are capable of generating images that are consistent with the distribution of the training data. Ultimately, there is no clear consensus on which datasets actually will be more suitable for styleGANs training. Nevertheless, Figure 3.6 demonstrate that the surroundings of FFHQ's synthetic images are blurry and highly distorted whereas each synthetic image in CFD offers little to no variation in the background. To conclude, CFD is the dataset that fulfils both of the two requirements. However, to make the diversity even more apparent that textual based here is a visualization to show what the diversity of the CFD contra the FFHQ, as shown in Figure 3.6. Figure 3.7 should only give a sense of how the distributions of gender, age, expressions is in



(a) CFD's synthetic images.

(b) FFHQ's synthetic images.

Figure 3.6: Figure (a) has a not so varying background for each of the images and multiple different backgrounds for the FFHQ's synthetic images (b).

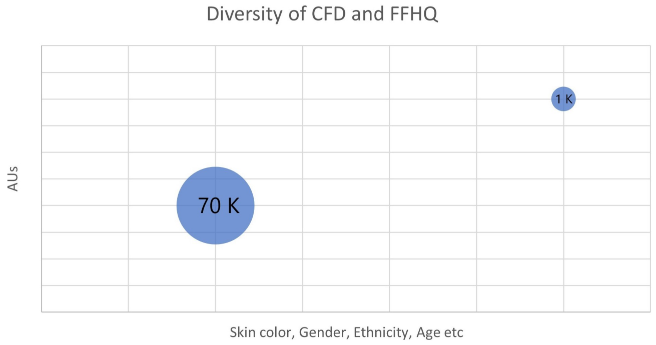


Figure 3.7: The variations in the two datasets FFHQ (to the left with 70,000 images) and CFD (to the right with 1,127 images).

each of the datasets and it should also give the impressions in a more understanding way. The y-axis conveys the AUs this thesis focuses on and the x-axis are the features that are the most interesting. Note that this is not an exact features diversity of all the image within each set only an simplified way of showing the diversity as is perceived by analysing both original datasets.

3.5 Dataset Construction and Preprocessing

In order for machine learning models to perform well, it is essential to ensure a thorough understanding of the general objective: developing models that can achieve high performance on novel data. Therefore, the chosen data sets in this project should meet such criterias for validity and generalizability. It is essential for the training data in this research to reflect the distribution of the typically expected demography in clinical trials. With this in mind one can find the right dataset that allows to fulfill certain criterias as stated in Section 3.3.1.

The first dataset DISFA has subjects that were captured by two cameras, positioned on the right and left sides of the subject. However, only the video data from the right camera was utilized for the study, while the left camera data was excluded to enhance data efficiency since it was almost identical to the right camera data. Additionally, the study found that horizontal flip augmentation, as discussed later in this chapter, which mirrors the target, compensates for the omission of the left camera sequence. The dataset was constructed by carrying out face localization and cropping, selecting successful crops (i.e capturing a face without cropping out section of the subjects face out of the image), sampling labels, binarizing labels, and subselecting frames.

The CFD or the Chicago Faces Dataset if a small-scale dataset and does not involve video in order to capture images. Instead individuals are posing in five different expressions and every image has the same kind of lighting.The CFD dataset undergoes pre-modification, particularly with regards to the synthetic images, before proceeding to the subsequent stages of the pipeline. This pre-processing phase involves operations such as cropping, face detection algorithms, and more.

Based on this analysis regarding dataset requirements, it was determined that FFHQ offered limited advantages in comparison to CFD. Moreover, CFD has to some extent encompassed all the strengths of FFHQ. Therefore, this research will be conducted exclusively using the real datasets DISFA and CFD. Henceforth, FFHQ, as a dataset option, has been eliminated.

3.6 Feature Extraction and Face Cropping

The first phase, **Phase 1**, of the pipeline and in the case of first processing the original dataset of CFD in the beginning was to first convert the images and data (each image that passes through the pipeline). The first phase of the pipeline is to firstly filter out images using the OpenCV [60] library in Python. Face-cropping is done by excluding irrelevant information in the frame such as background of unnecessary clothing, this way one can control the exposure to a minimal variation in both background but other surrounding features. The cropping is done by reading each image; determining where the centre of the each individual image is (since not all have the same height, width of the face etc).

Algorithm 1 Face Alignment and Cropping for Multiple Images

```
Requires $images \geq 0$
Face and detector loaded
DLib's detector
Path DLib's predictor
load images from folder
Detect faces in the image using DLib's frontal face detector
while $images \neq 0$ do Iterate over each detected face and align it
    Determine the facial landmarks for the face using DLib's shape predictor
    Extract the coordinates of the left and right eye landmarks;
    Extract the coordinates of the left and right corners of the mouth;
    Calculate the center of the mouth;
    Calculate the angle between the line connecting the eyes and the center of the mouth;
    Rotate the image by the calculated angle;
    Calculate the bounding box of the face using the facial landmarks;
    Calculate the center of the bounding box and the distance from the center to the edge of the box;
    Scale up the distance from the center to the edge by a factor to zoom in on the face;
    Calculate the new dimensions of the bounding box using the scaled distance;
    Crop the image using the new bounding box dimensions;
return Return aligned image
```

To determine the exact centre of an image one can use the dlib's facial landmark predictor. For every image and for every face it first aligns the image, then determines the facial landmarks for the face, extract left and right eye landmarks. It then extracts the mouth landmarks. Secondly, it calculates the angle between the eyes and the mouth, calculates the box of the face and then calculate the centre of the bounding box and the distance from centre to the edge of the box. Last step is to scale up the distance from centre to the edge by a factor to zoom in on the face. It then calculates the new dimensions of the bounding box using the scaled distance and then crops back the image to the size requested. Following this logic, here are the guides to a pseudo-code that follows the exact same steps but in a more general natural language that is easier to interpret, as shown in Algorithm 1.

The face extraction is used as the initial detector in order to assure that faces appear in pictures or if they are just random images with poor symmetric and/or wrongly cropped images. Dlib library is a popular open-source library that provides various machine learning algorithms and tools for developing applications related to computer vision, image processing, and machine learning. It includes functionalities for facial recognition, object detection, and facial landmark detection. OpenCV is a widely-used open-source computer vision library that offers a comprehensive set of functions and algorithms for image and video processing tasks. It provides numerous pre-trained models, including a face detector, which is commonly employed for detecting faces in images or video frames. The pseudocode describes a process that utilizes Dlib's face detection and facial landmark detection capabilities in conjunction with image manipulation techniques.

The face extraction process serves as the initial detector, ensuring that only images containing faces are captured. This is crucial for filtering out random images that lack proper symmetry or are incorrectly cropped. Figure 3.8 provides an illustration of the types of images that would be excluded. By analyzing the presence of key facial features, such as the eyes and mouth, and assessing the symmetry of facial landmarks, the face extraction process effectively identifies and filters out images that do not meet the criteria for inclusion.

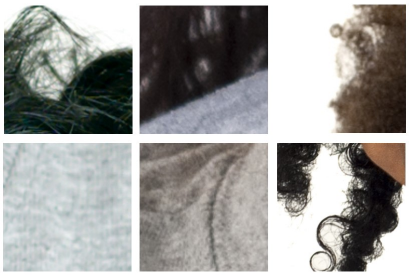


Figure 3.8: This figure displays the subset of images from the original CFD dataset that were not subjected to precise cropping during the initial processing stage.

Note that the figure above is the process done on the original dataset. Here both the cropping and the face detection was implemented. However, the same processing is not run of the synthetic images since the images are already resized to the desired shape, one solely needs to run the face detection algorithm on the synthetic dataset.

3.7 StyleGAN2-ADA Training on customised dataset of CFD and Pickle File Generations

The utilization of the generative model styleGAN2-ADA's potential was examined in a predecessor project, where the FFHQ dataset was utilized as the input data source. This decision was motivated by several factors, including the ease of the process, as the dataset had already been pre-trained on the model by NVIDIA. Additionally, various pre-trained models of different resolutions are also readily available for download. Thus, using the pre-trained model generated using the FFHQ dataset was chosen in order to facilitate the experimental process and to streamline data collection for subsequent analysis. However, no training was involved in that process.

In this current project, the lastly cropped version of CFD was utilized as a customised dataset for training styleGAN2-ADA. This process required a total training time of 35 hours and 45 minutes. To avoid potential loss of training progress due to cutoff run-times, pickle files were frequently generated. Specifically, after every 10 ticks, a pre-trained model was generated to log the current training progress.

In the context of this customised training process, there was considerable flexibility in determining the optimal stopping point for model training. Nevertheless, it is crucial to identify the appropriate time to terminate the training process. In this project, the primary criterion utilized for determining the optimal stopping point was visual assessment, where the training process was halted upon achieving a level of image quality that appeared visually realistic with a reasonable resolution. Thus, the training process was deliberately terminated at a point where the generated images were deemed to be of sufficient quality for the intended purpose of the project.

3.8 AU Subselection

As per [61], certain AUs are associated with one or more of the prototypical expressions discussed in section 2.1. Therefore, to guide the sub-selection of AUs for pipeline modeling, the prototypic expression framework is considered relevant, even though the actual prototypical expressions have less importance in actual prediction.

A subset of the total 12 AUs has been chosen for analysis based on their presumed association with emotions expected to be present during a clinical trial self-recording. The study hypothesizes that emotions such as happiness, sadness, and anger hold relevance in the trial setting, whereas emotions such as surprise and disgust are of lesser relevance.

This hypothesis is grounded in the belief that happiness may stem from internal emotional processes, while surprise is more likely to arise from external events unrelated to the subject's well-being. However, it is important to note that this hypothesis has not yet been substantiated. Table 3.2 provides an overview of the six chosen AUs for modeling purposes and their corresponding prototypical expressions.

Expression	Action Unit Number	Action Unit Name
Sadness	AU1	Inner Brow Raiser
	AU15	Lip Corner Depressor
Happiness	AU6	Cheek Raiser and Lid Compressor
	AU12	Lip Corner Puller
Anger	AU4	Brow Lowerer
	AU5	Upper Lid Raiser

Table 3.2: AUs Corresponding to Specific Facial Expressions

3.9 The Selection Pipeline

In computer science, a pipeline refers to a sequence of processes applied to an input, such as code or artifacts, for continuous testing purposes. Typically, a pipeline consists of multiple stages, each performing a specific task on the input data, with the output of each stage becoming the input for the next. Similarly, the pipeline used in this work simulates a TDS pipeline, aiming to test images on various pre-processing techniques. Pre-processes encompass both quality-related processes and necessary steps to ensure that the synthetic images processed through the pipeline are both naturalistic and proportional. Idea of the pre-processes is that they include both processes that convey numbers of quality but also some necessary steps to insure that the synthetic images that are prompted through the pipeline are indeed naturalistic images and proportional at the same time.

A commonly used classification pipeline consists of three main steps: preprocessing, feature extraction, and classification. Pre-processing involves transforming the raw image data into a format that can be used by the feature extraction algorithm. Feature extraction involves identifying relevant features in the image that can be used to classify or detect objects. Classification involves using a model to assign labels to the input image based on the extracted features.

Additionally, the pipeline may take into account temporal or spatial information, depending on the application. Real-world applications would integrate these steps into a series of processes that can be run automatically and at scale, without human intervention between pre-processing, feature extraction, and classification.

- Phase 1
  - Process of facial detection
  - Process of face cropping
- Phase 2
  - First-Stage
    - Process of measuring first-stage NR-IQA
  - Second-stage
    - Process of measuring second-stage IQA using FID
  - Third-stage
    - Process of measuring referenced IQA using deep learning (BRISQUE

via SVR)

- Phase 3
  - Selection of synthetic batches of images that preformed best in regards to quality score

Considering the aforementioned phases, each score in every phase can be validated through various methods, commencing with **Phase 1**. In this initial phase, a verification process is initiated, involving human intervention, to ensure the accuracy of the cropping stage results.

Proceeding to **Phase 2**, there exist three distinct processes, each with the objective of assessing image quality in different manners: reference-based image quality assessment and non-reference image quality assessment (NR-IQA). In the first process of **Phase 2**, NR-IQA is executed, followed by a subsequent validation process that utilizes the FID score to gauge the level of similarity between the two datasets. This validation step aims to analyze the coherence and consistency of the results obtained from both implementations.

Furthermore, the subsequent steps entail the utilization of various metrics to yield final outcomes. At this juncture, it becomes feasible to draw conclusions regarding the similarity of the two datasets when using the original datasets as reference points. This analysis enables the determination of whether the generated images possess comparable resolution and quality to the original images. Ultimately, after the execution of SVR, the final selection of test and train data is performed. The implementation of algorithms to select the highest ranked synthetic images is guided by five metrics (RMSE, PSNR, ISSM, SSIM, SAM and QIO). This process is executed in a batch script fashion, and its outcome serves as the ultimate output of the entire pipeline for integration into the classification model.

3.9.1 Selection of IQA metrics

Phase 2 involves the calculation of metrics for two datasets: the original dataset images and the dataset consisting of synthetic images. In the first phase, the metrics used are those employed in No-reference Image Quality Assessment (NR-IQA), namely Root Mean Squared Error (RMSE), Image Spatial Spectral Mutual Information (ISSM), and Spectral Angle Mapper (SAM). It is important to note that calculating these metrics may be irrelevant if the synthetic images lack real distortions. To address this, human annotation was conducted to confirm the presence of the necessary distortions and noise for the relevance of IQA metrics. Analysis reveals the presence of various distortions and significant levels of noise. In previous research, [62][63], on NR-IQA algorithms informed the selection of these metrics, indicating that all of them, including RMSE, ISSM, and SAM, can serve as objective quality metrics for image quality assessment.

Algorithm 2 Fréchet Inception Distance

```
Requires $images \geq 0$
Face and detector loaded
Import necessary libraries
function LOAD IMAGES FROM DIRS($dirs$)
    $images \leftarrow []$
    for $dir \in dirs$ do
        for $file \in dir$ do
            $image \leftarrow$ load image from $file$
            $images.append(image)$
        end for
    end for
    return $numpy.array(images)$
end function
$real\ dirs \leftarrow$ [list of directories containing real images]
$synthetic\ dirs \leftarrow$ [list of directories containing synthetic images]
$real\ images \leftarrow$ load images from dirs($real\ dirs$)
$synthetic\ images \leftarrow$ load images from dirs($synthetic\ dirs$)
while $images \neq 0$ do
    function CALCULATEFID($\mathbf{R}$, $\mathbf{S}$, $batch\ size$, $resize\ images$, $image\ size$)
        $\mathbf{M} \leftarrow$INCEPTIONV3MODEL()
        if $resize\ images$ is True then
            for $\mathbf{r}, \mathbf{s}$ in $(\mathbf{R}, \mathbf{S})$ do
                $\mathbf{r} \leftarrow$ RESIZEIMAGE($\mathbf{r}, image\ size$)
                $\mathbf{s} \leftarrow$ RESIZEIMAGE($\mathbf{s}, image\ size$)
            end for
        end if
        for $\mathbf{r}, \mathbf{s}$ in $(\mathbf{R}, \mathbf{S})$ do
            $\mathbf{r} \leftarrow$ PREPROCESSIMAGE($\mathbf{r}$)
            $\mathbf{s} \leftarrow$ PREPROCESSIMAGE($\mathbf{s}$)
        end for
        $\mathbf{A}_r \leftarrow$ COMPUTEACTIVATIONS($\mathbf{R}, \mathbf{M}, batch\ size$)
        $\mathbf{A}_s \leftarrow$ COMPUTEACTIVATIONS($\mathbf{S}, \mathbf{M}, batch\ size$)
        $\mathbf{m}_r \leftarrow$ COMPUTEMEAN($\mathbf{A}_r$)
        $\mathbf{m}_s \leftarrow$ COMPUTEMEAN($\mathbf{A}_s$)
        $\mathbf{C}_r \leftarrow$ COMPUTECOVARIANCE($\mathbf{A}_r, \mathbf{m}_r$)
        $\mathbf{C}_s \leftarrow$ COMPUTECOVARIANCE($\mathbf{A}_s, \mathbf{m}_s$)
        $FID \leftarrow$ COMPUTEFID($\mathbf{m}_r, \mathbf{C}_r, \mathbf{m}_s, \mathbf{C}_s$)
        return $FID$
    end function

    FID score ← CALCULATEFID(real images, $synthetic\ images, batch\ size$, True, $image\ size$)

    real plot images ← SelectFirstTen(real images)
    synthetic plot images ← SelectFirstTen(synthetic images)
    PLOT SCATTERPLOT
```

The code implemented in this phase incorporates several functions, such as *calculate_rmse*, *calculate_sam calculate_issm*. These functions accept lists of original and distorted images and compute the corresponding scores for each image quality metric. The calculated scores are stored in lists *(rmse_scores)*, *(issm_scores)* and *(sam_scores,)* and returned as numpy arrays. The results obtained will be presented in Chapter 4.

However, not all of them are suitable for no-reference image quality assessment (NR-IQA), which aims to predict the quality of an image without using any pristine, reference images. In **Phase 2**'s , **second stage** process includes calculating the FID scores. This includes both the baseline as well as the training data selection (more specifically 7 (included with neutral class) unique selection of classes and their corresponding number of data). The algorithm that follows the FID logic is shown below in Algorithm 2 and shows the distance between two Gaussian (distributions).

Algorithm 2 demonstrates the customized implementation of the Fréchet Inception Distance. Note that r stands for the real data and s stands for synthetic data. The results from the FID calculation on the synthetic dataset CFD and on the original datset CFD are reported in the upcomming Chapter 4.

The **third stage** of **Phase 2** involves the calculation of metrics for two datasets: the original dataset images and the dataset comprising synthetic images. However, in this stage, the scores are determined using a referenced approach via SVR which in order implements feature image quality prediction assessment. The scoring is based on the image quality metric called BRISQUE, which is employed to predict the quality of unseen data (i.e., synthetic images) by training on the training data (i.e., original images). The implementation of this assessment focuses on 8 main steps, which are succinctly summarized in an algorithm provided in the Appendix, as shown in Algorithm 4.

3.10 Annotation of synthetic data

Obtaining high-quality data is crucial to achieve superior performance in ML, and data annotation is considered the fundamental basis for achieving this goal. Professional annotation can greatly enhance the quality of the input but requires a significant investment. The baseline model solely uses a model's prediction to annotate the presence of AUs in the generated synthetic data. This prediction is subsequently employed as annotation method applied to the synthetic dataset Eigenfaces in the previous project. The outcome in term of performance was nearly identical to that of the baseline result. However, sevaral potential reasons for the outcomes were anticipated in the report but they do not involve this particular aspect of the method.

However, human annotation can still be excessively costly and somewhat unreliable, especially when performed unprofessionally and on a small scale such as the performed method in the previous predecessor project. Therefore, there is a interest in investigating a new approach to semi-automate the annotation of synthetic data based on the selected seeds. During this development process, the NPYViewer tool discussed in Section A.3 was employed to read the generated annotations stored in NumPy files.

To ensure high control when generating images, seed annotation was performed through the manual review of over 30,000 images, corresponding to 30,000 seeds. From this review, a set of 658 potential seeds was identified. These selected seeds, along with 150 chosen indexes and degrees, were then utilized to generate images depicting specific AUs of interest. Additionally, 144 seeds representing neutral expressions in the generated dataset were also chosen.

The inclusion of seeds in the annotation process allows for precise control and exploration of the latent space, enabling targeted generation of images with specific characteristics or expressions of interest.

3.10.1 Limitations with single AU-based Seed Selection

The initial method investigated in this study focused on the selection of seeds where each seed's synthetic image represents a single AU. Additionally, the approach involved monitoring the presence of six specific AUs within the annotated seeds. Consequently, if the generated image of a particular seed depicted a specific AU, the seed would be annotated in the corresponding AU list during the manual seed selection process.

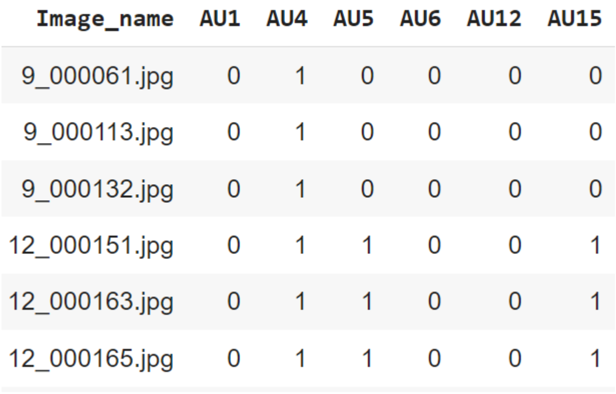
However, this approach has certain limitations. One major concern is the lack of an association method between the seed lists and the corresponding images, making it challenging to annotate images based solely on the seed information. This challenge stems from inadequate planning during the image review and seed list generation process. Moreover, facial expressions typically involve the simultaneous activation of multiple AUs, as observed in the original CFD dataset. Consequently, synthetic images generated from CFD often depict combinations of AUs within a single image. Selecting images that exclusively represent a single AU per image becomes exceptionally difficult in such cases. Additionally, the manual process of seed selection had to be constrained due to its reliance on human intervention, which inherently consumes a significant amount of time.

Considering these difficulties and the fact that real-life subjects often display several AUs, the chosen solution is to focus on single images depicting combinations of AUs. This decision aligns with real-life scenarios and allows for a more comprehensive analysis of facial expressions. By capturing the interplay between different AUs in a single image, the solution aims to obtain a realistic and representative dataset for this research.

3.10.2 Semi-automated Annotation for Synthetic Data

Given the limitations encountered in the previous approach, which hindered the implementation of annotation on scale, additional advancements were pursued to align the initial approach with the objective of human annotation. This approach is characterized as semi-automated, as it involves manual annotation for seed selection, while the generation of multiple images for each chosen seed is performed automatically without requiring further human intervention at the individual image level.

As a result of this development process, an algorithm was devised to automatically apply annotations to multiple images of varying degrees that are associated with the same seed during the generation stage. The semi-automated approach encompasses a systematic procedure for seed identification, referred to as seed-ID, which leverages the inclusion of seeds within the image names as prefixes. For instance, an image follows the format "seed-ID_numerical order.jpg" to ensure uniqueness and prevent overwriting. To illustrate, consider the image named "9_000001.jpg" depicted in Figure 3.9. In this example, "9" represents the seed styleGAN2-ADA utilized for image generation, while "000001" denotes the numerical order assigned to maintain distinct image names.



Image_name	AU1	AU4	AU5	AU6	AU12	AU15
9_000061.jpg	0	1	0	0	0	0
9_000113.jpg	0	1	0	0	0	0
9_000132.jpg	0	1	0	0	0	0
12_000151.jpg	0	1	1	0	0	1
12_000163.jpg	0	1	1	0	0	1
12_000165.jpg	0	1	1	0	0	1

Figure 3.9: This figure portrays an illustration of the Semi-automated AU Annotation approach for the generation of synthetic data.

Furthermore, the approach incorporates seed-tracking lists that are established based on the presence of specific AUs within the images during the manual seed selection process. Binary annotation is employed, where each image is annotated with a marker of either 1 or 0, indicating the presence or absence of a particular AU, respectively. Additionally, the approach involves monitoring the occurrence of the six AUs within each image, enabling comprehensive tracking of their presence or absence and thus enables a balanced distribution of the AUs of interest throughout the generated synthetic dataset SCFD.

The method involves mapping prefixes to corresponding AU. The algorithm retrieves each image's name and checks if the prefix exists in the associated AU-based seed-tracking lists. If the prefix is found in a certain AU lists, the current image will be annotated as having that AU in the annotation file using a number one. If a seed-ID is not found in a certain AU list, that current seed does not display that specific AU during the manual reviewing and seed annotation process mentioned above. Therefore, that seed has not been put into that certain AU list. In cases where a seed-ID is not associated with any AU list, it signifies that the seed does not possess any of the desired AUs and these seeds were categorized and placed in the neutral seed list.

This method is employed servers as a mechanism to facilitate an annotation approach that minimizes the reliance on labor-intensive manual efforts. By leveraging the seed-ID prefix embedded within each image name, this method enables streamlined data labeling. One significant advantage of integrating this annotation method with synthetic data generation is its potential to significantly increase the number of annotated synthetic images, while still requiring a relatively moderate number of initial seed annotations. Consequently, this capability allows for the generation of a substantially large synthetic dataset, which is well-suited for training and evaluating models, thereby facilitating the development and assessment of robust models.

3.10.3 Methods of evaluation model performance with Cross-Validation

This section presents the implementation of k-fold cross-validation and its procedure. The purpose of KCV in this study is to evaluate the performance of the DISFA, CFD, and FFHQ datasets. It involves dividing the dataset into k subsets or folds of approximately equal size, training the model k times, and assessing its performance on each fold using metrics such as ROC and ROC AUC.

In the first predecessor project, the cross-validation techniques was applied to combine various folds of the DISFA dataset with the entire curated version of the Eigenfaces dataset. In the current study, a similar methodology is adopted; however, there is a deviation in the dataset selection. Unlike the previous project, which incorporated the Eigenfaces dataset, the curated SCFD dataset is utilized as the synthetic data source in this research.

As shown in Figure 3.10, the video frames of twenty-six participants in the DISFA dataset are divided into subsets of thirteen folds, with two subjects per validation dataset for each fold. The first fold contains only the first two subjects as validation sets, while in the second fold, video frames from the next two subjects are used (i.e., subject 3 and 4). This process is repeated for each fold until reaching the last two subjects in DISFA for fold 13. After each split, the curated SCFD dataset is appended at the end of the remaining subsets of DISFA, and each of these curated sets is used as a training dataset for each fold training. Thus, the total number of images in the training set in each fold is

$$40,029 + (26 - 2) \cdot 300 = 47,229$$

number of images.

The implementation of k-fold cross-validation is demonstrated through the 'split_DISFA()' function. Firstly, the labels are loaded from a file, and then 'k' splits of the dataset are generated. For each split, the labels are loaded using the 'load_labels()' function, and the image file names are sorted to align with the label order using 'custom_sort()'. Subsequently, the data is divided into training and validation sets, and converted into

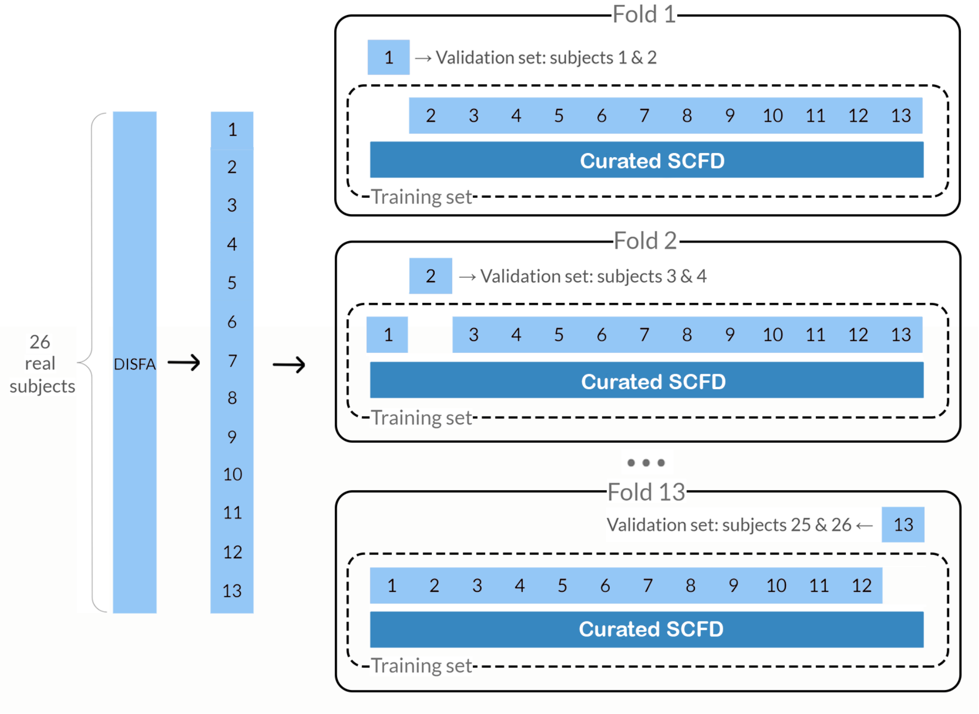


Figure 3.10: Illustration for 13-fold Cross-validation of curated DISFA and curated SCFD.

TensorFlow tensors using 'im_file_to_tensor()'. The function employs the 'yield' statement to generate a generator that produces the '(train, val)' datasets for each fold.

The advantage of k-fold cross-validation lies in its ability to provide a more robust estimation of the model's performance compared to evaluating it on a single random subset of the data. By training and evaluating the model on multiple subsets of the data, k-fold cross-validation offers a more reliable assessment that is less sensitive to the specific subset used for evaluation. Moreover, k-fold cross-validation aids in detecting overfitting, a situation where the model performs well on the training data but poorly on unseen data. By assessing the model on multiple folds of the data, k-fold cross-validation helps identify potential overfitting by measuring its performance on the validation data.

3.11 FER Pipeline Model

The subsequent sub-sections provide an account of the implementation of diverse FER models. This section will describe the methods that have been used in the predecessor project. Thus the summary will include the relevant methods used in the previous project [2].

The presented section provides a summary of the implementation details of FER models. These models were developed using Keras with the TensorFlow 2 backend and share common design choices. Color data with three channels is utilized, and the Adam optimizer algorithm is employed for training. Gradient clipping with a clipnorm value of 0.1 is used to address training instability, and early stopping based on the validation AUC metric is implemented.

The architecture of the models includes feature extractors augmented with two dense layers (of dimension 128) using Rectified Linear Unit (ReLU) activation and L1 regularization (with coefficient $1 \cdot 10^{-4}$).The final layer consists of six output neurons with sigmoid activation, representing different AU-classes. Binary cross-entropy (BCE) loss is employed for multilabel prediction, with each AU class treated independently. The loss function is calculated by summing the binary cross-entropy terms for all classes.

Finally, the choice of utilizing sigmoid activation and BCE enables comparisons between each output neuron's activation and its own probability distribution. This approach facilitates accurate representation of AU presence or absence.

3.11.1 Benchmark Model

The benchmark model utilizes an EfficientNet CNN feature extractor with an input tensor of size 224 x 224 x 3 and ImageNet-pretrained weights. The Keras EfficientNet B0 network with ImageNet weights is employed, and the model has approximately 4 million trainable parameters. The learning rate is set to $1 \cdot 10^{-6}$, and the batch size is 32. Training is conducted for 10 epochs, and the hyperparameter choices are determined through a brief grid search.

3.11.2 Model Enhancements and Training Approaches

To improve the model's performance, several enhancements and training approaches were implemented.

- **Class Imbalance Addressing:** A re-weighting approach was used to handle class imbalance. AU class weights were adjusted based on their occurrence, calculated using a formula considering positive and negative instances. The weights were applied in a customized weighted binary crossentropy loss function.

- **Subject-level Baselining:** False positives and neutral states mistaken as different emotions were corrected using subject-level baselining. Two approaches were employed: baseline subtraction subtracted the average model prediction of neutral frames, while siamese network baseline used pre-trained models and concatenated their outputs with neutral frames as baselines.

- **Advanced Models and Pre-training:** To overcome identity bias and limited data challenges, two pre-training methods were used. Supervised multi-stage pre-training involved pre-training VGG Face and EfficientNet B0 CNN backbones on RAF DB and AffectNet, followed by finetuning on the DISFA set. BYOL pre-training trained an EfficientNet B0 backbone using augmented views of the same image or pairs of images from the same emotion class.

- **Cross Validation Ensemble Construction:** Robust ensemble models were built by training multiple models using cross-validation. Each model was trained on a different group of training subjects, capturing various aspects of DISFA. The ensemble consisted of 13 models pre-trained on AffectNet combined through majority voting.

These enhancements and training approaches aimed to improve the model's performance by addressing the following: class imbalance, correcting false positives, utilizing pre-trained models, and constructing an ensemble model trained on different subsets of the dataset.

4

Results

This chapter provides an overview of the outcomes of the study. The result chapter will highlight the most significant discoveries and provide an interpretation of the results. Additionally, this chapter will include a discussion of the limitations and implications of the results. The result chapter will be structured and presented in a manner that supports the research questions and objectives and it also serves as a basis for future research in the field.

4.1 Process of Face Localization and Cropping

The process of identifying the face region in an image, known as face localization, is an important step in facial analysis tasks. Once the face region is recognised, it can be cropped and separated from the background, a technique known as face cropping. The purpose of face cropping is to remove irrelevant image information and to focus solely on the face region, making it easier to analyze and extract facial features that may be of interest. This also allows to minimise the background inclusion and the noise it introduces. The accuracy of face localization and cropping greatly affects the performance of subsequent facial analysis tasks, such as expression recognition and age estimation. Therefore, it is crucial to ensure that the face localization and cropping methods used are robust and reliable.

The face cropping pre-processing stage involves the use of both a landmark detector and OpenCV's HaarCascade for face localization. This ensures accurate detection of both symmetric and asymmetric faces. By using HaarCascade, it enables the detection of faces that cannot be identified solely based on landmarks. This approach allows for the detection of both symmetrical and asymmetrical faces, as some amount of asymmetry can be present in synthetic images to develop a more robust model. Here is the HaarCascade implementation that was implement as a process in the pipeline; see the algorithm below:

Algorithm 3 Face Detection for Multiple Images

```
Require: $images \geq 0$
Ensure: Import necessary libraries
Require: $image\_directories$: List of directories containing image files
Ensure: $aligned\_images$: Numpy array of loaded and aligned images
  procedure LoadAndAlignImages($image\_directories$)
      $aligned\_images \leftarrow []$
      Load the real and synthetic images
      while $images \neq 0$ do
          for $subfolder$ in $image\_directories$ do
              Set up the path for the current subfolder.
              Create a new subfolder in the new folder with the same name as the current subfolder.
              for $image$ in $subfolder$ do
                  Read in the image.
                  Detect faces with facial landmarks in the image.
                  if any faces are detected then
                      Save the image to the corresponding new subfolder in the new folder
                  end if
                  if no faces were detected in any of the images in the current subfolder then
                      Remove the new subfolder created for the current subfolder from the new folder
                  end if
              end for
          end for
      end while
      return $aligned\_images$
```

4.2 Preliminary Baseline Results without TSD and Semi-automatic Human Annotation

In the interest of ensuring a fair comparison of results using synthetic images, it is necessary to have a sufficiently large dataset. Therefore, in ths work the aim is to generate synthetic images that are at equivalent to the size of the original dataset to evaluate if there is a noticeable boost in performance attributed to the synthetic images. However, it is imperative to ensure an adequate number of images exist in both datasets to avoid overfitting and underfitting issues during k-fold cross-validation. The precise number of images necessary for this purpose depends on the complexity of the problem and the model employed. Given the moderate complexity of the baseline model and the absence of any requirement for the number of synthetic data to match that of the real data, an arbitrary quantity of synthetic data consisting of 43,280 images was generated in the SCFD.

The baseline model was trained using DISFA, SCFD and ROC AUC as evaluation method to obtain the baseline result. The results of the baseline model trained on 43,280 synthetic images in SCFD are presented in Figure 4.1, alongside the results from previous studies for comparison. The ROC curves for individual AUs indicate that certain AUs pose greater challenges for accurate classification compared to others. Specifically, AU1 (Inner Brow Raiser) exhibited low ROC values in comparison to the other AUs, indicating its difficulty in achieving highly accurate classification. Although the results indicate a moderate performance compared to previous studies that used Eigenfaces or only DISFA, the ROC values of various AUs remain the same or slightly worsen.

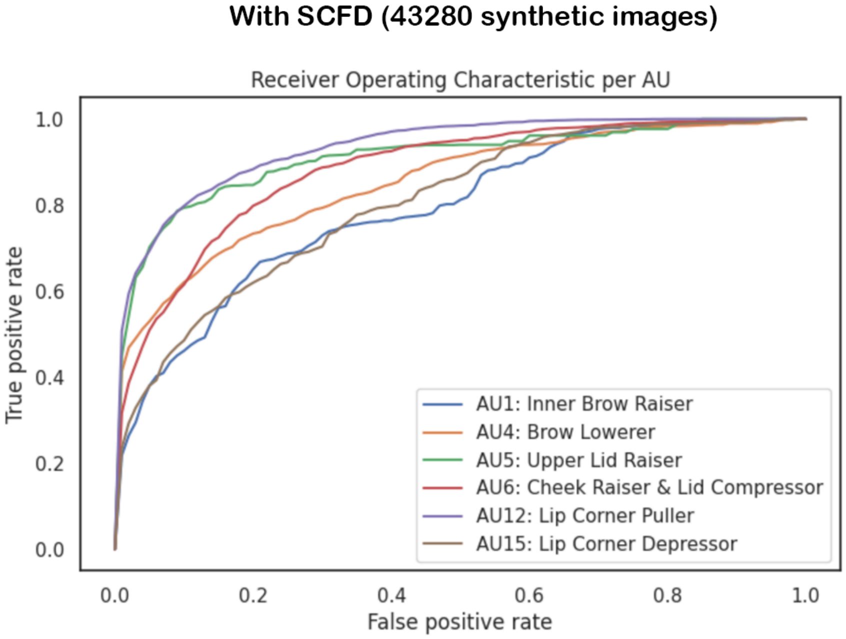


Figure 4.1: Receiver Operating Characteristic (ROC) baseline result of 43,280 synthetic images in the initial version of SCFD

Figure 4.2 illustrates the validation ROC AUC for each AU, averaged across all cross-validation folds. The results indicate that the baseline model already demonstrates good overall performance, even without the inclusion of the IQA pipeline. However, there is a noticeable disparity in performance between AU1 and AU12. This discrepancy suggests that the image quality of the class with poorer performance may be inferior to those with better performance.

These findings suggest that the increased number of synthetic images in SCFD, along with the present quality of the images, contribute to a good performance. However, the noticeable discrepancy between performances of the two classes AU1 and AU12 prompts an inquiry into the quality of the synthetic images employed. Additionally, the selection and annotation of seeds, indexes, and degrees may have been inadequate in certain aspects. Hence, there is a need to delve deeper into a

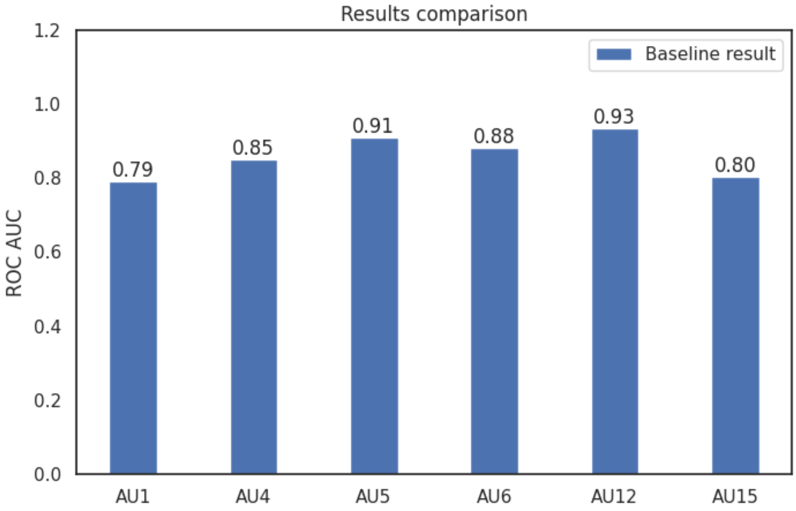


Figure 4.2: ROC AUC per AU baseline result using 43,280 synthetic images in SCFD.

more thorough examination of pipeline techniques such as Training Data Selection (TDS) to investigate the potential for enhancing the baseline result and potentially even improving the outcome of the classification model in this study.

4.3 The results from the Pipeline Phases

This section will discuss the results from each of the phases in the TDS pipeline. It will touch on the comparison between the original dataset and the synthetic dataset both in their structural differences and other image quality aspect. Moreover, each results is interpreted and will be analysed from a perspective of which subsets of classes (and accordingly the images) results in best quality. The last phase of the pipeline will have a clear outline in terms of what data needs to be the output of the pipeline, see Section 3.9 for more clarity.

4.4 Results from Pipeline

This section consists of the results from the pipeline, from Phase 1 (computing the NR-IQA) through Phase 3 (implementation of Support Vector Regression Image Quality Prediction). Detailed information regarding the setup of these phases is provided in Section 3.9.

4.5 Results from Phase 2

4.5.1 First-Stage Results - NR-IQA

The results from the first-stage NR-IQA was conducted by processing the original images and the synthetic images on a local resource on Chalmers University of Technology called Bayes. It includes both GPU an CPU computation resources and gives users options of submitting job to be run base on a scheduling schema with Slurm.

The results below are numbers of the first-stage NR-IQA processing. These numbers represent the evaluation scores for different metrics comparing an original dataset with a synthetic dataset. The metrics used are RMSE (Root Mean Squared Error), ISSM (Image Structural Similarity Measure) and SAM (Spectral Angle Mapper).

- **Root Mean Squared Error (RMSE)** - measure of the average difference between the values predicted by a model or algorithm and the actual observed values. In this case, the RMSE average score for the original dataset is 0.0, which means there is no difference between the original dataset and the predicted values. However, the RMSE average score for the synthetic dataset is 30.39, indicating a large difference between the synthetic dataset and the predicted values.

- **Image Structural Similarity Measure (ISSM)** - measure of the similarity between two images, where a higher value indicates higher similarity. In this case, the ISSM average score for the original dataset is 1.0, indicating a perfect similarity between the original dataset and the predicted values. However, the ISSM average score for the synthetic dataset is 0.33, indicating lower similarity compared to the original dataset.

- **Spectral Angle Mapper (SAM)** - measure of the similarity between two spectra, typically used for remote sensing or hyperspectral data. In this case, the SAM average score for the original dataset is 0.0, indicating a perfect similarity between the original dataset and the predicted values. Similarly, the SAM average score for the synthetic dataset is also 0.4, indicating a perfect similarity.

Based on the evaluation scores, it appears that the generated synthetic images have significant differences compared to the original images. The RMSE score of 30.38, and the ISSM score of 0.32 all indicate relatively low similarity or high dissimilarity between the synthetic images and the original images.

When comparing original dataset and synthetic dataset of images using SAM, RMSE and ISSM, the interpretation of what is considered a good score can vary depending on the specific context and requirements of the application. However, it's important to note that the SAM score, which measures spectral similarity, is perfect with a score of 0.4, indicating moderate similarity between the synthetic and original images.

This could suggest that the synthetic images may have similar spectral properties as the original images, but can differ in other image characteristics. RMSE and ISSM also indicate moderate similarity. The scores are also presented in Table 4.1, which is shown in Appendix A.

Metric	Score of Original	Score of Synthetic
RMSE	~ 0.1	~ 30.39
ISSM	~ 1.0	~ 0.329
SAM	~ 1.0	~ 0.4

Table 4.1: Results from the computation of NR-IQA algorithm for the metrics RMSE, ISSM and SAM.

4.5.2 Second-Stage Results - FID

The FID scores obtained for both the baseline models, namely the original dataset and the synthetic dataset, are approximately 26. These results suggest that the FID score benchmark display certain dissimilarities, although within an acceptable range. The observed dissimilarities need further reflection on the underlying reasons.

Primarily, it is important to note that the FID score is not directly associated with the implementation of quality assessment, unlike the process of NR-IQA. Instead, FID focuses on quantifying the similarity between two image datasets by implementing feature statistics extracted from a pre-trained deep learning model, typically this would be Inception. This methodology does not rely on reference images or explicit quality assessment measures. Figure 4.3 depicts the overall dispersion of the images, with red dots representing the synthetic images and blue dots representing the original images. Visually it demonstrates the spatial distribution of the images in the feature space. It also highlights the degree to which the synthetic images align with the original images. A higher degree of alignment indicates a better FID score. Overall, the results from the two processes, first-stage and second-stage in Phase 2 suggests that while there are dissimilarities between the synthetic and original images, but they fall within an acceptable range. Additionally, there are some discrepancies in terms of pixel-level dissimilarity, structural similarity, and spectral similarity, indicating that the synthetic images may not fully capture the characteristics of the original images.

4.5.3 Third-Stage Results - SVR

The last process of the pipeline that validates the image quality but in a reference image quality assessment is the SVR. With Support Vector Regression image quality prediction the image quality can be assessed from the perspective of the original images.

The code steps below performs an image quality assessment task using the BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) algorithm. The purpose is

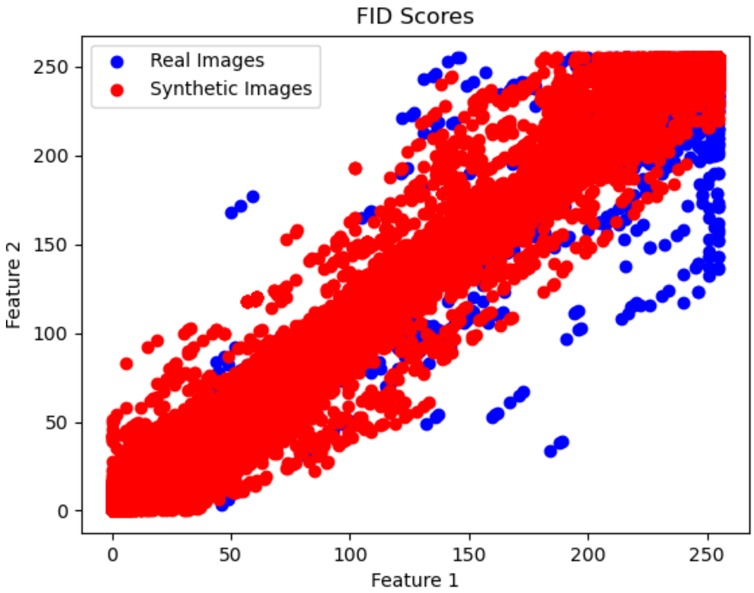


Figure 4.3: Figure displaying images in the feature space where red dots are the synthetic images and the blue dots are the original images.

to train a Support Vector Regression (SVR) model on a dataset of original images and then use the trained model to predict the quality scores for a separate set of synthetic images. The following steps summarize the process:

1. The code loads a dataset of original images from a specified folder.
2. Preprocessing is performed on the original images, and BRISQUE scores are calculated using the gray-level co-occurrence matrix.
3. The data is prepared for SVR training by reshaping the BRISQUE scores and generating dummy target values.
4. The original dataset is split into training and validation sets.
5. An SVR model with an RBF kernel is trained using the training data and feature scaling.
6. A separate dataset of synthetic images is loaded from a specified folder.
7. BRISQUE scores are calculated for the synthetic images.
8. The SVR model predicts the scores for the synthetic images.
9. The predicted scores for the synthetic images are printed.
10. Model performance is evaluated on the validation set using mean squared error (MSE).
11. The MSE is printed as a measure of the model's performance on the validation set.

The results from the SVR run show that BRISQUE scores for the synthetic images range from 0.355 to 0.619, with lower scores indicating better quality. The scores reveal that the quality of the synthetic images varies, with some images having relatively high scores (e.g., 0.619) and others having lower scores (e.g., 0.355). The mean squared error (MSE) is a metric used to evaluate the quality of a regression model and in this case, the MSE is used to assess the quality of the synthetic dataset compared to the original dataset. Further, in general, a lower MSE indicates better model performance, as it indicates that the perfromance of the model is closer to the actual values. Therefore, the MSE in this case indicates a good overall performance even though some outliers can have affected the last score somewhat. Lastly, note that the SVR algorithm implementation can be seen in the Appendix A.2.

4.6 Results from Phase 3

The final stage of the pipeline has been implemented using a batch script. It involves multiple steps to calculate No-reference Image Quality Assessment (NR-IQA) scores for all synthetic datasets, i.e all the 100,000 images. These scores are computed based on the five metrics explained in Section 3.9, two of them are not included in the first-stage process but are added as additional security. Those metrics are Quality Index Operator (QIO) and Peak Signal-to-Noise Ratio(PSNR). Subsequently, the images are ranked in ascending order based on their performance compared to some scores from images in the original dataset. This approach enables reduced reference image quality assessment and implements method of identifying images with the best quality and similar structural features.

The batch implementation follows a logic where each iteration processes a batch of 512 images. For each image, the relevant metrics are computed as described, and based on these scores, 180 images with the highest image quality are identified. Then their names and corresponding scores are recorded in a list. Simultaneously, these top-ranking images are relocated to a designated folder and the images that are not top ranked will also be discarded in the same way the top images are. The difference is that the rest of the 512 images are just discarded. It is worth noting that the randomly selected 512 images are permanently deleted from their original location, irrespective of their ranking. This step ensures that the images are not processed repeatedly and progressively reduces the total number of images in subsequent runs of the batch sizing method.

As a result of this process, 40,029 images has been retained from the initial pool of 100,000 generated images. These selected images is the curated set of TDS sample, serving as the input for the classification model. They represent the final outcome of the pipeline and are deemed to possess the highest quality among the generated images.

4.7 Semi-automatic AU annotation for curated dataset SCFD

Upon obtaining the curated SCFD dataset, comprising synthetic images with the highest IQA scores in the TDS pipeline, the dataset underwent annotation using the Semi-automated Annotation method. Consequently, a structured NumPy file was generated, adopting a six-column format as illustrated in Figure 3.9 (Section 3.10.2). The NumPy file encompasses binary annotations for a total of 40,029 images, exemplified by the illustration provided in Figure 3.9. In the subsequent phase of the project, this NumPy file together with the generated synthetic images within the curated SCFD, will serve as the input for the FER classification model.

4.8 FER Classification Model's Performance with Integrated TDS pipeline and Semi-automated Annotation

Following up to the first result benchmark that was obtained by utilizing uncurated synthetic dataset SCFD and without the TDS pipeline, this section presents the result and benchmark involving the integration of the TDS pipeline and the Semi-automatic Human Annotation method. The inclusion of TDS with IQA methods within the pipeline will be thoroughly examined to draw conclusive insights regarding their impact on performance improvement.

In Figure 4.4, the ROC curve for AU1 (Inner Brow Raiser) indicates that the enhanced classification model faces challenges in accurately predicting the presence of AU1. In comparison to other AUs, the model's performance in detecting AU1 is relatively lower. However, a comprehensive analysis of the curves reveals an overall improved performance compared to the baseline model. Notably, the orange curve representing AU4 (Brow Lowerer) and the green curve representing AU5 (Upper Lid Raiser) exhibit slightly better performance when compared to their respective curves in the baseline model. However, the ROC curve of of AU15 seems to have worsen quite a bit, indicating a slight degradation compared to the baseline result.

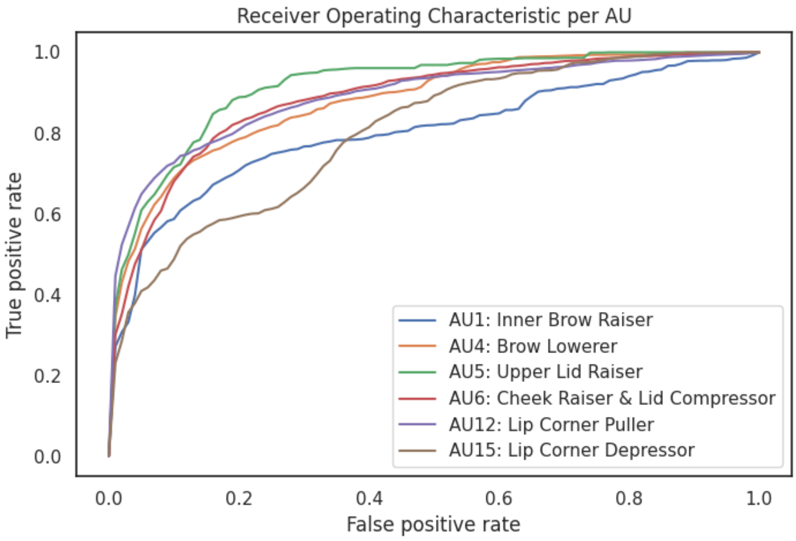


Figure 4.4: ROC AUC per AU baseline result of synthetic images in curated SCFD dataset

The application of the TDS pipeline in conjunction with semi-automated annotation resulted in new AUC values, reflecting the performance of the classification model.

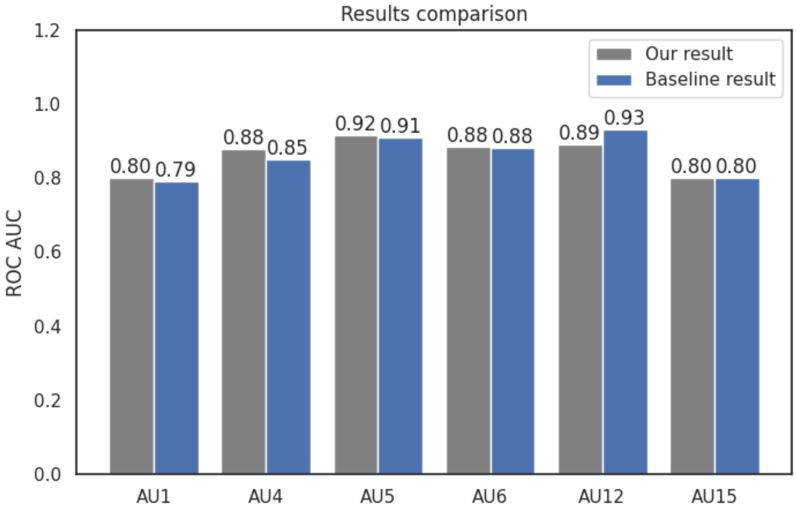


Figure 4.5: Summary of AUC values per AU of the original model, baseline model and the enhanced model with intergrated TDS pipeline and semi-automated human annotation using synthetic images in curated SCFD dataset

Notably, half of the classes exhibited improved performance, while 30% of them remained unchanged. The AUC values of AU1, AU4, AU5 have outperformed those of the baseline model, while the AUC of AU6 and AU15 remains the same as the result of both the baseline model that use the uncurated version of SCFD and the original model that only use 7,800 images of the real dataset DISFA. Furthermore, one class showed a decline in performance. This outcome suggests that there is potential for further refinement and optimization in both the TDS pipeline and the semi-automated annotation process to achieve even better performance across all classes.

Despite this outlier, the enhanced pipeline demonstrated an enhanced consistency in performance across the six classes. The deviation between their scores seemed to have been minimized, leading to a more uniform and consistent predictive performance across the classes. This observation implies an improved stability and reliability of the enhanced pipeline in accurately classifying instances from various classes.

5

Conclusion

This section presents an analysis of the research's findings and discusses the implications of the results and thereby draws conclusions. The section discussion focuses on important insights gained from the research and addresses limitations and challenges encountered during the research. Additionally, the conclusion summarizes the main findings and their significance, reaffirming the study's contributions to the study field of synthetic data generation and enhancing FER models' performance. Lastly, the section future work identifies potential areas for further investigation and suggests potential avenues to expand upon the current research.

5.1 Discussion

This section begins by revisiting the research questions and objectives, then assessing the extent to which they have been addressed. Furthermore, the main findings are discussed in details to identify relevance and disparities to the defined research questions. This chapter also addressed this reach's contribution to the existing knowledge in the field of synthetic image generation and enhancing FER classification model's performance.

5.1.1 TDS pipeline

The findings obtained from the different phases of the IQA pipeline have demonstrated a systematic approach for evaluating the quality of synthetic images. The evaluation has been implemented using both reference and no-reference IQA techniques. Furthermore, relevant domain knowledge of image assessment required in order to be implement the defined IQA pipeline. Additionally, it should be noted that during the development process, obtaining a larger synthetic dataset to serve as input for the classification model was one of the priorities. Hence, the IQA pipeline was designed to address not only the aspect of image quality but also the aspect of quantity in selecting images for the classifier. Therefore, the task involves balancing between collecting images with the highest reported scores across five IQA metrics and ensuring to obtain a sufficient number of images as training data. To achieve this, the method of batch sizing was implemented, which resulted in an amount of approximately 40,000 synthetic images.

Secondly, is important to highlight that the pipeline was implemented using stages as processes in order to validate the first run of the IQA, namely the NR-IQA. Once the NR-IQA was validated and shown to produce comparable results to RR-IQA, a number of NR-IQA methods were implemented, which also served as the final selection method. This approach allowed for a lightweight IQA that consisted of five different metrics, as opposed to using RR-IQA that sorely relied on the BRISQUE score.

Furthermore, the performance of the NR-IQA and reference-based IQA exhibited comparable results. This finding implies that the outcomes obtained from the TDS pipeline should not have an adverse impact on the benchmarks established for the baseline and final results of the classification model. Therefore, the focus should shift from selecting specific metrics for image selection in the classification model to implementing accurate methods for image assessment.

Consequently, a methodology wherein the input for each phase was derived from the output of the preceding phase was employed in the TDS process. This approach ensured that the results obtained from each phase sequentially validated one another, leading to a consistent conclusion. Through the execution of three analyses, a significant disparity in image quality between the synthetic and original images was identified. These findings serve to reinforce the reliability and effectiveness of the implemented IQA methods.

Moreover, the incorporation of the TDS pipeline has facilitated the identification and inclusion of training data of highest quality available in SCFD based on the evaluation of five selected metrics. As a consequence, the application of the TDS pipeline has led to an enhanced classification outcome of the final model in comparison to the baseline result. This result of the employed TDS approach substantiates the effectiveness of the methodology in training the model and optimizing the classification performance through the utilization of high-quality synthetic data.

Overall, these findings highlight the contribution of the pipeline towards evaluating as well as selecting of high-quality synthetic images, and thereby improved classification result. By improving the understanding and implementation of IQA techniques, this study provides valuable insights for future research endeavors in this domain.

5.1.2 Semi-automated Annotation method

The exploration of a trial approach involving the annotation of a single AU for each annotated seed has yielded valuable findings and insights. This investigation has shed light on several disadvantages associated with the approach in particular and human annotation in general. The idea of approach itself leads to a lack of association between the seed and the AU annotation, which makes it impossible to annotate at the later stage after data generation. With other words, it is unpractical and thus is low in robustness. Further, challenges with human annotation are the labor-intensive nature of manual annotation, logistical challenges arising from the absence of a well-designed data storage system, implications stemming from human errors and limited expertise in the field of annotation, and, notably, the time-consuming nature of the process. Hence, it is desired to automate annotation to the greatest extent possible, particularly when dealing with large-scale datasets such as those generated through synthetic data generation, to address the scarcity of available real data. However, previous research [3] has demonstrated a lack of compelling evidence supporting the effectiveness of leveraging machine learning model predictions as a means of annotation in the current state-of-the-art synthetic data generation. Therefore, development of a more robust and effective method than both of the method mentioned above is crucial for the aim of enhancing the FER classification model, as the quality and quantity of data have a significant impacts on model performance.

In order to tackle this, the proposed approach of Semi-automated Annotation method addresses the limitations of the previous project's approach, allowing for annotation on a larger scale. It involves manual annotation for seed selection, followed by automatic annotation during synthetic data generation for multiple images associated with each chosen seed. The approach incorporates a systematic seed identification procedure, known as seed-ID, which utilizes prefixes in image names to create association between the seeds and the AU annotations of the synthetic images. Moreover, binary annotation is used to mark the presence or absence of specific AUs in each image, enabling comprehensive tracking of AU occurrences. The method efficiently maps prefixes to corresponding AUs and leverages seed tracking lists to annotate images accordingly.

By integrating this annotation method with synthetic data generation, it minimizes the need for labor-intensive manual efforts and enables streamlined data labeling. It allows for the generation of a significantly large synthetic dataset, facilitating the development and evaluation of robust models for AU prediction in a FER classification model. Most importantly, the proposed method exhibits a high level of practicality, as it does not impose the requirement of obtaining annotations simultaneously with the data acquisition process. This characteristic makes it highly compatible with the TDS pipeline.

5.1.3 Performance of the FER classification model

Analysing the preliminary results of the baseline model, it was observed that even when the number of synthetic images was increased fourfold compared to the number of real images, the overall performance of the model decreased for multiple classes of AUs, while experiencing only a slight improvement for one AU class. One possible explanation for this phenomenon is that the initial SCFD dataset of synthetic data, which has not undergone processing using the TDS pipeline, could have contained a set lower quality instances. It is plausible that the inclusion of lower-quality instances within the initial dataset adversely affected its overall quality and subsequently contributed to the decline in performance across various classes. the small size of the original CFD dataset could have posed challenges for the StyleGAN2-ADA network to effectively learn its distinctive features within the limited training time allocated. This limitation may have caused the presence of low-quality images within the initial synthetic dataset used on the baseline model.

As for the enhanced classification model, it has yielded improvements in performance for the majority of classes, which is 50 percents of the six classes of interest. In particular, the AUC values of AU1, AU4, and AU5, have increased compared to both the baseline and original models, indicating an enhancement in the classifier's ability to distinguish between positive and negative cases. In the case of AU12 emerges as an outlier performance-wise, exhibiting a significantly lower AUC value compared to both the baseline and original models, could depend on the less frequent occurrences of class AU12. Moreover, the interpretation of facial expressions is highly individual and can vary significantly among unprofessional annotators. Thus, it is plausible to consider that a less comprehensive standard may have been applied during the seed annotation for this specific class.

Nevertheless, these outcomes still indicates that the performance's advancements are achieved through the enhancement processes such as TDS pipeline and Semi-automated Annotation method. Although this research only studies the effectiveness of TDS pipeline and annotation for performance enhancement of a FER classification model, as they are deemed to show effectiveness, they can even be applied to other types of model that are in need or synthetic data due to scare, suitable data.

Furthermore, as highlighted in section 4.8, the enhanced model demonstrates a higher consistency in performance, as evidenced by the minimal deviation observed among the AUC values across different classes. This balance in performance indicates that the TDS pipeline method has contributed with a selection of uniformly distributed dataset of high quality synthetic images. This consequently results in a higher level of stability and reliability in the classification outcomes provided by the enhanced model. Despite the subtle differences in performance observed among individual classes, the overall trend showcases a more harmonized and balanced performance across the entire spectrum of classes. This uniformity in performance further enhances the credibility and robustness of the enhanced model in effectively addressing the classification tasks at hand.

5.2 Conclusion

Firstly, this thesis demonstrates a systematic approach to evaluate synthetic image quality using both reference and no-reference Image Quality Assessment (IQA) techniques. The TDS pipeline validates results across phases and consistently reveals differences between synthetic and original image quality. This aims to reinforce the reliability of the IQA methods. Overall, this study emphasizes the pipeline's effectiveness in evaluating and selecting high-quality synthetic images, contributing to the understanding of image quality assessment in synthetic face generation for FER model improvement.

Secondly, the exploration of Latent Space and the implementation of the Semi-Automated Annotation method have contributed to the improvement of the enhanced model's performance, surpassing both the baseline result. Despite encountering challenges in accurately predicting AUs that have lower occurrences in the synthetic dataset SCFD, the enhanced model still outperformed the baseline benchmark. However, there is potential for further enhancement by fine-tuning the FER classification model. Additionally, the utilization of synthetic data generation, such as StyleGAN2-ADA, has already demonstrated increased performance, particularly in capturing less frequent instances of AUs. These findings indicate promising avenues for future research and the continued improvement of AU prediction models.

Further, it was observed that the enhanced model with the integration of TDS pipeline and Semi-automated Annotation method exhibits a higher degree of consistency in performance across the compared to the baseline and original models. This indicates that the integration of TDS pipeline and Semi-automated Annotation method contributes to a more stable and reliable classification outcome. Moreover, this increased consistency in performance enhances the credibility and robustness of the enhanced model compared to the baseline model. Despite subtle variations in performance among individual classes, the overall trend highlights the model's ability to effectively address the classification tasks with a more balanced performance.

In conclusion, TDS has effectively evaluated both the image quality and performance of synthetic images, leading to improved results when incorporating them into the FER model. The relatively small qualitative improvement indicates room for further enhancements. However, the results are still significant as they demonstrate that modifying the model's input can lead to substantial performance improvements. Finally, as mentioned, the fact that we used a larger number of images and selecting them in a systematic way via IQA process had considerable positive effect on the final result.

5.2.1 Contributions

Having summarized the findings and their significations in the previous section, the research's contributions have been identified. Thus, this section will be dedicated to provide a comprehensive description of the main contributions archived in this research.

The primary contribution of this master thesis is the development of a novel pipeline that incorporates IQA methods. The approach taken in this work, known as the Train Data Selection (TDS) pipeline, aims to address the challenge of effectively assessing a large dataset comprising over 100,000 generated images. One of the key considerations in this process is striking the right balance between data quality and data quantity, which was lacking in the predecessor project and may have impacted the initial benchmark performance. Due to the prevalence of this issue, the approach was taken involves mitigating the scarcity of data by leveraging GAN-generated datasets while simultaneously addressing quantity-related challenges through Latent Space exploration and the TDS pipeline. These strategies are employed to tackle the inherent challenges and optimize the performance of the classification model in this study.

Another contribution of this research is the development and integration of a methodology for Semi-automated Human Annotation. This approach not only aims to reduce human intervention in the annotation process but also seeks to achieve a more systematic approach to manual annotation. Since the use of pre-trained models' predictions as annotations can potentially cause complications such as the performance of relying sorely on the performance of the pre-trained. This resulting annotation methodology aims to enhance the efficiency of the annotation process, ultimately improving the overall quality and accuracy of the conducted annotation.

5.3 Future Work

The subsequent section of this thesis project is dedicated to outlining potential avenues for future research. These subjects of inquiry serve to extend and build upon the findings and outcomes attained in the current thesis research, providing new possibilities for further exploration and advancement in related fields of synthetic data generation and applications within FER.

Firstly, an indexing tracking embedded in image names can also be integrated in seed-ID for future research. This can be used to annotate images to specific classification pools. A suggested approach is that image with certain AU and with certain index could be uniquely identified to a certain classification pool. This index tracking approach can hold potential for enhancing the accuracy and efficiency of the image classification, as it allows for further control over AU identification and categorization in synthetic images. Further, due to time constraints, the current project did not explore the inclusion of the index and degrees in image names when comprising the images. However, incorporating this information in future research can contribute to a further control and thereby accuracy in annotation of various AUs presences in synthetic data instances, thereby leading to further improvements in the classification performance.

Furthermore, while this thesis has made moderate progress in terms of performance, there is still untapped potential for further enhancing the performance of the TDS pipeline. Specifically, exploring the utilization of TDS as a method to evaluate the generative capabilities of the styleGAN2-ADA model holds promise for achieving improved classification results. Therefore, future researches can explore more refined methods to further enhance the capabilities of TDS pipeline in evaluating image quality and generative models performance. Furthermore, it is worth considering the possibility of incorporating specific thresholds for IQA metrics as additional quality criteria in future reseraches. Additionally, this approach offers flexibility by allowing the thresholds to be adjusted according to the desired level of image quality. By implementing these thresholds, the system can also ensure that only images meeting or exceeding the predefined quality standards are included in the synthetic dataset. This should serve the purpose of maintaining a high standard of image quality aligned with the specific requirements and objectives of future studies.

In addition, data annotation is the foundation of high-quality datasets and human annotation often provides the ideal control. However, this method can be a resource-intensive and potentially unreliable process, especially in this case, where it is conducted on a limited scale. Additionally, professional annotation services can greatly enhance the quality of data, but they require substantial investment in time and other resources. Moreover, the preliminary results of this study shed light on the challenges associated with controlling and validating predictions from pre-trained models when employed for annotation purposes. As said before, future research lies in investigating a hybrid annotation approach that combines the semi-automated annotation technique proposed in this thesis with the baseline annotation method utilizing classification model to predict the presence of AUs.

Moreover, the classification model consists of some augmentation techniques among other functions that it inherits from the previous projects. However, the augmentation strategy primarily focused on flipping images to introduce a few variations to the training process. Nevertheless, it is worth considering incorporating additional augmentation methods for images such as gamma noise, grayscale transformations, and other similar noise-based methods for even more advanced image augmentation and thus can potentially enhance the model's robustness. Notably, images quality in both synthetic dataset and real dataset can potentially be improved using the suggesting augmentation techniques, as the real dataset DISFA can also benifit from further image quality enhancement. Furthermore, these mentioned techniques can potentially bridge the existing quality gap between synthetic images and their original counterparts and thus can potentially contribute to more robust IQA and model training, and thereby enhance classification outcome.

In summary, these areas of research hold potential for advancing the field of image classification and validation. The suggested recommendations, including the integration of index tracking in seed identification, further exploration of TDS with thresholds for IQA metrics, development of hybrid annotation systems, augmentation refinement, incorporating other models such as regression model or other non-linear models, present exciting avenues for future investigations. Therefore, there are potential for delving into these aspects to contribute with improvement in accuracy and efficiency as well as reliable validation methodologies for the FER domain.

Bibliography

[1] A. De and A. Saha, "A comparative study on different approaches of real time human emotion recognition based on facial expression detection," *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 483–487, 2015.

[2] C. von Numers, *Facial Expression Recognition for Clinical Trial Self-recordings - Image-based Automated Emotion Prediction with Scarce Data*. Department of Computer Science, Engineering, Chalmers University of Technology, and University of Gothenburg, 2022.

[3] L. Arevalo, S. Platakidou, D. Enström, and L. Le Tran, "Facial expression recognition using deep neural networks," 2022.

[4] S. Guan and M. Loew, "A novel measure to evaluate generative adversarial networks based on direct analysis of generated images," *Neural Computing and Applications*, vol. 33, no. 20, 13921–13936, 2021. DOI: `https://doi.org/10.1007/s00521-021-06031-5`.

[5] J.-T. Liu, F.-Y. Wu, W.-J. Lu, and B.-L. Zhang, "Domain adaption for facial expression recognition," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019, pp. 1–6. DOI: `10.1109/ICMLC48188.2019.8949178`.

[6] K. Akhmetov, *Domain adaptation for facial expression classifier via domain discrimination and gradient reversal*, 2021. arXiv: `2106.01467 [cs.CV]`.

[7] A. De and A. Saha, "A comparative study on different approaches of real time human emotion recognition based on facial expression detection," in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015, pp. 483–487. DOI: `10.1109/ICACEA.2015.7164792`.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative Adversarial Nets*, 2014.

[9] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Van den Broeck, "A semantic loss function for deep learning with symbolic knowledge," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 5502–5511. [Online]. Available: `https://proceedings.mlr.press/v80/xu18h.html`.

[10] N. Kodali, J. D. Abernethy, J. Hays, and Z. Kira, "How to train your DRAGAN," *CoRR*, vol. abs/1705.07215, 2017. arXiv: `1705.07215`. [Online]. Available: `http://arxiv.org/abs/1705.07215`.

[11] T. Karras, S. Laine, and T. Aila, *A style-based generator architecture for generative adversarial networks*, 2019. arXiv: `1812.04948 [cs.NE]`.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *CoRR*, vol. abs/1812.04948, 2018. arXiv: 1812.04948. [Online]. Available: http://arxiv.org/abs/1812.04948.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, *Analyzing and improving the image quality of stylegan*, 2020. arXiv: 1912.04958 [cs.CV].

[14] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, *Training generative adversarial networks with limited data*, 2020. arXiv: 2006.06676 [cs.CV].

[15] G. P. Way, M. Zietz, V. Rubinetti, D. S. Himmelstein, and C. S. Greene, "Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations," *Genome Biology*, vol. 21, no. 1, 2020. DOI: https://doi.org/10.1186/s13059-020-02021-3.

[16] E. Trunz, M. Weinmann, S. Merzbach, and R. Klein, "Efficient structuring of the latent space for controllable data reconstruction and compression," *Graphics and Visual Computing*, vol. 7, 2022. DOI: https://doi.org/10.1016/j.gvc.2022.200059. [Online]. Available: http://graphics.tudelft.nl/Publications-new/2022/TWMK22.

[17] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, *Ganspace: Discovering interpretable gan controls*, 2020. arXiv: 2004.02546 [cs.CV].

[18] F. Leeb, S. Bauer, M. Besserve, and B. Schölkopf, *Exploring the latent space of autoencoders with interventional assays*, 2023. arXiv: 2106.16091 [cs.LG].

[19] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Computer Vision – ECCV 2018*, ser. Lecture notes in computer science, Cham: Springer International Publishing, 2018, pp. 556–572.

[20] 2022. [Online]. Available: https://github.com/NVlabs/stylegan2-ada.

[21] M. Woodland, J. Wood, B. M. Anderson, *et al.*, "Evaluating the performance of stylegan2-ada on medical images," *arXiv:2210.03786 [cs, eess]*, vol. 13570, 142–153, 2022. DOI: https://doi.org/10.1007/978-3-031-16980-9_14. [Online]. Available: https://arxiv.org/abs/2210.03786.

[22] A. Gavrilov, A. Jordache, M. Vasdani, and J. Deng, "Preventing model overfitting and underfitting in convolutional neural networks," *International Journal of Software Science and Computational Intelligence*, vol. 10, pp. 19–28, Oct. 2018. DOI: 10.4018/IJSSCI.2018100102.

[23] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, *Veegan: Reducing mode collapse in gans using implicit variational learning*, 2017. arXiv: 1705.07761 [stat.ML].

[24] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein gan*, 2017. arXiv: 1701.07875 [stat.ML].

[25] 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, 2010. DOI: 10.1109/TKDE.2009.191.

[26] F. Nielsen and F. Barbaresco, Eds., *Geometric Science of Information*. Springer International Publishing, 2021. DOI: 10.1007/978-3-030-80209-7. [Online]. Available: https://doi.org/10.1007%2F978-3-030-80209-7.

[27] S. Mo, M. Cho, and J. Shin, "Freeze the discriminator: A simple baseline for fine-tuning gans," 2020. DOI: 10.48550/ARXIV.2002.10964. [Online]. Available: https://arxiv.org/abs/2002.10964.

[28] M. Abadi, P. Barham, J. Chen, *et al.*, *Tensorflow: A system for large-scale machine learning*, 2016. arXiv: 1605.08695 [cs.DC].

[29] E. Rojas, A. N. Kahira, E. Meneses, L. B. Gomez, and R. M. Badia, *A study of checkpointing in large scale training of deep neural networks*, 2021. arXiv: 2012.00825 [cs.DC].

[30] V. Riccio and P. Tonella, *When and why test generators for deep learning produce invalid inputs: An empirical study*, 2022. arXiv: 2212.11368 [cs.SE].

[31] P. Zhang, W. Zhou, L. Wu, and H. Li, "Som: Semantic obviousness metric for image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[32] H. Cong, L. Fu, R. Zhang, *et al.*, *Image quality assessment with gradient siamese network*, 2022. arXiv: 2208.04081 [eess.IV].

[33] C. Hao, Z.-X. Yang, L. He, and W. Wu, "Texture synthesizability assessment via deep siamese-type network," en, *Secur. Commun. Netw.*, vol. 2022, pp. 1–11, Feb. 2022.

[34] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, *No-reference image quality assessment via transformers, relative ranking, and self-consistency*, 2022. arXiv: 2108.06858 [eess.IV].

[35] J. Ma, J. Wu, L. Li, W. Dong, and X. Xie, "Active inference of gan for no-reference image quality assessment," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6. DOI: 10.1109/ICME46284.2020.9102895.

[36] S. Dost, F. Saud, M. Shabbir, M. G. Khan, M. Shahid, and B. Lovstrom, "Reduced reference image and video quality assessments: Review of methods," en, *EURASIP J. Image Video Process.*, vol. 2022, no. 1, Jan. 2022.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training gans*, 2016. arXiv: 1606.03498 [cs.LG].

[38] [Online]. Available: https://www.image-net.org/index.php.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training gans*, 2016. arXiv: 1606.03498 [cs.LG].

[40] S. Barratt and R. Sharma, *A note on the inception score*, 2018. arXiv: 1801.01973 [stat.ML].

[41] S. Zhou, M. L. Gordon, R. Krishna, A. Narcomey, L. Fei-Fei, and M. S. Bernstein, *Hype: A benchmark for human eye perceptual evaluation of generative models*, 2019. arXiv: 1904.01121 [cs.CV].

[42] E. Denton, S. Chintala, A. Szlam, and R. Fergus, *Deep generative image models using a laplacian pyramid of adversarial networks*, 2015. arXiv: 1506.05751 [cs.CV].

[43] C. Mailis, *Npyviewer*, https://github.com/csmailis/NPYViewer, 2021.

[44] S. Menard, *Applied Logistic Regression Analysis.* SAGE Publications, Incorporated, 1995, 90–100.

[45] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning, second edition : data mining, inference, and prediction*, 2nd ed. New York: Springer, 2009, ISBN: 9780387848570.

[46] A. Liu, W. Lin, H. Chen, and P. Zhang, "Image retargeting quality assessment based on support vector regression," *Signal Processing: Image Communication*, vol. 39, pp. 444–456, 2015, Recent Advances in Vision Modeling for Image and Video Processing, ISSN: 0923-5965. DOI: https://doi.org/10.1016/j.image.2015.08.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596515001265.

[47] S. Lei, H. Zijian, Y. Jiebin, and F. Fengchang, "Super resolution image visual quality assessment based on feature optimization," *Computational Intelligence and Neuroscience*, vol. 2022, D. Zhang, Ed., 1–10, 2022. DOI: https://doi.org/10.1155/2022/1263348. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9236850/.

[48] J. Huang and C. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005. DOI: 10.1109/TKDE.2005.50.

[49] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," *MachineLearningMastery.com*, Aug. 2020. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/.

[50] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, Apr. 2011. DOI: 10.1007/s11222-009-9153-8.

[51] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137–1143, Aug. 1995. [Online]. Available: http://ijcai.org/Proceedings/95-2/Papers/016.pdf.

[52] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research*, vol. 11, no. 70, pp. 2079–2107, Mar. 2010. DOI: 10.5555/1756006.1859921. [Online]. Available: http://jmlr.org/papers/volume11/cawley10a/cawley10a.pdf.

[53] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, no. none, Jan. 2010. DOI: 10.1214/09-ss054. [Online]. Available: https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.pdf.

[54] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, 151–160, 2013. DOI: https://doi.org/10.1109/t-affc.2013.4.

[55] [Online]. Available: https://www.chicagofaces.org.

[56] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *CoRR*, vol. abs/1904.11150, 2019. arXiv: 1904.11150. [Online]. Available: http://arxiv.org/abs/1904.11150.

[57] D. McDuff, J. M. Girard, and R. e. Kaliouby, "Large-scale observational evidence of cross-cultural differences in facial behavior," Jun. 2018.

[58] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, pp. 151–160, 2013.

[59] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll, "Efficient recognition of authentic dynamic facial expressions on the feedtum database.," in *ICME*, IEEE Computer Society, 2006, pp. 493–496, ISBN: 1-4244-0367-7. [Online]. Available: `http://dblp.uni-trier.de/db/conf/icmcs/icme2006.html#WallhoffSHR06`.

[60] [Online]. Available: `https://docs.opencv.org/4.x/d1/dfb/intro.html`.

[61] P. Ekman and W. V. Friesen, "Facial action coding system," *PsycTESTS Dataset*, 1978. DOI: `https://doi.org/10.1037/t27734-000`.

[62] I. Stępień and M. Oszust, "A brief survey on no-reference image quality assessment methods for magnetic resonance images," *Journal of Imaging*, vol. 8, no. 6, 2022, ISSN: 2313-433X. DOI: `10.3390/jimaging8060160`. [Online]. Available: `https://www.mdpi.com/2313-433X/8/6/160`.

[63] S. Yang, T. Wu, S. Shi, *et al.*, *Maniqa: Multi-dimension attention network for no-reference image quality assessment*, 2022. arXiv: `2204.08958 [cs.CV]`.

A

Appendix 1

A.1 Data Aquasition Related Goals: DISFA

The DISFA dataset exhibits several similarities with the target task, making it a suitable choice. The elicitation process is naturalistic, and the recordings are predominantly frontal. Moreover, each frame of the dataset contains annotations for 12 Action Units (AUs). These 12 AUs are listed below:

Inner Brow Raiser	AU1
Outer Brow Raiser	AU2
Brow Lowerer	AU4
Upper Lid Raiser	AU5
Cheek Raiser and Lid Compressor	AU6
Nose Wrinkler	AU9
Lip Corner Puller	AU12
Lip Corner Depressor	AU15
Chin Raiser	AU17
Lip Stretcher	AU20
Lips Part	AU25
Jaw Drop	AU26

Table A.1: Action Units Corresponding to Specific Facial Expressions

The dataset includes four minutes of video footage at a rate of 20 frames per second for each subject. The demographic distribution of the subjects is satisfactory in terms of ethnicity, age, and gender, although the majority are white. Given the relatively good fit with the previously discussed requirements, DISFA was chosen over as the baseline for this thesis.

A.2 Support Vector Regression for Image Quality Prediction Using BRISQUE Metric

Algorithm 4 Image Processing, Data Split, and SVR Training

```
Require: Original images folder, Synthetic images folder, Test size ratio
  procedure IMAGEPROCESSINGDATASPLITSVR
      Load original images from folder
      for each original image do
          Preprocess the image and calculate BRISQUE score
          Convert image to grayscale
          Scale image to 8-bit unsigned integer
          Calculate gray-level co-occurrence matrix (GLCM)
          Compute contrast, correlation, energy, and homogeneity from GLCM
          Calculate BRISQUE score as the average of contrast, correlation, energy,
  and homogeneity
      end for
      Prepare data for SVR training
      Split the dataset into training and validation sets using train_test_split
      Train SVR model
      Load synthetic images from folder
      for each synthetic image do
          Preprocess the image and calculate BRISQUE score
          Convert image to grayscale
          Scale image to 8-bit unsigned integer
          Calculate gray-level co-occurrence matrix (GLCM)
          Compute contrast, correlation, energy, and homogeneity from GLCM
          Calculate BRISQUE score as the average of contrast, correlation, energy,
  and homogeneity
      end for
      Predict scores for synthetic images using SVR model
      Print predicted BRISQUE scores for synthetic images:
      for each synthetic image do
          Print "Image" and image index
          Print "Score:" and predicted score for the image
      end for
      Evaluate model performance on the validation set
  end procedure=0
```

A.3 NumPy File Viewer as a Tool

The NPYViewer [43] is a pre-existing npy-file viewer written in Python using the PyQt5 library. It allows the user to browse and view multiple npy files at once, as well as display metadata about the selected file. It has a simple GUI which offers a viewer as 3D point cloud, grayscale images, as heightmaps and various timeseries visualizations as well as adjacency matrices (weighted graphs etc). Other current features supported are:

- View and manipulate 2D NumPy arrays and lists saved in .npy files as spreadsheets
- Convert .npy files to .csv format and vice versa
- Export .npy files as .mat files for use in MATLAB and Octave
- Display 2D NumPy arrays as grayscale images or 3D point clouds for arrays with 3D coordinates
- Visualize 2D NumPy arrays as heightmaps and 1D NumPy arrays as timeseries data
- Load .npy files as command line arguments with support for visualization of adjacency matrices as directed edge weighted graphs
- Print NumPy arrays in the terminal using the -noGUI argument
- The program uses PyQT5 to create a user-friendly graphical user interface (GUI)

The NPYViewer tool is primarily used to validate the annotation process and make sure that the npy-file has been created correctly. That includes checking that the size of the npy-file dimension is correct and that the overall annotation looks equally distributed; since one already has control over the generation of images.

A.4 Thesis Project Roadmap

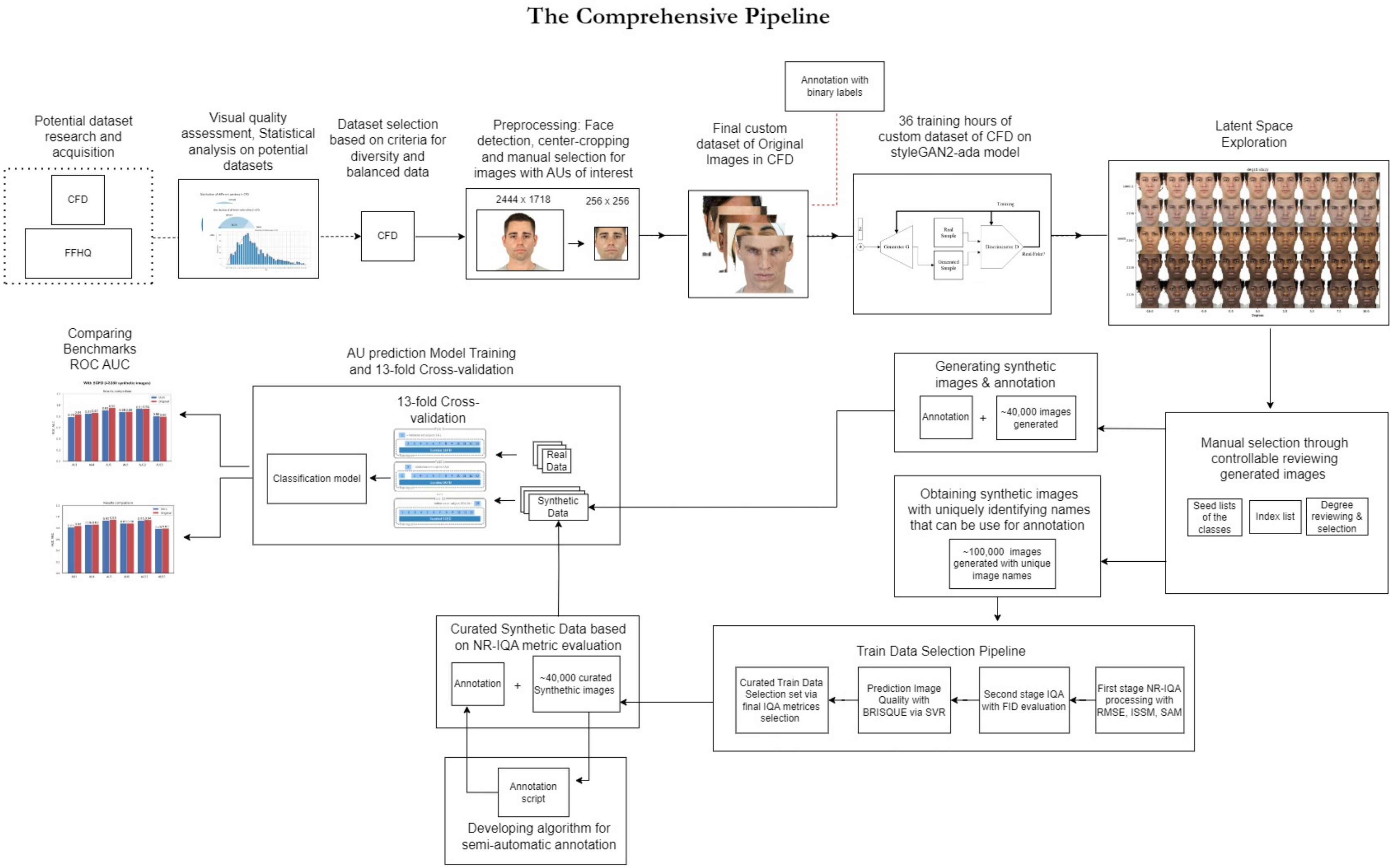

The Comprehensive Pipeline
Potential dataset research and acquisition
CFD
FFHQ
Visual quality assessment, Statistical analysis on potential datasets
Dataset selection based on criteria for diversity and balanced data
CFD
Preprocessing: Face detection, center-cropping and manual selection for images with AUs of interest
2444 x 1718
256 x 256
Annotation with binary labels
Final custom dataset of Original Images in CFD
36 training hours of custom dataset of CFD on styleGAN2-ada model
Training
Generator G
Real Sample
Generated Sample
Discriminator D
Real/Fake?
Latent Space Exploration
Comparing Benchmarks ROC AUC
AU prediction Model Training and 13-fold Cross-validation
13-fold Cross-validation
Classification model
Real Data
Synthetic Data
Generating synthetic images & annotation
Annotation
+
~40,000 images generated
Manual selection through controllable reviewing generated images
Seed lists of the classes
Index list
Degree reviewing & selection
Obtaining synthetic images with uniquely identifying names that can be use for annotation
~100,000 images generated with unique image names
Curated Synthetic Data based on NR-IQA metric evaluation
Annotation
+
~40,000 curated Synthethic images
Train Data Selection Pipeline
Curated Train Data Selection set via final IQA metrices selection
Prediction Image Quality with BRISQUE via SVR
Second stage IQA with FID evaluation
First stage NR-IQA processing with RMSE, ISSM, SAM
Annotation script
Developing algorithm for semi-automatic annotation

Figure A.1: Thesis Project Roadmap.