School of Computer Science
Faculty of Science
University of Nottingham

Multimodal Analysis of Depression in Unconstrained Environments

Keerthy Kusumam

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science of The University of Nottingham. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

October 2023

Acknowledgements

I am deeply grateful for the opportunity to pursue my PhD degree which has been an incredible learning experience that has challenged me intellectually, personally, and professionally.

I would like to express my sincere appreciation to my advisors who played significant roles in shaping my research ideas and methodology. Dr. Georgios Tzimiropoulos, provided invaluable expertise and support during the PhD. Prof. Michel Valstar, offered invaluable insights and support during the PhD. Prof. Elvira Perez Vallejos provided tremendous support and exceptional kindness all through out the PhD and was instrumental to my thesis completion. I thank all of my supervisors for their support, valuable insights, and their intellectual guidance have been crucial in enabling me to complete this thesis successfully.

I would like to express my sincere thanks to my brilliant colleagues in AstraZeneca especially my proxy-supervisor Dan, without whose immense kindness, understanding and encouragement I would not have been able to balance the PhD thesis writing with work. I thank Dr. Tomas Krajnik and Prof. Tom Duckett who inspired my research thinking and ideology which carved my identity as a researcher. I would like to express my gratitude to my colleagues and friends, Jing, Aaron, Joe, Siyang, Dimitris, Joy, Shashank, Gustavo, Marie, Simon, Johann, Joao, Claudio, Christian, Tryphon and George for their support, encouragement, and intellectual exchange during my PhD program and my time in Lincoln.

I am deeply grateful to my family, my mother especially, for the unwavering support, love, and encouragement during my PhD. Her radical approach to life and work, as a self-made lawyer and the unquestioned trust in my choices have been the source of my motivation. I am eternally grateful to my lifelong mentor, teacher and uncle Kamaneesh, whose guidance and astute view of the world instilled in me the courage and inspiration to create my own path since childhood. I am grateful to my loving sisters, Poonam, Aswathy and Anutty, my dear friends Rakhi, Rita, Anders, Tanya, Hannah, Andrew, James, Lucia for always being there, whenever I was in need.

Finally, I would like to thank my partner Anestis for everything - his calming presence, exceptional love for his bees and nature, and his acceptance and constant belief in me have been my steadfast anchor in all turbulence.

Abstract

Mental health disorders, such as depression and anxiety, are a significant global problem affecting millions of people, leading to disability, increased mortality from suicide, and reduced quality of life. Traditional diagnostic and evaluation methods rely on subjective approaches and are limited by resource availability, driving the need for more accessible and efficient methods using technology. Digital mental health, a rapidly growing field, merges digital technologies into mental health care, utilizing the Internet and mobile phone software to deliver mental health services. The use of mobile health technologies, such as Ecological Momentary Assessments and digital phenotyping, can improve depression diagnostics by generating objectively measurable markers in natural environments. Technological progress in computer vision, natural language processing, and affective computing has also led to the emergence of automated behavior analysis methods, improving depression assessment and understanding.

This thesis addresses the problem of mood assessment and analysis for detecting depression from multimodal data in unconstrained, natural environments. This thesis presents a novel, multi-modal dataset collected from a purpose- built smartphone app for depression recognition in real-world, unconstrained environments and proposes a state-of-the-art, automated depression recognition system leveraging advancements in multimodal analysis. The research outcomes have the potential to be applied in automated patient monitoring or therapy administering platforms. The thesis contributes by: 1) collecting a novel, longitudinal, and multi-modal, Mood-Seasons dataset in real-world settings, 2) benchmarking state-of-the-art video analysis techniques on newly collected and publicly available datasets, 3) building a multimodal spatio-temporal transformer model for automated depression severity prediction, 4) presenting a new framework for face generation that learns to synthesize novel face images that adhere to a given pose and appearance from exemplar image in a semantically meaningful way and 5) applying the face manipulation method for anonymizing the Mood-Seasons dataset for privacy preservation.

In conclusion, this thesis addresses the limitations of current depression diagnostics and assessments by integrating smartphone-driven digital phenotyping technologies to advance and personalize depression care. By collecting a novel dataset, proposing state-of-the-art methods for depression recognition, and addressing privacy concerns, this work has the potential to significantly improve mental health care delivery and accessibility.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Mental health disorders are a significant global problem, with more than 300 million people affected and 800,000 suicides occurring each year (WHO, 2020). Behavioural or mood disorders such as depression, anxiety, etc. continues to be the main drivers of disability leading to significant morbidity, increased mortality from high suicide risk, decreased functioning, and poor quality of life. The Survey of Mental Health and Wellbeing in England (McManus et al., 2016) found that 1 in 6 people aged 16+ had experienced symptoms of a mental health problem, such as depression or anxiety, in the last week, but only one in three of these individuals sought treatment for these conditions over a 12-month period.

Depression is characterised by distinct observable behavioural symptoms associated with general psychomotor functioning, emotional expression, and interpersonal interactions. Prompt diagnosis and timely personalised intervention are crucial for people suffering from mood disorders such as depression to receive maximum benefit from treatment. Unfortunately, current diagnosis and evaluation of depression rely on subjective approaches such as self-reporting and clinical interviews, which are subject to limitations related to their dependence on explicit definitions and reliable evaluation. Furthermore, there is a lack of resources, both in developed and developing countries, to provide mental health services in person. The increasing numbers of affected individuals have led to long waiting times for mental health screening and treatment delivery, highlighting the need for more accessible and efficient methods for depression assessment and treatment. Therefore, governments and health facilities are exploring using technology, such as automatic diagnosis or monitoring, for increased accessibility and improved treatment processes.

Digital mental health is a rapidly growing area of research that merges advances in digital technologies with mental health care such as e-Health (Parikh and Huniewicz, 2015) and mHealth (Ameringen et al., 2017), (Price et al., 2014) using the Internet and mobile phone software to provide mental health services. Mental health care currently benefits from the use of mobile phone or Internet-based software in various clinical care stages, such as symptom assessment, patient engagement, and psychoeducation, tracking, and monitoring treatment progress (Luxton et al., 2011). Many evidence-based applications or technologies are currently available to deliver assessment, monitoring, and interventions in mental health such as schizophrenia, substance abuse, eating disorder, sleeping disorder, mood disorders such as bipolar, anxiety, and depressive disorders (Bakker et al., 2016).Apps such as Mobilyze, Purple robot (Ameringen et al., 2017) offer a way for patients to track their daily mood, by completing questionnaires and provide them with positive affective feedback when necessary.

Digital technologies present a promising opportunity to individualise depression care. With the advanced capabilities of digital sensors and computing on smartphones, they can function as "human sensors" to monitor minute changes in behavioural patterns. Electronic medical records can collect extensive data from various medical fields, produce custom reports, and transmit data between healthcare systems without disruption. Telepsychiatry can facilitate real-time interactions with patients in their natural environment.

Mental health technologies use different mechanisms for operation, including self-assessments of mood (Kauer et al., 2012). Ecological Momentary Assessments (EMA) is a method of sampling user experiences in real time and using it to provide better treatment. Digital phenotyping (DP) through EMA, especially using assessments conducted using mobile health technologies, has the potential to greatly improve the accuracy of depression diagnostics by generating objectively measurable markers, analysed in natural environments. It can promote user engagement (Sloan et al., 2011) and help deliver evidence-based treatment such as Cognitive Behavioural Therapy (CBT) (Spek et al., 2007).

Technological progress in computer vision, natural language processing, audio and

multimodal analysis, and affective computing has led to the emergence of automated behaviour analysis methods, which could offer substantial improvements in depression assessment and comprehension. By employing these automated methods, researchers can gain new insights into behavioural indicators of depression and develop more reliable screening and diagnostic tools. Additionally, these methods can be used to measure the effectiveness of interventions and test clinical theories about the underlying mechanisms of depression. Despite the challenges that remain, the development and use of automated methods for behaviour analysis represent an exciting and promising direction for the field of clinical psychology.

Preliminary investigations have also shown that linguistic and behavioural clues from social media data and data extracted from electronic medical records can be used to predict depression status. Burns et al.(Burns et al., 2011) provide ecological momentary interventions for depression by using mobile sensors, including GPS for location data, accelerometers for movement detection, phone microphones for voice data from calls, phone call history, and activity, to analyse behavioural patterns and predict user's mood. In similar studies such as (Saeb et al., 2015), the authors use contextual data and passive monitoring using GPS to predict depressive symptoms, and in (Depp et al., 2010) the authors design interventions for bipolar disorder using self-reported real-time mood of users.

The conversational agent, Woe bot, presented by Fitzpatrick et al. (2017) delivers CBT with daily mood tracking and prompts active user engagement through natural language processing. Studies use contextual information to assess the mood or emotional state of the patient and correlate it with self-reported mood scores, such as PHQ-9 (Sajatovic et al., 2015). The PHQ-9 questionnaire is a well-adopted metric for reviewing depression severity for screening, diagnosing, and monitoring patients. It is used in primary care and clinical use to assess the state of depression during the preceding two weeks. Mental health technologies use apps like DepressionMonitor (Ameringen et al., 2017) to provide PHQ-9 or PHQ-8 (excluding the suicide symptom for ethical reasons) for clinical use. Many digital interventions serve as a tool to record mood data to deliver therapy or monitor progress based on the longitudinal behavioural pattern of patients during or post treatment.

Current depression diagnostics and assessment have significant limitations due to

heterogeneity of clinical presentations, lack of objective assessments, and assessments based on patients' perceptions, memory, and recall. Integration and application of smart phone-driven digital phenotyping technologies have the potential to significantly advance and personalise depression care.

This thesis introduces a novel multimodal dataset collected from a purpose-built smartphone app for the recognition of depression in unconstrained real-world environments. The data set includes longitudinal data over three weeks. The thesis proposes a state-of-the-art automated depression recognition system that takes advantage of the latest advances in multimodal analysis. The thesis also proposes a novel approach to address privacy concerns in the data set using generative methods to anonymise face images.

1.1 Motivation

Observable traits of depression as stated by Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (APA) (Association et al., 2013) include both visual indicators (facial expression and demeanor) and speech (increased pauses, muteness). Facial expressions along with speech are prominent behavioural observations that are strong indicators of mood disorders (Girard and Cohn, 2015a), (Hollis et al., 2015), (Ringeval et al., 2017a) including depression. In the past, many approaches have used different signals to detect depression and other mood disorders, such as facial expression (Ringeval et al., 2017a), (Jan et al., 2014), gaze (Alghowinem et al., 2013a), head movement (Joshi et al., 2013b), body pose (Joshi et al., 2013a). Many studies have shown that reduced expressibility, eye contact, eyelid activity, iris movement, smile intensity, smile duration, lack of smile, listening smiles (when not speaking) are common traits in people diagnosed with depression (Pampouchidou et al., 2017b). Voice and speech analysis are used as reliable means of estimating and tracking mood disorders (Cummins et al., 2015a), (Faurholt-Jepsen et al., 2016), with studies having established accuracies through clinical trials. Depressive markers in speech include vocal prosody, ie, pause duration and vocal frequency (loudness) and were used to detect depression in (Cohn et al., 2009a).

The audiovisual emotion recognition challenge AVEC (Ringeval et al., 2017a) or-

ganised a series of depression challenges and provided the contestants with clinical interviews of people diagnosed with and without depression using self-reported depression questionnaire scores to assess the severity of depression. Studies such as in (Ringeval et al., 2017a) show a strong correlation between detected facial expressions for affective states and depression. The common pipeline for automated facial and speech analysis, as depicted in (Girard and Cohn, 2015b) starts with image preprocessing, registration, feature extraction, and learning models for classification and prediction of depression.

The analysis of multiple modalities can provide even better results in the detection of symptoms of depressive or mood disorders, as shown by (Ringeval et al., 2017a), (Pampouchidou et al., 2017b). Different methods use multiple modalities, such as in (Dibeklioğlu et al., 2015) where the authors combine facial, postural, and vocal measures to detect depression.

An accurate characterisation of facial behaviour that can assess mood in real-time can be used as a reliable sensor in mental health technologies for managing mood disorders. This would open more opportunities to deliver behavioural interventions based on multimodal, audio, vision, and language data, prompting seamless user engagement during video sessions. The focus of this PhD is on the problem of mood assessment and analysis for detecting mood disorders like depression from multimodal data containing face, voice, and spoken words in unconstrained, natural environments. The results of the research can be applied successfully to deliver mental health care on automated patient monitoring or therapy administration platforms.

Many published studies addressing the problem of mood analysis for mental health disorders point to the difficulty of obtaining labelled data on a large scale (Pampouchidou et al., 2017b) mainly attributed to the clinical expertise needed and its sensitive nature. This makes most studies resort to collecting their own data sets in the laboratory, where most of the available data are in the form of clinical interviews with a limited number of subjects and in restricted clinical settings. It limits the detection of mood disorders using facial behaviour in previously unseen environments or in-the-wild. To develop a system that has real-world impact, it should be able to assess depression in natural environments. This PhD thesis addresses the problem of the lack of real-world data necessary for such a system that learns to recognise

depression by collecting a large, longitudinal and multimodal dataset collected using a smartphone app.

The state-of-the-art methods in computer vision problems, such as image recognition and detection, are based on deep learning. One of the main advantages of deep learning-based methods are the models that can learn representations from data, without the need for defining handcrafted features. Deep learning-based methods have shown to achieve high accuracies in tasks such as object recognition (Krizhevsky et al., 2012), human pose estimation (Newell et al., 2016a), face alignment (Bulat and Tzimiropoulos, 2016), semantic segmentation (Girshick et al., 2014) etc. Several neural network architectures such as Convolutional neural networks (CNN), Recurrent neural networks (RNN), long short term memory networks (LSTM), have been used successfully in tasks like facial expression recognition (Yu and Zhang, 2015), (Kim et al., 2015), and emotion recognition from audiovisual data (Ng et al., 2015).

The thesis will analyse the real world data set using state-of-the-art and novel techniques to characterise the severity of depression. The thesis will focus on providing a comprehensive benchmark of state-of-the-art video analysis techniques on newly collected and publicly available datasets. The thesis will build a framework using cutting-edge multimodal transformers for automated depression severity prediction and provide quantitative evaluation of the effectiveness of the approach on different datasets.

Generating face images under varying poses and expressions is a rapidly growing area of research in computer vision. This is because face images are a rich source of information about a person's identity, emotions, and state of mind. Generative Adversarial Networks (GANs) are a class of deep learning models that are particularly well-suited for generating realistic synthetic data. GANs work by training two competing networks: a generator network and a discriminator network. The generator network tries to create fake data that is indistinguishable from real data, while the discriminator network tries to distinguish between fake and real data.

Advances in GAN-based techniques for face manipulation offer promising opportunities for automated depression analysis. For example, GANs can be used to (i) develop privacy-preserving models by anonymising the identity of face images while

preserving pose and expression for the downstream depression analysis task, without compromising the privacy of real patients, and (ii) create data augmentation by generating additional synthetic face images to augment existing datasets, particularly helpful in cases where datasets are small or imbalanced. Augmenting data sets with synthetic data can help to improve the performance of depression analysis models by making them more robust to different variations in face images (iii) evaluate model performance in edge cases by testing developed depression models in synthetic face images of people with specific edge cases, such as different poses and expressions, gender or race that are not available in the dataset.

Preservation of privacy is a significant issue in the storage and dissemination of depression recognition data sets. Data sets for automated analysis of depression can involve sensitive personal information, especially facial and voice data that reveal one's identity. Therefore, privacy preservation is crucial to ensure that private information from individuals is not shared or analysed without their consent, misused, or discriminated against. Privacy preservation in these datasets can ensure (i) protection of anonymisation rights of the participant, (ii) prevent potential discrimination against age, sex, race, or mental health condition, and (iii) maintain trust between the participants and research authorities or healthcare systems by complying with regulations.

This thesis introduces a novel method for manipulating facial images using GANs that alters their pose and facial expressions with or without preserving identity. The proposed approach for face generation synthesises novel face images that conform to a given pose while transferring appearance and style information from an exemplar image. The thesis proposes an application of this method to anonymise the identities present in the Mood-Seasons dataset and discusses its potential future applications for automatic depression analysis.

1.2 Contributions

This thesis proposes the following contributions to the field of automated depression analysis.

- Address the problem of the scarcity of real-world labelled data necessary for a system that learns to recognize depression by collecting a novel video dataset

that captures longitudinal participant behaviors with markers of depression in unconstrained, in-the-wild environments. Alongside the data set, the research community is using a bespoke cross-platform smartphone application that facilitates data collection. This also includes a comprehensive description of the ethics approval process and the ethical guardrails that were put in place before collecting private and sensitive personal mental health data.

- Provide an extensive benchmark of state-of-the-art video analysis techniques on the newly curated Mood-Seasons and publicly available AVEC 2014 datasets.
- Propose a multi-modal spatio-temporal transformer model that fuses the behavioural cues such as facial appearance, voice, spoken words (audio and language) from the videos that are relevant for estimating an individual's depression severity levels reliably in natural environments. The novel differential loss was introduced to improve performance by leveraging multiple videos from one person.
- Present a new framework for face generation that learns to synthesise novel face images that adhere to a given pose, whilst transferring appearance and style information from an exemplar image in a semantically meaningful way.
- Show the viability of applying the face manipulation method for anonymising the identities present in the Mood-Seasons dataset as a proof-of-concept, and laying out a roadmap on how this can be further leveraged for automatic depression analysis.

1.3 Publications

The following is the list of articles published during the PhD

1. Kusumam, K., Sanchez, E., and Tzimiropoulos, G. (2021). Unsupervised face manipulation via hallucination. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2406–2413. IEEE
2. Jaiswal, S., Valstar, M., Kusumam, K., and Greenhalgh, C. (2019b). Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 81–87

3. Haddon-Hill, G., Kusumam, K., and Valstar, M. (2021). A simple baseline for evaluating expression transfer and anonymisation in video transfer. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–08. IEEE

4. Xu, J., Song, S., Kusumam, K., Gunes, H., and Valstar, M. (2021). Two-stage temporal modelling framework for video-based depression recognition using graph representation. *arXiv preprint arXiv:2111.15266*

Chapters 4 and 5 of this thesis will be submitted for publication to reputable affective computing journal in the near future.

1.4 Outline of the Thesis

The thesis structure comprises seven chapters. The thesis commences by defining mood disorders from a clinical perspective, with an emphasis on depression in both clinical and general populations.By zooming in on the prevalence of depression in the general population, the thesis examines the increasing reliance on technology in assistive diagnostics and treatment delivery. The thesis then provides an overview of the current art in digital mental health, approaches, and deployment. It further channels its discussion around the role of automated behaviour analysis using different data modalities, such as audio, images, and language, in delivering reliable approaches that help address existing issues. Then it extensively reviews research works from past and present that address automated depression analysis, identifies recurring themes, and discusses common issues and pitfalls in the field.

The third chapter provides a background on Generative Adversarial Networks, which is a crucial technological paradigm used in the thesis methodology. The background chapter provides an in-depth discussion of generative adversarial networks and techniques using synthetic data.

Chapters 4, 5 and 6 are the main contributions of the thesis. Chapter 4 comprehensively describes a large-scale data collection study that provided a novel, multimodal (audio-visual-text) and longitudinal Mood-Seasons dataset, collected in natural, in-the-wild conditions from a smartphone. The dataset comprises video, audio, and textual transcriptions from the general population, with the depression severity

recorded from their responses to a PHQ-8 questionnaire. The chapter describes the meticulous and ethical design of the data collection methodology, the development of the app, and its deployment. It also reviews the lessons from deploying such an app and the general population's perception of it. It describes how participation and adoption of smartphone-based mental health data collection methods can be maximised.

Chapter 5 describes several approaches, employing multimodal data to provide a benchmark for automated depression analysis on the Mood-Seasons dataset, which was introduced in chapter 4. The chapter provides an extensive evaluation of state-of-the-art unimodal and multimodal approaches for video recognition applied for the task of depression severity prediction. The chapter includes our novel approach to depression recognition using multimodal transformers that employs audio-visual-language fusion to learn depression severity scores from videos, at different temporal granularity. The chapter includes detailed experimental evaluation and verification of the efficacy of the methods presented in two datasets, the Mood-Seasons and the AVEC datasets.

In Chapter 6, the thesis presents a novel method for manipulating face images using conditional generative adversarial networks and its application for data anonymisation to address privacy preservation in sensitive datasets such as those used in automated depression recognition.

The thesis finally concludes in Chapter 7, summarising the main chapters and giving directions for future research.

Chapter 2

Depression Recognition: Assessment, Data and Methods

This chapter offers an overview of depression recognition, with an emphasis on contemporary techniques for diagnosing and assessing depression, data gathering approaches, and state-of-the-art automated depression recognition methods. The focus is on providing an understanding of the current landscape of depression estimation and monitoring, as well as highlighting advances in technology and methodologies that shape the future of depression care.

2.1 Depression Assessment

Major Depressive Disorder (MDD) is a complex condition characterised by a diverse array of potential behavioural markers and etiological factors (Zimmerman et al., 2015). Research shows that depression stems from the interactions between genetic and environmental factors that alter physiological systems (Organization et al., 2017).

The diagnostic approach involves determining symptomatic thresholds, assessing patient distress, evaluating functional impairments, and eliminating other potential causes (Association et al., 2013; Kamath et al., 2022). It is also crucial to eliminate the possibility of other factors, such as psychiatric, substance use, and medical disorders (Association et al., 2013).
The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (5th edition; [DSM-5]) and the World Health Organisation's International

Statistical Classification of Diseases and Related Health Problems (11th edition; [ICD-11]) define depression by two core symptoms - persistent low mood and reduced interest - lasting at least two weeks with at least 4 additional symptoms (Association et al., 2013; Kamath et al., 2022). Symptoms are shown in Table 2.1. The

Summary of Major Depressive Disorder Criteria
Five (or more) of the following symptoms are present for a period of at least 2 weeks: 1. Depressed mood 2. Anhedonia i.e., diminished interest or pleasure 3. Weight loss or weight gain 4. Sleep disturbances (insomnia or hypersomnia) 5. Psychomotor agitation or retardation 6. Fatigue 7. Feelings of worthlessness or excessive inappropriate guilt 8. Cognitive difficulties 9. Suicidal thoughts and/or behaviours
Other Criteria: Symptoms cause clinically significant distress or functional impairment. Symptoms are not better explained by other psychiatric or medical diagnosis.

Table 2.1: Major Depressive Disorder Criteria

presence of significant distress and impairment in daily functioning are also required (Association et al., 2013; Organization et al., 2017). Symptoms fall into psychological, neurovegetative, and neurocognitive categories (Kendler, 2016). Psychological symptoms are predominantly subjective as they depend on the patient's personal experiences, leading to alterations in behaviour. On the contrary, neuro-vegetative and neurocognitive symptoms present more objectively and are associated with measurable behavioural consequences that influence functioning (Kamath et al., 2022). Patient accounts of subjective symptoms are based on their unique experiences and interpretations. Understanding the distinctions between subjective and objective symptoms, as well as their expressions in voluntary or involuntary behaviour, is crucial to advance digital biomarker identification in the field of depression diagnostics.

2.1.1 Depression Questionnaires

Depression questionnaires, both self-rated and clinically rated, are widely used to evaluate and diagnose Major Depressive Disorder (MDD) (Lakkis and Mah-

massani, 2015) . Popular self-rated instruments include the 9-item Patient Health Questionnaire (PHQ-9), Beck Depression Inventory (BDI), 16-item Quick Inventory of Depression Symptomatology Self-rated (QIDS16-SR), and the Depression Scale of the Centre for Epidemiological Studies (CES-D) (Lakkis and Mahmassani, 2015). In real-world settings, self-rated tools are often preferred due to ease of administration and lower resource demands (Maurer et al., 2018). These instruments play a crucial role in the continuum of depression care and contribute to personalised patient care.

Two common assessment tools are HAMD and BDI (Baer and Blais, 2010), which differ in terms of administration time, focus, and measures. HAMD requires a 20-30 minute interview with a clinician, focussing on neurovegetative symptoms, while BDI is a 5-10 minute self-reported questionnaire underlining negative self-evaluation symptoms. Both tools have demonstrated consistency in distinguishing depressed from nondepressed patients (Baer and Blais, 2010). However, HAMD has been criticised for neglecting some typical depression symptoms (Baer and Blais, 2010; Gibbons et al., 1993).

Self-report scales and inventories (SRSIs) such as the BDI / BDI-II, PHQ-2 / 8/9 and the Depression and Somatic Symptom Scale (DSSS) are also used for the assessment of depression. Despite their widespread use and high specificity and sensitivity (80%-90%), SRSIs have several drawbacks, including not considering the clinical meaning of observed symptoms, allowing individual variability in reporting, and susceptibility to reporting bias (Pichot, 1986; Williams et al., 2005). However, SRSIs are widely adopted in primary health care and research, emphasising their cost effectiveness. There is no consensus on which tool, PHQ-9 or BDI-II, is more effective. Both have been found to have adequate reliability, convergent/discriminant validity, and responsiveness to change Titov et al. (2011).

2.1.2 Limitations of Current Assessments

However, current diagnostic and assessment approaches have limitations. DSM-based depression diagnosis relies on subjective factors, such as patient reports, clinical observations, and clinical judgement (Kamath et al., 2022). Time constraints and high variability in symptomatic presentations with multiple comorbidities are major limitations Tom et al. (2014).

Although depression rating scales can introduce some objectivity to clinical assessments, their use is limited due to resource constraints and their reliance on patient memory, which may only capture a narrow aspect of a patient's general mental state (Hong et al., 2021). They may also not fully capture the neurological or functional impacts of depression (Robinson et al., 2017).

The DSM-5 highlights various observable audiovisual indicators of depression, such as crying, facial expressions, psychomotor agitation, and psychomotor retardation (Association et al., 2013). Automated measurement techniques are increasingly being employed to operationalise these behaviours, pinpoint those that reliably signal depression, and determine the distribution of typical and atypical behaviours. Statistical approaches to depression analysis have been useful in this context, as they help identify differences in specific behaviours between groups.

A variety of potential depression indicators have been discovered through recent studies, with some examples in the visual domain being smaller distances between eyelids, shorter blink durations (Alghowinem et al., 2013b), slower and less frequent head movements (Alghowinem et al., 2013c; Girard et al., 2014), longer periods of looking down (Scherer et al., 2014), and reduced smiling (Girard et al., 2014), gaze directions, listening smiles, self-adaptors, fidgeting behaviours, and foot tapping or shaking behaviours [Waxer, 1974; Hall et al., 1995]. Acoustic indicators include increased voice tension (Scherer et al., 2014), decreased coordination among formant frequencies and cepstral channels (Williamson et al., 2013), and longer and more variable switching pauses Yang et al. (2012b).

2.1.3 Digital Phenotyping

The term "digital phenotyping" has been used to refer to a range of concepts related to the analysis of digital data to measure behavioural and psychological traits. Some key definitions are (i) Objective measurement of human behaviour and experience using personal digital data streams from smartphones and wearables (Onnela and Rauch, 2016) (ii) Moment-by-moment quantification of the individual-level human phenotype using data from personal digital devices (Insel, 2017). (iii) passively collected digital breadcrumbs that can signal mental health states based on interactions

with technology (Torous et al., 2018). While terminology varies, digital phenotyping generally involves collecting and analysing different types of digital data, such as phone usage, typing patterns, mobility patterns, social interaction, and sleep data, to measure behavioural and cognitive markers relevant to mental health. The core premise is that passively collected digital data can serve as sensitive biomarkers to complement or predict outcomes from traditional clinical assessments.

In the context of this thesis, digital phenotyping refers specifically to the analysis of audio-visual and language data in the form of free-form videos of participants in order to extract and integrate relevant behavioural markers that may be indicative of severe depression. More research is still needed to determine the reliability and validity of different digital phenotyping approaches for the monitoring of depression. DP has the potential to significantly improve the accuracy of depression diagnosis and assessment by adding objectivity to the process, delivering digital behavioural biomarkers for personalised treatment. The generated phenotypes provide an ecological and continuous representation of the physical, emotional, behavioural, social, and cognitive activities of a patient in real time (Huckvale et al., 2019).

2.1.4 Ecological Momentary Assessments

DP relies on data from Ecological Momentary Assessments (EMA), which are conducted using personal digital devices such as smartphones, wearable sensors, and data collected from human-computer interactions (Colombo et al., 2019). EMA can be classified into active and passive categories (Dogan et al., 2017). Active EMA involves data reported directly by the user, such as electronic assessments using depression questionnaires such as Patient Health Questionnaire (PHQ-9), Hamilton Depression Rating Scale (HDRS), Quick Inventory of Depressive Symptomatology (QIDS), and Beck Depression Inventory (BDI) (Colombo et al., 2019). Passive EMA consists of data automatically collected from digital devices and platforms without the user's active input, such as phone usage, GPS, and sensor data (Dogan et al., 2017).

EMA aims to minimise recall bias, maximise ecological validity, and investigate processes that influence behaviour in real-world settings (Shiffman et al., 2008). Active EMA reduces recall bias and allows clinicians to gain insight into the situational and social context of patients (Kim et al., 2020). It can also help patients recognise

mood patterns, triggers, and coping strategies, as well as monitor suicidal ideation Wichers et al. (2011). Audio samples and language analyses can be used to assess mood disorders in active EMA (Gratch et al., 2021).

Passive EMA employs passive sensing using smartphones and wearables to capture multiple facets of human behaviour (Dogan et al., 2017). It is especially useful for capturing symptoms such as fatigue, sleep, focus, etc (Ware et al., 2020). Studies have shown significant correlations between objective behavioural characteristics collected through mobile phones and wearable devices and depressive symptoms (Rohani et al., 2018; Ware et al., 2020).
The challenges and limitations of active and passive EMA include the degree of technical understanding of patients, socioeconomic status, inconvenience, and burden on participants, device-dependent issues, missing data, unconstrained data collection, data security and privacy concerns, and potential discrepancies between active and passive EMA data (Farhan et al., 2016; Lu et al., 2018a). Despite these challenges, identifying and monitoring digital biomarkers through EMA can provide valuable insights.

2.2 Data Collection Methods

The field of automated depression recognition faces significant challenges in terms of data quality and availability, which can impede the accuracy and robustness of automated systems. The efficacy of such systems is heavily reliant on the data sets available for training, which are often difficult to obtain due to the complex and multifaceted nature of depression.

Most automated depression recognition systems use self-reported depression screening questionnaires as ground truth, necessitating careful study design to collect accurate data. Additionally, relying on data collected in a laboratory setting can limit the ability to capture underlying behaviours specific to an individual with depression, as well as restrict the number of subjects and longitudinal data points available. Further complicating matters is the difficulty of sharing data, especially in the case of clinical trials aimed at monitoring treatment effects for multiple endpoints. To advance the field, it is crucial to identify and utilise publicly available high-quality data sets, as well as consider novel data collection methods that can overcome the

challenges associated with depression research.

In general, designing the experimental environment and selecting the appropriate modalities are critical factors in collecting depression data. This section examines the data collection settings and various data sets that are currently available to the research community and provides a detailed analysis of their characteristics, including public or private accessibility, modality, annotation type, number of subjects, etc. The section also highlights the unique features of a new data set described in Chapter 4, which offers a scalable approach to data collection for research on depression recognition.

2.2.1 Data Collection Study Set-up

Depression research often involves collecting data from clinical or general populations. Clinical data for depression are collected by recruiting participants from hospitals or psychological clinics. In these studies, participants are assessed using DSM-IV or HAMD standards, as well as various other diagnostic tools, such as Mini International Neuropsychiatric Interview (MINI) and BDI. Different recruitment approaches, such as flyers, posters, social networks, personal contacts, and mailing lists, have also been used in some studies. In the general population, participants are recruited from the public who are not under any clinical treatment for depression. Self-reported questionnaires form diagnostic or screening measures, such as PHQ-9 or BDI-II, for participants to self-identify symptoms, which then inform a severity scale.

Designing the experimental environment is crucial to obtain valuable patterns for predicting depression. To ensure consistency in data collection, inclusion criteria must be met, agreements, protocols, and consent forms are signed prior to the experiment, and devices such as cameras, microphones, and sensors are arranged. Screening procedures, including exclusion criteria, are different for clinical and general populations. In clinical populations, healthy control groups are also included to have a more balanced sample space and clinical labels. Interviews designed by human and virtual humans are used to assess symptoms related to depression. Ethics approval requirements must be met before the commencement of studies.

Different modalities, including speech and video samples, physiological signals, and text, have been employed to improve the performance of depression assessment. For audio clips, a computer or laptop is used to record the data samples, while for the video modality, the number of cameras and other attributes are used to record the face and whole body separately from different angles. A careful study protocol design is required for eliciting certain behaviour using tasks that are scripted vs. nonscripted or free-form. Passive monitoring, mobile phone data, GPS locations, and longitudinal endpoints are also used to remotely measure depression.

2.2.2 Datasets

The majority of existing databases have been restricted to the research for which they were initially developed (He et al., 2021), without public access to depression recognition studies. However, several publicly available databases have been created for depression recognition purposes, including the ones released by the renowned AVEC data sets (Continuous Audio / Visual Emotion and Depression Recognition Challenges) of 2013 (Valstar et al., 2013) and 2014 (Valstar et al., 2014), and the Stress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ).
The data sets AVEC2013(Valstar et al., 2013) and AVEC2014 (Valstar et al., 2014) data sets are both subsets of the Audio-Visual Depression Language Corpus. AVEC2013, annotated with the Beck Depression Inventory II (BDI-II), consists of 340 videos in German. Participants engaged in various human-computer interaction tasks in front of a webcam and microphone, including free speech, reading, singing, and picture-based association tasks. The challenge released 150 longer-duration videos as AVEC 2013 to the public. AVEC2014, a subset of AVEC2013, comprises 300 videos with shorter durations than AVEC2013. In AVEC2014 the organisers chose to only retain two tasks, Freeform and Northwind, based on its relevance to recognising depression. For each of these tasks, 150 videos were made available. The depression annotations are BDI-II scores that range from 0-63, annotated at an event or video level. Audio and video clips are available for both tasks.

DAIC-WOZ (Gratch et al., 2014), part of the Distress Analysis Interview Corpus has been used for AVEC2016 (Valstar et al., 2016), and AVEC2017 depression recognition challenges. It is a hugely popular and (Ringeval et al., 2017b) widely used data set that is part of a larger corpus. The data set comprises 189 semi-structured clinical

interviews designed for the diagnosis of psychological distress conditions such as anxiety and depression. The data set utilizes a virtual interviewer with strictly controlled emotional status during interviews. Four different modes of interviews were designed to collect data, including face-to-face, teleconference, wizard-of-oz, and automated. It consists of audio, video, and deep sensor modalities and includes galvanic skin response (GSR), electrocardiogram information (ECG), and participant respiratory data. Text transcriptions of the interviews are also made available in the data set. It is available for public access, although direct access does not provide any video data anymore, only the features that were extracted from the frames.

E-DAIC is an extended version of DAIC-WOZ (Ringeval et al., 2019), collected from semiclinical interviews designed for the diagnosis of psychological distress conditions such as anxiety and depression. This database has 163 development samples, 56 training samples, and 56 test samples, with age, gender, and PHQ-8 scores labelled. E-DAIC was provided as part of the AVEC2019 (Ringeval et al., 2019) challenge.

The VHQ-1 data set (Jaiswal et al., 2019a,c) comprises audio and video recordings of 55 participants who participated in structured interviews under different mediation modes, namely face-to-face, teleconference and virtual human. Self-report questionnaires for depression (PhQ-9), personality (Big-Five) and anxiety (GAD-7) were used. A comparative study was conducted between different modes of administration, including self-administration. The videos were filmed in a laboratory environment with another person present and feature a general population. The data is not available for public access.

Speech is a highly informative modality that contains markers of mental health disorders. The field of depression recognition using void or audio samples shows great advances in collecting natural, real-world data. There are several studies in the literature that collect continuous audio samples using accessible devices like smartphones to collect ecologically valid data from subjects with depression.

The Early Mental Health Uncovering (EMU) data set (Tlachac et al., 2021) is designed to detect mental health disorders using passive and active modalities. The EMU data set was collected by a team of researchers in 2019-2020 and contains heterogeneous

data, including scripted and unscripted voice recordings, smartphone logs including text messages, and Twitter data, all labelled with demographics and depression screening scores (PhQ-9) and anxiety (GAD-7). To collect data, the EMU app was used and the data are labelled with depression severity and anxiety level scores. The data set is publicly available and features around 60 participants recruited via Amazon Mechanical Turk.

The study conducted using the moodable framework (Dogrucu et al., 2020) employed a non-intrusive or passive method utilising an Android app to collect data from 335 MTurk participants for the evaluation of depression. Participants were instructed to record a predetermined phrase, PhQ-9 questionnaires, as well as historical sensor data from their smartphone and recent social media posts. The implemented random forest models were able to detect depression and suicidal ideation with satisfactory F1 scores and accuracy levels. 226 participants provided scripted voice samples along with different levels of passive data. The research shows a potential opportunity to change the current approach to the screening of depression by utilising machine learning on readily available biometrics, such as voice samples, and historical data from smartphones and social media, making the screening process more efficient and accessible.

Huang et al. (2018) investigated depression assessment using real-world voice samples and PhQ-9 responses collected from 887 participants using smartphones and showed promising results. The app included elicited and freeform tasks such as free speech, read out loud, sustained vowel, and diadochokinetic repetition. The data set is not available for public use. The MODMA data set (Cai et al., 2020) is a publicly available data set that includes EEG and audio data from clinically depressed patients and normal controls, carefully diagnosed and selected by professional psychiatrists in hospitals in China. The EEG data set consists of data collected using both a traditional 128-electrode EEG recorder and a new wearable 3-electrode EEG recorder for pervasive usage collected from 55 participants. Audio data was collected from scripted and unscripted interviews, story reading, and emotional picture watching.

RADAR-MDD (Matcham et al., 2019, 2022) is a large-scale data collection study that spanned multiple sites involving more than 600 participants from the clinical population. The study aimed to remotely monitor the participants to reliably track

and predict relapse in MDD. The participants wore wearable devices that tracked activity levels and interacted with various apps on their smartphone. The collected data consisted of both active and passive modalities, including data relating to sleep patterns, physical activity, stress, mood, sociability, speech samples, and cognitive function. Depression labels were defined by self-reported questionnaires IDS-SR. Active app interaction also collected PHQ-8 scores. This large-scale study took place from 2017 to 2021 and is a highly promising direction for remote monitoring of mood disorders.

A plethora of studies exist in the literature that collects and analyses passive data through wearables and smartphone sensor data. Student Life (Wang et al., 2014) is a continuous monitoring app to track mental health and well-being, including smartphone data such as GPS, call logs, activity of 48 college students over a 10-week period, and self-reporting PHQ-9 was used as labels. The data set is available for public access. Other studies such as LifeRhythm (Lu et al., 2018b) followed similar protocols and collected passive smartphone data from 79 college students for more than six months.

The ORYGEN database (Ooi et al., 2011) contains video and audio data samples recorded from discussions between parents and their children. The BlackDog database (Alghowinem et al., 2012) was collected from a clinical study conducted by the BlackDog Institute in Sydney, Australia. The data set contains speech data recorded during conversations between interviewers and participants who had met the criteria for DSM-IV. The clinical interaction was performed by asking specific questions about events stimulated by specific emotions, allowing for the exploration of emotional regulation in individuals with mental illness.

The Pittsburgh database (Yang et al., 2012a) is another notable data set that comprises 57 depressed participants in clinical treatment for depression. Participants were required to meet the DSM-IV criteria for MDD and the severity of their condition was assessed at different intervals by multiple clinical interviewers. This database is publicly available. The BD database (Çiftçi et al., 2018) is a data set consisting of 46 patients and 49 healthy controls from a mental health service hospital. The data set contains semi-structured interviews and gathers sociodemographic and clinical patterns. Additionally, the Young Mania Rating Scale (YMRS) and MADRS were used

to estimate depressive and manic features, and audiovisual samples were recorded during video sessions. The data set is annotated by bipolar mania/depression ratings and has been used as challenge data in AVEC2018.

From the above discussion, it is clear that collecting data for depression research is a challenging endeavour due to several limitations. Conventional lab-based data collection methods face many constraints, such as having to carefully design protocols, which are usually task groups that may not reflect or elicit natural behaviour, and collecting well-distributed data sets with healthy and control groups. Capturing environmental factors such as ambient lighting, noise and the presence of other people can reflect mood and provide valuable information for a holistic scene understanding identifying the markers of low mood, imperceptible otherwise by the clinicians. Therefore, collecting data in-the-wild, where individuals can behave naturally, with minimal obstruction, is crucial for depression research.

From an openness perspective, most databases are only used for in-house research and not released publicly for depression recognition studies due to privacy concerns. However, some databases such as AVEC2013, AVEC2014, DAIC-WOZ, Pittsburgh data set, and MODMA data set are available for researchers. Most databases were collected by the US and EU regions, except for MODMA, which is a Chinese database. Furthermore, recruiting subjects for depression studies conducted in a lab environment is very challenging, which leads to limited data samples in all databases.

Wearable activity trackers have the ability to collect detailed sensor data that characterise users' behaviour and physiology, known as digital biomarkers. This information could be used to detect depression in a timely, unobtrusive, and scalable manner. However, it is important to consider the acceptance of wearable technology and people's willingness to share their personal data. Despite this, many people are more willing to provide videos of their face and voice rather than data from full-time activity trackers, which may contain highly sensitive information such as real-time GPS or phone correspondence. Therefore, videos of subjects' face and voice recorded from video diaries can provide critical information as digital biomarkers for automated mood analysis.

To address the scarcity of labelled face and voice data for identifying mood disorders,

this thesis presents a novel video data set that captures participants' longitudinal behaviours in unconstrained, in-the-wild environments recorded using active smartphone interactions. The data set includes recordings of 139 subjects collected over a period of 3 weeks, which is considerably high compared to studies with similar settings in the field. This data set is a valuable resource for the development of machine learning algorithms and the identification of physiological biomarkers of mood disorders. A comparison of the above data sets to the newly collected MoodSeasons data set is provided in Table 2.2. The data set distinguishes itself through several key attributes. First, it incorporates diverse data modalities, including audio, video, and smartphone interactions, enhancing the understanding of participants' emotions. Additionally, Mood-seasons provides longitudinal data collected over three weeks, enabling the study of mood and behaviour changes over time, a valuable insight into mood disorders' dynamics. Moreover, it stands out by collecting data in natural, uncontrolled environments, capturing genuine behaviour and subtle mood markers. Furthermore, the data set will be made publicly accessible, promoting wider research and development in depression recognition. Lastly, with data from 139 participants, Mood-seasons offer a substantial data set size, supporting evaluation of robustness and generalizability in depression models. These attributes make Mood-seasons a valuable resource in the study of depression recognition and mood disorders.

2.3 State-of-the-art Automated Depression Recognition

Since depression has been associated with neurophysiological and neurocognitive abnormalities, which is evident in facial and voice behaviour [37], [137,] audiovisual signals for Automatic Depression Estimation (ADE) have become a main area of research. Recognition of depression can be formulated as a classification and/or regression problem based on audiovisual cues. The following sections describe the state-of-the-art methods in automated depression recognition classified based on the input modalities such as audio, video, or multimodal data.

2.3.1 Audio Modality

In the field of audio-based Automatic Depression Estimation (ADE), feature extraction has relied primarily on hand-crafted features. Since the late 1990s, a range of feature representation approaches have been proposed to assess depression severity.

Dataset Name	Size	Data Modalities	Annotation	Accessibility	Longitudinal	Environment
AVEC2013	340 videos	Audio, Video	BDI-II scores	Publicly available	No	In the lab
AVEC2014	300 videos	Audio, Video	BDI-II scores	Publicly available	No	In the lab
DAIC-WOZ	189 sessions	Audio, Video, Sensors	Clinical interviews	Limited video access	No	In the la
E-DAIC	163 samples	Audio, Video	PHQ-8 scores	Publicly available	No	In the lab
VHQ-1	55 participants	Audio, Video	Questionnaires	Not publicly available	No	In the lab
EMU data set	60 participants	Audio, Text, Social Media	Screening scores	Publicly available	Yes	in-the-wild
Moodable framework	335 participants	Audio, Social Media	Machine Learning models	Limited accessibility	Yes	in-the-wild
Huang et al.	887 participants	Audio, PhQ-9 responses	Smartphone-based data	Not publicly available	No	in-the-wild
MODMA data set	55 participants	EEG, Audio	Clinical diagnosis	Publicly available	No	In the lab
RADAR-MDD	600+ participants	Audio, Wearable Data	Self-reported questionnaires	Research access	Yes	in-the-wild
Student Life	48 students	Smartphone data	PHQ-9 self-reporting	publicly available	Yes	in-the-wild
ORYGEN database	N/A	Video, Audio	Clinical interviews	Not publicly available	No	In the lab
BlackDog database	N/A	Speech data	Clinical interactions	Limited accessibility	No	In the lab
Pittsburgh database	57 participants	Video, Audio	DSM-IV criteria	Publicly available	No	In the lab
BD database	46 patients	Video, Audio	Clinical patterns	Research access	No	In the lab
Mood-seasons	**139 participants**	**Audio, Video, Smartphone**	**PhQ-8 Self-reporting**	**Yes (in the process)**	**Yes**	**in-the-wild**

Table 2.2: Comparison of different depression recognition data sets.This table compares different depression recognition data sets in terms of their size, data modalities, annotation, accessibility, and longitudinal data. The introduced Mood-seasons data set is a new and promising longitudinal, multimodal (audio, video, transcripts) data set for depression recognition, collected from a diverse range of participants in natural settings.

Some traditional (shallow) methods include the duration of pause (Stassen et al., 1998), the speech rate and pitch variation (Cannizzaro et al., 2004), and prosodic, voice quality, spectral, and glottal features (Moore II et al., 2007). Low-Level Descriptor (LLD) indicators, such as prosodic, source, formant, and spectral features, have been found to be effective predictors of depression (Cummins et al., 2015b).

However, hand-crafted features require manual tuning and expert domain knowledge for feature selection, which can be time-consuming and limited in scope for applications and domain. On the other hand, deep learning-based representations have shown considerable performance margins across various disciplines, including ADE (Alpert et al., 2001), suggesting that deep learning methods could serve as a valuable alternative to traditional hand-crafted features in ADE research.

In 2016, a pioneering deep learning model, DepAudioNet, was introduced by Ma

et al. (2016a) to extract depression representations from vocal cues. This model employs LSTM and CNNs to encode discriminative audio representations for depression recognition. DCNN is used to model spatial feature representations from raw waveforms, while LSTM learns short-term and long-term feature representations from mel-scale filter banks (Shannon and Paliwal, 2003). DepAudioNet extracts various scale representations, including high-level, short-term, and long-term features along with sampling methods to alleviate class imbalance for depression analysis.

Despite the limited size of the available depression databases, deep learning-based depression recognition methods have attracted significant interest from researchers. In 2018, a fusion of deeply learnt and hand-crafted features was proposed to effectively measure the severity of speech depression (He and Cao, 2018). This framework used DCNN to learn and fuse shallow and deep patterns, extracting hand-crafted features such as LLD features and Median robust extended local binary patterns (MRELBP) from audio and spectrograms. Raw audio and spectrograms were input into the model to obtain deep learned features. Joint fine-tuning was employed to learn complementary representations between hand-crafted and deep-learnt features.

To increase data samples and improve the accuracy of the ADE task, in (Yang et al., 2020) a Deep Convolutional Generative Adversarial Network (DCGAN) was proposed that used a two-level learning strategy to improve the convergence speed of training by dividing the feature maps into blocks and applying a DCGAN model to each block. Researchers have used DCNN and LSTM, to assess depression severity, in Niu et al. (2020) sought to overcome the limitations of traditional feature design methods by converting audio segments into spectrograms to feed into a deep architecture. In 2021, Niu et al. (2021) introduced a novel framework that combined Squeeze-and-Excitation (SE) components and Time-Frequency Channel Attention (TFCA) blocks with DenseNet's Dense blocks and Transition Layers, creating the Time-Frequency Channel Attention and Vectorisation (TFCAV) network.

In Dong and Yang (2021), a deep architecture for ADE from speech was achived by fusing Speaker Recognition (SR) and Speaker Emotion Recognition (SER) features to enhance ADE performance, and introducing the Feature Variation Coordination Measurement (FVCM) algorithm. The ResNet-50 network was used as the foundation for the SR and SER models, offering insights into different patterns for depressed

individuals in audio and video frames. The studies encourage further exploration in ADE, emphasising the importance of considering various modalities and feature extraction methods in this field.

2.3.2 Visual Modality

Visual cues are crucial in the recognition of deep depression, leading researchers in affective computing to investigate discriminative patterns in videos for ADE. An early attempt to employ deep learning for the detection of depression from static images was made by Al Jazaery and Guo (2018), who developed a two-stream network using facial images and optical flow features to learn depression patterns. They introduced Appearance-CNN and Dynamics-CNN to model static and dynamic patterns for depression recognition. The first step involved training a model from scratch on the public CASIAWebFace Database with 494,414 images from 10,575 subjects (Yi et al., 2014). The second step fine-tuned the pretrained model for ADE. The study (Al Jazaery and Guo, 2018) inspired subsequent works in the field of deep learning for depression recognition and analysis. Zhou et al. (2018) proposed a novel deep architecture called DepressNet to learn representations from images for depression recognition. They pre-trained different deep architectures (AlexNet, ResNet, GoogleNet) on the CASIA database and constructed DepressNet by changing the softmax layer into a regression layer, followed by a global average pooling (GAP) layer.

In De Melo et al. (2019), the researchers adopted a 2D-CNN and distribution learning to model depression patterns, using the expected loss function to predict depression levels. Their method outperformed most state-of-the-art methods in AVEC2013 and AVEC2014. Song et al. (2020) presented a novel multiscale architecture for depression recognition and considered human behaviour primitives (AUs, gaze direction, and head pose) as frame-wise feature representations. Spectral heatmaps and spectral vectors were used to mine multiscale representations of expressive behaviour, which were then input into DCNN for ADE. The method achieved promising results on the AVEC2013 and AVEC2014 databases.

Furthermore, studies based on the AVEC2013 and AVEC2014 databases (De Melo

et al., 2020) introduced a two-stream DCNN framework to learn patterns from RGB images and encoded images from video clips. The appearance stream took static images as input, while the temporal stream took image sequences as input. They used the mean squared error function to address the regression issue and a simple fusion method (average pooling) to combine the outputs of the two networks for the ADE task.

Many Works (Dibeklioglu et al., 2017),(De Melo et al., 2020) pre-train deep models on large-scale databases (e.g.ImageNet, VGG, VGGFACE etc.) using deep architectures (e.g., VGG, ResNet, etc.) and fine-tune them on depression databases, e.g., AVEC2013 and AVEC2014, to enhance performance. Novel loss functions are also proposed to improve depression recognition performance.

He et al. (2021) introduced a novel network that combined 2D-CNN networks and attention mechanisms for the recognition of depression. They proposed a DCNN with attention mechanisms and weighted spatial pyramid pooling to model global features. The architecture consists of two branches that focus on local patches and global features from the entire facial region. . In terms of pre-processing, researchers primarily use MTCNN, OpenFace, RetinaFace, and Dlib toolkits to detect and crop the facial region, providing a solid foundation for depression detection.

Although single image features have been widely employed in ADE tasks, yielding promising results, these approaches often neglect the temporal information that may be beneficial to ADE. To address this issue,Al Jazaery and Guo (2018) proposed using C3D and RNN to extract spatio-temporal features from video clips in two different scales for depression recognition. Their framework includes two components: loose- and tight-scale feature extraction components, which involve fine-tuning of deep models and temporal feature aggregation. The C3D Tight-Face model learns high-resolution features, while the C3D Loose-Face model focusses on larger face regions to capture global features. An RNN is then used to model the temporal features learnt by both C3D models, and a mean operation is applied for prediction.

De Melo et al. (2020) proposed a combination of different C3D architectures to learn spatio-temporal patterns from the full face and local regions, which are further combined with 3D Global Average Pooling (3D-GAP) for the prediction of depres-

sion. The local C3D architecture focusses on discriminative information in the eye region, while the global C3D architecture targets spatiotemporal patterns based on the entire facial region. The proposed method was tested on the AVEC2013 and AVEC2014 databases and achieved high performance.

Uddin et al. (2020) employed LSTM to model sequence information from video data. Deep facial expression features were extracted using a CNN and pooled by Temporal Median Pooling (TMP) method to feed the LSTM module for ADE. Experiments conducted on the AVEC2013 and AVEC2014 data sets indicated the efficacy of the proposed methodology. They extracted dynamic features to model subtle emotions from facial regions. In de Melo et al. (2020), a 3D framework called the multiscale spatiotemporal network (MSN) was developed to learn characteristic information from video clips. The model employed several parallel convolutional layers to learn substantial spatio-temporal variations from facial expressions, and used multiple receptive fields to maximise the use of distinct spatial areas from the facial region for AD.

In 2021, several works (de Melo et al., 2021; He et al., 2022) proposed predicting the severity of depression. In (He et al., 2022), the authors presented an end-to-end pipeline to generate discriminative representations of entire video clips. Specifically, a 3D-CNN combined with a Spatiotemporal Feature Aggregation Module (STFAM) was trained from scratch on the AVEC2013 and AVEC2014 data sets, allowing the model to learn informative depression patterns. The STFAM integrates channel and spatial attention mechanisms as well as a 3D DEP-NetVLAD aggregation method to capture compact characteristics based on feature maps.

In (de Melo et al., 2021), a new deep learning architecture was proposed, called the Maximisation and Differentiation Network (MDN), to model facial expression variations closely related to depression. The MDN was designed without 3D convolutions and exploited discriminative temporal patterns learnt by two different blocks that modelled smooth or sudden facial variations. The models were validated in the AVEC2013 and AVEC2014 databases.

In comparison to static features or image-based features, image sequences can capture short-term and long-term spatio-temporal information from videos, leading to

improved training of deep discriminative models for depression recognition. From a training perspective, most of the works include a two-staged pipeline, which includes pre-training and fine-tuning stages.

Several methods have been proposed to aggregate segment-level features into audio or video-level features for depression severity prediction. Average pooling was adopted in (Valstar et al., 2014, 2013), while Meng et al. (2013) used MHH to process each component of audio segment-level features to aggregate the temporal sequence. Dhall and Goecke (2015) built upon the Bag-of-Words (BoW) approach in action recognition and facial expressions by constructing visual words from video segment-level features and generating aggregation results through frequency histogram calculations. They also examined the performance of depression detection using alternative statistical techniques, including mean, maximum, and standard deviation.

He and Cao (2018) observed the difficulty in tuning the Gaussian components during the aggregation process and integrated the Dirichlet process to automatically learn the number of Gaussian components based on the observed data and obtain video-level features for depression detection. Niu et al. (2019) demonstrated that average-pooling and max-pooling were special cases of L_p-norm pooling. By combining the L_p norm pool with the least absolute shrinkage and selection operator (LASSO), they identified the suitable parameter p for the detection of depression. The aggregate results were then obtained by calculating the L_p norm of each dimension of the features at the segment level.

2.3.3 Multimodal depression recognition methods

The integration of multimodal data has emerged as a promising approach to the recognition of depression, as it can leverage the complementary information from various sources to provide a more comprehensive and accurate understanding of an individual's mental state. This section aims to present the state-of-the-art techniques in depression recognition using multimodal fusion, focussing on the types of data used, the fusion methodologies employed, and the challenges and future directions in the field.

Researchers have used data related to facial expressions, eye movements, gestures, and posture to analyse depression-related behaviours. Speech and language features have been used to recognise depression, focussing on prosodic, spectral, and linguistic characteristics. Physiological data, such as electroencephalogram (EEG), heart rate variability (HRV) and skin conductance, have been used to examine the correlation between depression and physiological signals. Text-based data from social media platforms, including tweets, posts, and comments, have been analysed to identify depression-related patterns in language use.

In (Yang et al., 2016) the authors use decision trees to infer multimodal input features from audio, video, and patient transcripts that related to personality type, sleep, mood, etc. The state-of-the-art in detecting depression won the AVEC challenge using a topic modelling approach to exploit context-aware recognition and combine audio, video, and semantic features to achieve prediction tasks (Gong and Poellabauer, 2017). Deep learning approaches are also highly popular, (Dibeklioglu et al., 2017) using stacked auto-encoders to predict depression severity, (Zhu et al., 2017c) using deep convolutional neural networks only in images, and(Ma et al., 2016b) using convolutional neural networks in speech signals. Multimodal methods using deep learning also showed promising results on the DAIC WOZ data set featuring depression (Yang et al., 2017b).

In an attempt to predict depression scores, Gupta et al. (2014) combined the audio baseline characteristics of AVEC2014 with the acoustic characteristics of (Van Segbroeck et al., 2013) using late fusion. They integrated AVEC2014 video baseline features with supplementary video representations, such as LBP-TOP, optical flow features, and facial landmark motion (Gupta et al., 2014). The final multimodal result was obtained by linearly fusing the prediction scores of the audio and video modalities. In contrast, Pérez et al. (2014) generated predictions for affective dimensions, which were used as attributes for audio segment-level features. They also used facial landmarks to extract motion and velocity information from video segments, implementing a majority strategy for the predicted results from all segments.

Jain et al. (2014) applied Principal Component Analysis (PCA) to reduce the dimensionality of LLDs and used Fischer Vector encoding to obtain audio features. The multimodal representation was generated by concatenating audio and video features

for depression detection. In another study, audio features consisting of LLDs and MFCCs, as well as video features containing hand-crafted descriptors (e.g., LBP, LPQ, and Edge Orientation Histogram) and deep representations extracted by VGG-Face, were combined (Jan et al., 2017). The concatenated features were then input into the linear regression (LR) and partial linear regression (PLR) models, with the results of the two regressors weighted as individual depression scores.

Chao et al. (2015) demonstrated the effectiveness of deep learning methods using a multimodal framework with audiovisual cues in 2015. Visual features taken from the pre-trained 2D-CNN model and an LSTM-RNN is used to learn the temporal context from the audiovisual features. In 2017, a combination of 1D-DCNN and DNN methods for ADE was proposed (Yang et al., 2017d). This method used different models to merge audiovisual features and textual inputs from transcripts. Each modality used hand-crafted features that were entered into a 1D-DCNN and then into a DNN to assess the PHQ-8 scores. The three single models (audio, visual, text) were fused and entered into a DNN to determine the severity of the depression based on the PHQ-8 scale Yang et al. (2017c,d). The same method was also adopted in (Yang et al., 2017a, 2018a) with promising results.

The AVEC2018 Bipolar Disorder Sub-Challenge used a Bipolar Disorder Corpus (Çiftçi et al., 2018) and several methods were proposed to analyse bipolar depression, for example (Yang et al., 2018b) uses a DNN and fusion architecture using Random Forest. IncepLSTM, combined an Inception module and LSTM that was designed to handle bipolar disorder (BD) with irregular variations in different episodes (Du et al., 2018). Zhao et al. (2019) introduced a method that integrates unsupervised learning, transfer learning, and hierarchical attention from speech to assess the severity of depression, yielding promising results on the AVEC2017 depression challenge. In 2020, a new Spatio-Temporal Attention (STA) architecture and Multi-modal Attention Feature Fusion (MAFF) method were proposed to extract multi-modal features from audiovisual cues for depression severity evaluation (Niu et al., 2020). The method used 2D-CNN, 3D-CNN, and attention mechanisms to learn deep features, and the experiments showed that the proposed architecture outperformed most existing ones.

Multimodal fusion methods have generally produced optimal performance for ADE in various databases, though fusion of complementary information between au-

dio and video cues can be complicated (Niu et al., 2020). Classical methods have also been proposed for depression estimation since 2015, such as ordinal logistic regression (Jayawardena et al., 2020) and median robust LBP-TOP (MRLBP-TOP) (He et al., 2018). Deep learning techniques, such as 1D-CNN, 2D-CNN, and 3D-CNN, have been commonly used to learn discriminative patterns from static images and hand-crafted features. Moreover, attention mechanisms have been employed to learn salient patterns from deep-learnt features from multimodal data. The literature shows a clear trend of methods adopting multimodal fusion to leverage complementary signals to gain a comprehensive understanding of depression levels.

2.4 Summary and Research Gaps

This literature review has provided an overview of depression assessment methods, widely used depression data sets, and state-of-the-art techniques for automated depression recognition using audio, visual, and multimodal cues. Several key limitations motivate the research contributions presented in this thesis.

- Most existing depression databases are limited in size and diversity, comprising audiovisual data collected in controlled lab environments. This work introduces a novel longitudinal multimodal depression database with 139 participants recorded in natural settings over multiple sessions.
- Current models are heavily based on audiovisual signals captured during a single session by participants. However, there is no work that integrates longitudinal data into the prediction model. The work presented in this thesis utilises the longitudinal aspect of the Mood Seasons data set to provide robust models.
- Complex end-to-end multimodal fusion is underexplored, especially for integrating behavioural sensing with audiovisual cues. This work proposes an end-to-end architecture tailored for the traits of multimodal depression data considering short-term and long-term modelling of videos via sentence level and video-level predictions.
- Evaluation of model robustness and generalisability is limited, since models are usually evaluated on training data sets. This work utilises multiple data sets to evaluate the generalisation of the model.

- Although several methods in the previous sections use state-of-the-art methods and provide customised solutions, there is no benchmark available on different baseline methods that are easily adoptable. Although AVEC addresses this to a certain extent, the methods are siloed and harder to access. This work provides an extensive benchmark of state-of-the-art video analysis techniques on the newly curated Mood-Seasons and publicly available AVEC 2014 data sets.

In summary, while progress has been made in automated depression analysis, limitations remain regarding data diversity, modelling approaches, fusion techniques, evaluation protocols, and annotation requirements. This thesis collects a novel longitudinal multimodal depression data set and proposes new model architectures, self-supervised learning strategies, and rigorous cross-data set evaluation to advance the state-of-the-art in this important research domain.

Chapter 3

Background On Image Synthesis using GANs

One of the main contributions of this thesis is the development of a face manipulation method (Chapter 6) that can manipulate the pose and expressions in a given image, which can then be used to anonymise the identities of subjects for the analysis of depression without harming the process. This method relies on a class of deep learning models called generative adversarial networks (GANs) to synthesise high-quality and high-fidelity face images.

GANs were first introduced by Goodfellow et al. (2014), and since then several variants of the model have been proposed in the literature. It is important to review GANs in depth and detail, as the choice of architecture, loss functions, and training methodology is key to building high-quality generative models. This chapter presents the key theoretical concepts of GANs and their variants. Specifically, this chapter will cover the following topics: (i) the basic principles of GANs, including generator and discriminator networks, (ii) the training and optimisation of GANs and the challenges and tricks to train them, (iii) the different types of GAN, such as vanilla GANs, conditional GANs, etc., and (iv) the evaluation of GAN-generated images.

3.1 Generative Adversarial Networks

Generative adversarial networks, GANs, are a type of generative models that can learn an estimate of a given distribution p_{data}, representing the training set (Goodfellow, 2016). Their ability to represent high-dimensional data distributions implicitly makes them a suitable choice for semi-supervised and unsupervised learning (Radford et al., 2016).

3.1.1 Formulation

GANs consists of two players, set up against each other, the generator network which synthesises images and a discriminator network, which differentiates between samples coming from real and generated/fake data. The goal of the discriminator is to output with a high probability that the input is coming from the real data distribution, while the goal of the generator is to generate samples that can fool the discriminator. The generator does not have access to the real images directly, whereas the discriminator uses supervised learning by accessing both real and generated images that are labelled real or fake. The generator uses the error signal of the discriminator to improve the quality of the samples it generates. A common analogy (Goodfellow, 2016) for this scenario is that of a money forger and a police, where the generator acts as a forger trying to create fake money while the discriminator tries to detect the fake money from the real ones. This min-max game between the generator G and the discriminator D, with the optimisation objective to train the two models together, is summarised as

$$\min_G \max_D \quad \mathbb{E}_{x\sim p_{data}} \log D(\mathbf{x}) + \mathbb{E}_{z\sim p_g} \log(1 - D(G(\mathbf{z}))) \tag{3.1}$$

where p_{data} is the distribution of images from the training set and p_g is the distribution of the generated samples, G is the generator model, D is the discriminator model, $\mathbb{E}_{x\sim p_{data}} \log D(\mathbf{x})$ is the expected value of the discriminator's log probability of a real image and $\mathbb{E}_{z\sim p_g} \log(1 - D(G(\mathbf{z})))$ is the expected value of the discriminator's log probability of a fake image. This is the standard cross-entropy cost, which is minimised while training a standard binary classifier with sigmoid output.

For training a GAN, the optimisation involves finding the discriminator parameters

that maximise its classification accuracy and the generator parameters for which the discriminator produces high confidence for the generated samples. In the vanilla GAN version, (Goodfellow et al., 2014) the discriminator is trained using two mini-batches, one consisting of **x** sampled from the real images and the other from the noise distribution. The two network gradients are updated simultaneously, one updating θ^D to reduce the discriminator's cost and the other updating θ^G to reduce the generator's cost.

3.1.2 On the optimal training of GANs

In the original vanilla GAN, the authors show that for a fixed G, the optimal discriminator is given by,

$$D^*_G(\mathbf{x}) = \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}$$

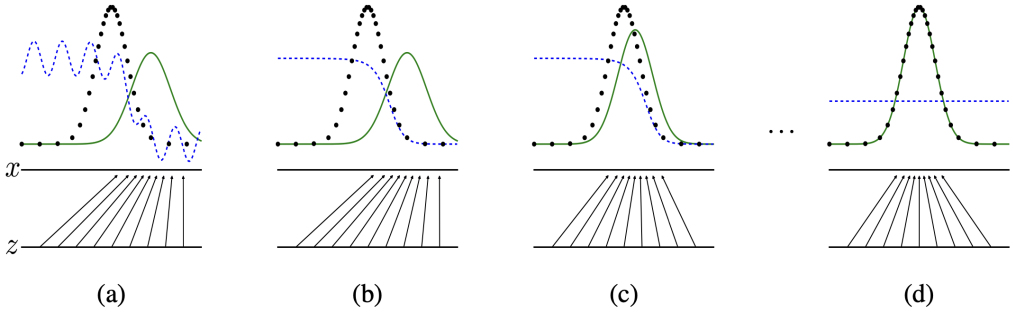


Figure 3.1: Generative adversarial networks (GANs) are trained by concurrently updating the discriminative distribution $D(x)$ (blue dashed line) to distinguish between real data samples $x \quad p_{data}(x)$ (black dotted line) and generated samples x' $p_g(x')$ from the generative distribution $G(z)$ (green solid line). The domain from which the noise vectors z are drawn uniformly is shown as the lower horizontal line. The upward arrows demonstrate how the mapping $x' = G(z)$ imposes the non-uniform $p_g(x')$ on transformed z. $G(z)$ morphs $p_g(x')$ by contracting in high-density regions and expanding in low-density areas to match $p_{data}(x)$.(a) When $p_g(x') \approx p_{data}(x)$, $D(x)$ becomes an imperfect classifier. (b) $D(x)$ is updated to converge to the optimal $D^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x')}$.(c) The gradients of D guide $G(z)$ to map to areas more likely classified as real by $D^*(x)$. (d) At convergence, $p_g(x') = p_{data}(x)$ so $D(x)$ nears $\frac{1}{2}$. The discriminator cannot differentiate between real and generated data.

The above figure illustrates the training strategy of the discriminator. Assume that both z and x are one dimensional. The black arrows show the mapping from z to x and model density p_g is represented by the green curve and the data distribution p_{data} is represented by black dots and the discriminator density is represented by the blue line. The optimally trained discriminator estimates the ratio between the model density and the sum of the model and data densities (Goodfellow et al., 2014). When the discriminator output is large, the model density is low and where the discriminator output is too low, the model density is high. The generator is updated to create better model density by pushing towards the discriminator uphill, i.e., the generator's mapping $G(z)$ should move in the direction that maximises $D(G(z))$. After training for several steps, where $p_g = p_{data}$ and, the discriminator outputs the same value for x and z, $D(\mathbf{x}) = \frac{1}{2}$.

Algorithm 1: The algorithm for training GANs. The number of steps k refers to the discriminator updates.

for *number of training iterations* **do**
  for *k steps* **do**
    { Sample minibatch of m noise samples $\{z_{(1)}, z_{(2)}, \dots z_{(n)}\}$ from the noise prior $p_g(z)$} { Sample minibatch of m examples $\{x_{(1)}, x_{(2)}, \dots x_{(n)}\}$ from the data distribution $p_{data}(x)$} { Update the discriminator by ascending its stochastic gradient: { $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right]$}
  {Sample minibatch of m noise samples $\{z_{(1)}, z_{(2)}, \dots z_{(n)}$ from the noise prior $p_g(z)$} {Update the generator by descending its stochastic gradient: $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)$ }

For an optimal discriminator, the training objective of the generator can be written as,

$$\min_G V(D^*, G) = \int_x \left(p_{data}(x) \log D^*(x) + p_g(x) \log(1 - D^*(x))\right) dx$$

$$= \int_x \left(p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} + p_g(x) \log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right]$$

Given Jenson-Shannon divergence,

$$JSD(P_{data} \parallel P_g) = \frac{1}{2} D_{KL}\left(p_{data} \parallel \frac{P_{data} + P_g}{2}\right) + \frac{1}{2} D_{KL}\left(p_g \parallel \frac{P_{data} + P_g}{2}\right)$$

where D_{KL} is the Kullback-Leibler divergence.

$$= \frac{1}{2}\int_x p_{data}(x)\log\left(\frac{p_{data}(x)}{\frac{p_{data}(x)+p_g(x)}{2}}\right) + \frac{1}{2}\int_x p_g(x)\log\left(\frac{p_g(x)}{\frac{p_{data}(x)+p_g(x)}{2}}\right)$$

$$= \frac{1}{2}\int_x p_{data}(x)\log\left(\frac{2p_{data}(x)}{p_{data}(x)+p_g(x)}\right) + \frac{1}{2}\int_x p_g(x)\log\left(\frac{2p_g(x)}{p_{data}(x)+p_g(x)}\right)$$

$$= \frac{1}{2}\Big(\log 2\int_x p_{data}(x)\log\left(\frac{p_{data}(x)}{p_{data}(x)+p_g(x)}\right) + \frac{1}{2}\log 2\int_x p_g(x)\log\left(\frac{p_g(x)}{p_{data}(x)+p_g(x)}\right)$$

$$= \frac{1}{2}\Big(\log 4 + \min_G V(D^*, G)\Big)$$

Therefore,

$$\min_G V(D^*, G) = 2\dot{J}SD(P_{data} \parallel P_g) - \log 4$$

It shows that when training D is optimal, training G is equivalent to minimising Jensen-Shannon divergence between p_{data} and p_g. Since the Jensen-Shannon divergence is nonnegative and zero only when they are equal, the optimal training criterion is $\min_G V(D^*, G^*) = -\log 4$, which is the global minimum and $p_g = p_{data}$.

However, these theoretical results cannot be used to guarantee convergence because adversarial networks represent a limited family of p_g distributions via the function $G(z;\theta_g)$ and optimise the parameters θ_g of the distribution instead of p_g(Goodfellow et al., 2014).

The non-saturating loss function

In the mini-max game, the cost function of the generator is $J^{(G)} = -J^{(D)}$. The generator's cost function becomes,

$$J^{(G)}(\theta^{(G)}, \theta^{(D)}) = -[\frac{1}{2}E_{x\ p_{data}}\log D(x) + \frac{1}{2}E_z \log(1 - D(G(z))])$$

The discriminator minimises the cross entropy loss while the generator tries to maximise the same function. When the discriminator trains for several iterations, it may come to a point where it successfully classifies a generated sample as fake, the gradient of the generator's loss becomes zero, thus exposing the problem of vanishing gradients in the vanilla GAN training (Goodfellow, 2016). In order to

mitigate this, the authors propose a non-saturating loss function for the generator which is largely heuristically motivated where,

$$J^{(G)} = -\frac{1}{2}E_z \log(D(G(z)))$$

$$= \max_{\theta_g} E_z \log(D(G(z)))$$

This will ensure a strong gradient in case the discriminator becomes better while training. In the previous setting, the generator minimised the likelihood of the discriminator being correct, and in the modified game, the generator maximises the likelihood of the discriminator being wrong. The authors (Goodfellow, 2016)(Mescheder et al., 2018) show that the non-saturating version of the vanilla GAN loss works well in practice. Ideally, the discriminator should be trained optimally before updating the generator; in practise, the discriminator is trained a few iterations before a generator's update (Radford et al., 2016). There are many known training difficulties for GANs (Goodfellow, 2016) (Radford et al., 2016) (Salimans et al., 2016) such as non-convergence, occurring where an optimiser like stochastic gradient descent is used for the task of finding the Nash equilibrium of the game (Salimans et al., 2016).

3.1.3 Challenges in GAN training

Although the authors in (Goodfellow et al., 2014) provide the theoretical convergence of GANs, that is, the existence of global minimum when $p_{data} = p_g$, when the discriminator is optimal, in practise this is hardly the case (Arjovsky and Bottou, 2017a)(Arjovsky et al., 2017)(Salimans et al., 2016). The two prominent issues related to training of GANs are instability and mode collapse. Mode collapse or partial mode collapse is a phenomenon where the generator creates samples with same composition (texture, colour, etc.) for different inputs. A wide range of research focusses on addressing these issues and proposes ways to overcome the limitations in the original formulation.

In a notable work presented in (Arjovsky and Bottou, 2017a), the authors show that despite the theoretical guarantee of global minimum, when the discriminator is trained to convergence, its error goes to zero which means that the Jenson-Shannon Divergence in $2\log 2 - 2JSD(p_{data} \parallel p_g)$, maxes out to $\log 2$ and the function $V(D^*, G)$ tends to zero. The authors further point out that this phenomenon occurs when

the distributions are disjoint or their supports lie in low-dimensional manifolds. In this case, it is always possible to train a perfect discriminator that can distinguish between the real and fake samples. The discriminator therefore provides constant output for the two distributions, and the gradient update to the generator vanishes. The generator does not have any useful information from the discriminator to train in a meaningful way, leading to training instability.

The authors (Arjovsky and Bottou, 2017a) also show that the alternate non-saturated formulation of generator's cost function also leads to similar instabilities and mode collapse. The gradient of the non-saturated cost function is given by $[\ \Delta\theta = \bigtriangledown_\theta E_{z\tilde{p}(z)}\big[-\log(D((_\theta(z)))\big]$

$$= \bigtriangledown_\theta KL(P_{g_\theta} \parallel P_{data}) - 2JSD(P_{g_\theta} \parallel P_{data}),$$

where P_{g_θ} and P_{data} represent two distributions. Here, cost minimises the KL divergence and maximises the JSD, which are two opposites. This causes instability in the generator's gradient updates. Furthermore, the loss function minimises the reverse KL divergence $KL(P_{g_\theta} \parallel P_{data})$, which means that the discriminator assigns a very high cost to samples resembling fake data and tries to find observations that are more likely to be generated from the real distribution. This formulation is susceptible to mode collapse or partial mode collapse. However, it accounts for the good quality of images generated by the GAN (Arjovsky and Bottou, 2017a).

3.1.4 Heuristics and measures to stabilise GAN training

Following are some measures suggested by recent research (Salimans et al., 2016) to help alleviate the above-mentioned instabilities and mode collapse associated with GAN training:

- **Instant noise** - proposed in (Sønderby et al., 2016),(Arjovsky and Bottou, 2017a) suggests adding Gaussian noise to both generated and real samples during training to in order to have a well-defined divergence between the real and fake distributions without a common support.

- **Minibatch discrimination** - where the discriminator has an additional input feature which defines the distance of each sample with respect to other samples in a minibatch, identifying generator samples that are too close to each other for reducing mode collapse

- **Feature matching** - where the generator has a new cost function trained to match the expected intermediate features of the discriminator
- **Historical averaging** - where the cost function includes a penalty term for preventing parameters from deviating from the average parameter values from the previous times
- **One-sided label smoothing** - where the targets for the discriminator 1 is replaced with 0.9 so that the highly confident discriminator does not provide weak gradients to the generator
- **Virtual batch-normalisation**- where each sample **x** is normalised with respect to the statistics of a reference batch sample which was fixed at the start of the training.

3.1.5 Variants of GAN

DCGAN

The Deep Convolutional GAN (DCGAN) architecture, proposed by Radford et. al (Radford et al., 2016) was one of the first stable architectures for training a GAN to generate images. The main contributions of the DCGAN paper include the CNN architecture they proposed after extensive evaluation. Spatial pooling functions were replaced with strided convolutions and up-sampling was carried out using fractionally strided convolutions. This allowed the networks to learn its own up sampling and down sampling functions leading to better image generation (Radford et al., 2016). The second observation was to use batch normalisation in both generator and discriminator networks to stabilise training. Batch normalisation (Ioffe and Szegedy, 2015) was applied except in the output layer of the generator and the input layer of the discriminator.

The authors proposed removing the fully connected layers after the convolutions for achieving deeper models. This was achieved by providing the output of the last convolutional layer of the generator as the input to the discriminator. The discriminator uses a sigmoid function on its final convolutional features. The work also showed that using leaky ReLU instead of ReLU activations in the discriminator produced better quality samples. The DCGAN architecture is shown in Figure 3.2.

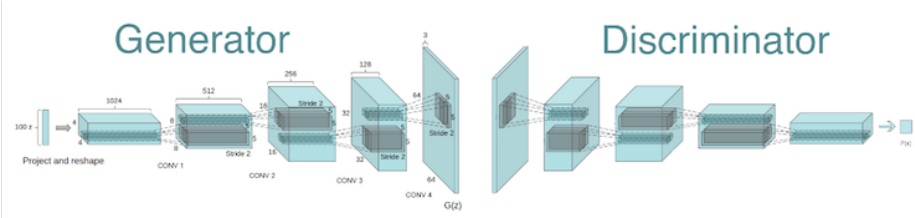


Figure 3.2: The fully convolutional architecture of the DCGAN generator and discriminator networks.

DCGAN showed that the latent code was learnt with semantic information by performing vector arithmetic in the latent space that showed semantic changes in the output. The authors also presented a supervised setting for using the learnt GAN representations for classification problems in popular datasets such as CIFAR-10 and SVHN. To achieve this, the authors concatenated the last layer of the discriminator's features trained on the Imagenet dataset and further used an L2-SVM classifier to train on the features, resulting in high accuracy.

Wasserstein GAN

The Wassertein GAN was proposed by Arjovsky et al. (Arjovsky et al., 2017) with an alternative formulation for the cost function of the GAN, using an approximation of the Wasserstein distance. They proposed minimising the Wasserstein -1 distance or Earth Mover's distance between the real and model data distributions. The authors argue that this new distance function is a more reasonable choice for approximating distances between disjoint distributions that may lie in low-dimensional manifolds using gradient descent (Arjovsky et al., 2017). The Wasserstein distance $W(P_{data}, P_g)$ is defined as the minimum cost of transporting mass to transform the distribution P_{data} to P_g. The distance is defined as

$$W(P_{data}, P_g) = \inf_{\gamma \in \prod(P_{data}, P_g)} E_{(x,y)\sim\gamma}\left[\parallel x - y \parallel \right],$$

where $\prod(P_{data}, P_g)$ represents the set of all joint distributions $\gamma(x, y)$ whose marginals are defined by P_{data} and P_g. For an intuitive point of view, probability distributions can be seen as the amount of mass placed at each point, and EM distance is the minimum work required to transform $p_{data}(x)$ to $p_g(\theta)$. Therefore, the function $\gamma(x, y)$ is the optimal transport plan defined by the joint probability distribution

$\prod(P_{data}, P_g)$ with marginals, P_{data} and P_g. To find the EMD, we multiply $\gamma(x, y)$ by the Euclidean distance at each point x,y. This is derived as 3.1.5. Since the infimum is intractable, according to Kantorovich-Rubinstein duality, the WGAN optimisation function is defined as

$$\min_{G} \max_{D \in \mathcal{D}} E_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] \; - \; E_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{\mathbf{x}})] \tag{3.2}$$

where $\mathcal{D}$ is a set of 1-Lipschitz functions and $\mathbb{P}_g$ is a model distribution defined by $\tilde{\mathbf{x}} = G(\mathbf{z}),\ \mathbf{z} \sim p(\mathbf{z})$. This is implemented in practice by clipping the parameters of the discriminator, to ensure that the parametric space of the critic lies within a compact space after each gradient update. This is a simple way of enforcing K-Lipschitz constraint, but works well in practice. The discriminator or the critic is trained well before each generator update. The loss metric, Wasserstein distance, was shown to correlate with the generated image quality and convergence of the generator, providing a meaningful loss metric. The WGAN provides a more stable way of training GANs with the Wasserstein distance metric that is continuous and differentiable. It does not suffer from vanishing gradient problems (Arjovsky et al., 2017) and can reduce mode collapse. The authors report experiments using DCGAN architecture and conventional GANs. The experiments showed that WGAN was able to train and generate quality samples without batch normalisation and was not very sensitive to the choice of non-linear activation functions.

Improved Wasserstein GAN (WGAN-GP)

Following the research on stabilizing GAN training, Gulrajani et al. (2017a) introduced an improved method for training Wasserstein GANs. They focus on the problems arising from using weight clipping in WGAN to enforce Lipschitz constraints for optimisation and demonstrate that weight clipping biases the critic to learn simpler functions, reducing its capability. The authors propose the use of a gradient penalty to eliminate this behavior and replace weight clipping. The gradient penalty term enforces a penalty on the gradient norm of the critic with respect to the training samples.

A differentiable function is 1-Lipschitz if and only if it has gradient with norm at most 1 everywhere. They propose to use gradient penalty as a way to constrain the norm of the gradient of the critic's output with respect to its input. The modified

cost function is,

$$L = E_{\tilde{x}\sim P_g}[D(\tilde{x})] - E_{x\sim P_{data}}[D(x)] + \lambda E_{\hat{x}\sim P_{\hat{x}}}[(\parallel \Delta_{\hat{x}} D(\hat{x}) \parallel_2 -1)^2]$$

, where $\hat{x}$ is sampled from straight lines between pairs of points sampled from the real and generated distributions (Gulrajani et al., 2017b). The implementation also removes batch normalisation as the critic's gradient norm is penalised with respect to each input independently, and suggests using layer normalisation instead of batch normalisation. The authors report experiments on various types of architecture including DCGAN, ResNet-101, to generate up to 128 x 128 samples in the LSUN dataset. They reported an improved performance of using gradient penalty over weight clipping in the CIFAR-10 dataset using inception scores as evaluation metric.

Hinge-Loss

Simulataneous works by (Lim and Ye, 2017a) (Tran et al., 2017) propose the hinge loss for adversarial training with the following objective function,

$$V_D(\hat{G}, D) = E_{x\sim p_{data}(x)}[\min(0, -1 + D(x))] + E_{z\sim p_g(z)}[\min(0, -1 - D(\hat{G}(z)))]$$

$$V_G(G, \hat{D}) = -E_{z\sim p_g}[\hat{D}(G(z))]$$

Hinge Loss as shown in (Lim and Ye, 2017a) is based on the geometrical interpretation of the discriminator as a linear classifier, such as an SVM that uses maximal margin hyperplane to separate two classes. Here the intuition is that the discriminator updates away from the hyperplane while the generator updates towards the hyperplane. Recent results and experiments in (Zhang et al., 2018) ,(Miyato et al., 2018) show good quality samples, stable training, and reduced mode collapse with this cost function.

Spectral Normalisation

Spectral normalisation was introduced by (Miyato et al., 2018) as a weight normalisation technique to stabilise GAN training in the discriminator. The spectral normalisation method also tries to ensure Lipschitz continuity in the functions learnt by the discriminator. It controls the Lipschitz constant of the discriminator function by restricting the spectral norm of each layer. The Lipschitz constant is given by the

largest singular value of the matrix, W or the spectral norm of W. Therefore, the method normalises the spectral norm of the weight matrix W of each layer so that it satisfies the Lipschitz constraint $\sigma(W) = 1$,

$$W_{SN} = \frac{W}{\sigma(W)}$$

Also, the power iteration method is used to accelerate the computation singular value decomposition thereby making spectral normalisation efficient and lightweight. Recent experiments and results in (Zhang et al., 2018)(Miyato et al., 2018)(Brock et al., 2018) show the effectiveness of using spectral normalisation in GAN training. The work, (Zhang et al., 2018) examines the training of GANs using a combination of different training choices such as Hinge + SN, Hinge + GP, etc.

Least Squares GAN

Least Squares GAN (Mao et al., 2017) addresses the problem of training instability and the generation of better quality samples by proposing a new cost function of the least squares in the discriminator. The authors show that minimising the objective function is equivalent to minimising the Pearson X^2 divergence. The training objective is given by:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}E_{x\sim p_{data}(x)}\left[(D(x)-b)^2\right] + \frac{1}{2}E_{z\sim p_z(z)}\left[(D(G(z))-a)^2\right]$$

$$\min GV_{LSGAN}(G) = \frac{1}{2}E_{x\sim p_z(z)}\left[(D(G(z))-c)^2\right]$$

where a and b and fake and real labels and c is the label that the generator should assign to the sample in order to fool the discriminator. Training becomes stable because the LSGAN assigns a high penalty to the generated samples that lie far away from the correct side of the decision boundary, thus pushing the generated samples to lie close to the decision boundary or the real data manifold (Mao et al., 2017). This will also eliminate the case of vanishing gradient because the high penalty of correctly classified, yet far apart samples will yield higher gradients for the generator to learn. The authors demonstrate in their experiments the superior quality of the samples produced by LSGAN.

Conditional GAN

Mirza and Osindero (2014) showed that GANs can be extended to include conditional information if the generator and discriminator received additional information on class / label, y. They achieve this by concatenating both generator and discriminator input with the labels, y, where y represents a one-hot label vector containing the class information corresponding to the real images. The new objective function becomes

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}\big[\log D(x \mid y)\big] + E_{z\sim p_z(z)}\big[1 - \log D((G(z \mid y))\big]$$

Conditional GAN has been successfully applied in image-image translation problems (Isola et al., 2017). Auxiliary classifier GAN introduced in (Odena et al., 2017) the authors introduce a classifier to enforce class conditional information in the discriminator. Instead of feeding the discriminator input with the class information from the ACGAN output, the authors modified the discriminator with an auxiliary decoder network that outputs the class label for the training data. The authors show high-quality images produced by this approach at higher resolutions, such as 128 x 128, where the images are conditioned on the respective class.

3.1.6 Applications of GANS and notable works

Image-to-Image translation with Cycle GAN

Image-to-image translation is a class of problems in computer vision where the goal is to translate images from a domain X to images from another domain Y. There are two main settings for image-to-image translation, (i) paired and (ii) unpaired. In paired setting, it is supervised since the mapping from X to Y is well defined or for every image $x \in domainX$ there exists $y \in domainY$, (Creswell et al., 2018), (Hong et al., 2019). In an unpaired setting, the translation problem becomes unsupervised, where such a mapping is not explicitly provided to the network during training using aligned training data, for e.g., for pose translation the image of the person in the required pose would not exist in the training dataset. This can be challenging, as there may be multiple mappings of $domainX \rightarrow Y$.

Cycle GAN (Zhu et al., 2017a) addresses the problem of image-to-image translation in an unpaired setting, where they have two generators and two discriminators

to learn the mapping of domain $X \rightarrow Y$ and domain $Y \rightarrow X$. They use two types of losses, adversarial and cycle-consistent. The adversarial losses try to match the distribution of real images to that of the generated images, while the cycle loss forces the learnt mappings to be semantically meaningful and correspond to each other. The cycle loss is formulated as follows.

$$\begin{aligned} \min_{G_{X\rightarrow Y}, G_{Y\rightarrow X}} \max_{D_X, D_Y} \lambda_1 L_{GAN}(D_Y(G_{X\rightarrow Y}(X)), D_Y(Y)) + \\ \lambda_2 L_{GAN}(D_X(G_{Y\rightarrow X}(Y)), D_X(X)) \\ +\lambda_3 L_{cycle}(G_{X\rightarrow Y}(G_{Y\rightarrow X}(X)), X) \\ +\lambda_4 L_{cycle}(G_{Y\rightarrow X}(G_{X\rightarrow Y}(Y)), Y), \end{aligned} \tag{3.3}$$

where λ controls the balance between the importance of different losses. Cycle GAN uses least-squares GAN formulation and employ patch GAN architecture to the discriminator. They also find the identity map loss $\parallel G(X) \rightarrow Y(X) \rightarrow x \parallel$ is beneficial for retaining the color composition.

Image stylisation with Adaptive Instance normalisation

Image stylisation is the problem of transferring the inferred style from one image to another image while preserving the content of the latter (Huang and Belongie, 2017). In the style transfer literature, Given a content image I and a style image S, adaptive instance normalisation aligns the channel-wise mean and variance of a feature map I to match that of S (Huang and Belongie, 2017). There are no learnable parameters and the ADAIN is given by,

$$ADAIN(I, S) = \sigma(S)\left(\frac{I - \mu(I)}{\sigma(I)}\right) + \mu(S)$$

Here the normalised image I is scaled according to the variance of the style image S and shifted according to its mean. This simple mechanism essentially performs style transfer as shown in (Huang and Belongie, 2017),(Park et al., 2019a) forcing the local texture of content and style images to be similar.

Progressive Growing of GANs

Progressive growing of GANs was proposed by Karras et al. (2017) for the generation of high-quality images at high resolutions by progressively growing both the gener-

ator and discriminator networks. The framework starts from a low resolution and adds new layers to the networks, G and D progressively during training, capturing finer details at multiple scales. This method also increases training stability and reports the state-of-the-art inception score on unsupervised CIFAR-10 evaluation. The configuration adds layers progressively, at multiple resolutions, by fading them in smoothly, and increases the resolution of generated images. All the existing layers remain trainable during the entire process.

A few heuristics used by the authors include (i) minibatch standard deviation – where the standard deviation for each feature in each spatial location in a minibatch is computed and averaged over all features and locations at a single scale. This value is replicated and used as an additional feature map towards the end of the discriminator. It is similar to minibatch discrimination, to ensure variation in the generated samples. (ii) normalisation of pixel-wise features in the generator. The authors use WGAN-GP loss mainly for generating high-resolution images 1024x1024, in CELEBA and LSUN bedroom datasets. Results are also reported on least squares LSGAN loss to demonstrate the model reliability despite the choice of loss function.

3.1.7 GAN Evaluation

Inception score

The inception score, first introduced in Salimans et al. (2016) uses a pre-trained inception model on generated images to evaluate the presence of meaningful objects. The generated images should ideally have the desirable properties of being highly classifiable and diverse with respect to the class labels. The method first applies the inception model to the generated sample, $\mathbf{x}$, to retrieve its conditional label distribution $P(y|\mathbf{x})$. Images with meaningful content are expected to have a conditional distribution with low entropy; at the same time $p(y)$ should have high entropy to ensure high variation within class samples. The inception score is given by

$$InceptionScore = \exp\left(E_x KL(p(y|\mathbf{x})||p(y))\right) \tag{3.4}$$

Although the inception score is shown to be correlated with human judgment, there are several drawbacks, including its inability to detect mode collapse or over fitting.

Also, it is computed over a general inception model ignoring the real data distribution; the quantitative evaluation may not be representative.

Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) was proposed by Heusel et al. (2017) to have a reliable metric to evaluate the quality of generated samples. The FID is calculated between the real data distribution and the generated data distribution. The generated and real samples are into a feature space given by a specific layer of the Inception Net and assume that the representation follows a multidimensional Gaussian. The mean and covariance are estimated for the generated and real data. FID estimates the distance between these two Gaussian as

$$FID(P_r, P_g) = \parallel \mu_r - \mu_g \parallel \; + Tr(C_r + C_g - 2(C_r C_g)^{1/2}) \tag{3.5}$$

where μ_r, μ_g, C_r, C_g represent the mean and covariance of the real and generated feature embeddings, respectively. FID has been reported to be a reliable measure of discrimination ability, robustness, and efficiency (Heusel et al., 2017). The experiments show correlations with human judgment as well as with sample quality. However, drawbacks exist for FID, such as its inability to detect over-fitting by GANs.

3.2 Conclusion

All GAN variants presented in this chapter were used to develop the GAN framework presented in Chapter 7. However, each variant had its own limitations. For example, training DCGAN under a conditional loss formulation resulted in mode collapse, a phenomenon where the generator learns to produce a small number of very realistic images but is unable to generate a variety of different images. The WGAN and WGAN-GP variants were stable in training, but generated lower-quality samples when used in the conditional GAN setting. This is because the WGAN and WGAN-GP variants are designed to minimise the difference between the real and generated data distributions, but they do not explicitly encourage the generator to produce realistic images. On the other hand, the Hinge Loss GAN with spectral normalisation was the only variant that was able to train stably under the conditional GAN setting and generate high-quality and high-fidelity face images. This is because the Hinge Loss GAN with spectral normalisation is explicitly designed to encourage the generator to produce

realistic images. It does this by using a hinge loss function, which penalises the generator for producing images that are not realistic. Spectral normalisation is also used to stabilise the training process and prevent the discriminator from overfitting the training data.

Chapter 4

Design Of Real-world, Multi-modal And Longitudinal Data Collection Study

Mental health assessments usually include subjective reporting of symptoms at clinical consultations or as part of a study procedure. This is susceptible to recollection bias and does not account for the fluctuation of symptoms over time or in response to various contextual or situational triggers. Ecological Momentary Assessment (EMA) is a method that addresses these issues by allowing patients to report behaviours and experiences at a high frequency, in real time, and in more realistic settings. EMA, (Shiffman et al., 2008) which began with using paper and pen, is now a well-established digital approach given via smartphone app-delivered notifications.

Ecological Momentary Treatments (EMI) (Wichers et al., 2011) allows the administration of interventions to patients that are personalised to their EMA and administered throughout their daily lives. As personalised, accessible, and scalable treatments, they have the potential to revolutionise mood disorder management. Incorporating EMA into clinical practise provides the opportunity to evaluate treatment effects and outcomes using its fine-grained approach, improving clinical decision making and expanding its potential to have real-world consequences for managing mental health issues (Kamath et al., 2022).

Automated depression analysis uses multimodal data to predict depression from expressive affective behaviour captured in video recordings. High-quality anno-

tated data are required for cutting-edge machine learning approaches such as deep learning, which can infer complex behaviours like mood states. EMA is an excellent technique for delivering such high-quality data that adds context to objective clinical assessments and a fine-grained and personalised lens to an individual's expressive behaviour.
All existing data sets for depression are collected in the laboratory environment, with carefully designed tasks designed to elicit targeted psychological responses from subjects. We have understood that the current trend toward reliably monitoring depression involves easily accessible, convenient, real-time, and real-world data delivered through EMA. The natural in-the-wild conditions are important to capture daily mood states and its dynamics, as opposed to the limited lab-based environment.When you enter a lab where the study will need interaction with a researcher, people may become defensive of their sensitive information and wish to stay confidential, or they may just find it uncomfortable expressing themselves freely. As a result, laboratory investigations use carefully designed activities to remove emotional masking and obtain trustworthy results. Voice modulation is something individuals perform in public, and the voice tones may alter when you do an in-person in-the-lab study.

A smartphone app liberates one from having to meet a stranger and engage in a series of activities: One can access it from anywhere, at any time of their choosing, and at their own pace. A method such as this may elicit honest signals about the participants' underlying emotional states. This is one of the first in-the-wild data collection app for collecting depression scores that will be made available for further research to the community. This is an important contribution that will help push the boundaries of automated depression research where we can validate the potential of the state-of-the-art methods in real, in wild data sets rather than a carefully curated and specially designed task-based lab environment data set.

This chapter introduces the conceptualisation, design, implementation, and collection of a large-scale smart-phone-based data collection study, Mood-Seasons. The study collects Ecological Momentary Assessments (EMA) by recording clinically validated PHQ-8 questionnaire and audio-visual-language data from the subjects, using a smartphone in real-world and real-time conditions. The use of EMA provides a subjective contextualisation of the objective questionnaires used for assessments,

and an automated analysis of audiovisual language information from these EMA is highly valuable in assessing depression.

The first section describes the motivation and need to design an app for a data collection for real-world mood analysis. The second section describes the Virtual Human Questionnaire study which sought answers for the validity of questionnaire administering mediums through a comparison of different modes of questionnaire administration such as human-mediated, self-reported and virtual-human mediated questionnaires. The third section introduces the Mood-Seasons app study, the specifications, implementation, deployment, protocol, the ethical implications and the perceptions of the app and user engagement. The final section gives an extensive overview and specifications of the Mood-Seasons data set, which is a major contribution of this thesis.

4.0.1 Contributions

The contributions towards the VHQ study involved the design of the study, application for ethics approval, recruitment of participants, and conducting of the study.

The Mood-Seasons study was conceived, designed, and implemented solely by the author, including app development, ethics application, recruitment, follow-up, data collection, storage, and analysis. The author led the end-to-end execution of the Mood-Seasons study from initial design to final data analysis. This involved designing the study, developing the mobile app for data collection, obtaining ethics approval, recruiting participants, following up with participants during the study, collecting and securely storing the data, and analysing the final data set. The Mood-Seasons study represents a comprehensive individual contribution spanning the full arc of study design, implementation, and analysis.

4.1 VHQ Study As A Proof Of Principle

As a first step toward designing a large-scale data collection study that gathers mental health assessment data for the automated diagnosis of mood disorders, it is important to understand the validity of the collected assessment data. One of the questions we set out to answer through the Virtual Human Questionnaire (VHQ) study is if the

depression severity scores obtained from self-administration of the questionnaires (through electronic form) were significantly altered when administered face-to-face by another person or a virtual human.

The VHQ study showed that there is no significant difference in the way people answer questionnaires when interviewed by a human or virtual human compared to when self-administering the same. This enables direct use of the questionnaire scores collected using a smartphone app, to aid in mental health evaluation in combination with behaviour analysis.

The study examined the distribution of scores obtained from a number of questionnaires often used in the diagnosis of depression (PHQ-9), anxiety (GAD-7) and in personality assessment (BFI-10). The study, which consisted of 55 participants, involved the administration of these questionnaires in three different modes: self-administration using an electronic form, human interviewer (face to face and videoconferencing) and virtual human interviewer. The hypothesis is that the answers to the questionnaires are not significantly affected by the different modes of administration. Through a statistical analysis of the questionnaire scores obtained from each of these modes, the study revealed if the self-administration method of obtaining the questionnaire scores can be considered equivalent to obtaining the same through a human interviewer. The question aforementioned is fundamental to the design of a data collection study that will then base the automatic assessment of mental health on self-reported online questionnaire responses, deployed at a much larger scale.

VHQ study also emphasised another facet; if the human is replaced by a virtual human agent, does the hypothesis still hold? The main contributions of the VHQ study are as follows:

- demonstrates the use of virtual human agents as interviewers administering standard psychological questionnaires.

- shows that the answers given to the questionnaires do get significantly affected if the questionnaires are administered by a human interviewer compared to being self-administration.

- shows that there is no significant effect even if the human interviewer is replaced by a virtual human agent.

4.1.1 Virtual Human Questionnaires (VHQ) study

The study involved three interactive sessions where one of the three questionnaires, PHQ-9, BFI-10 and GAD-7 were chosen for administration via:

- Face to Face (FF): Here, the interviewer is a real human sitting in front of the participant who asks questions from the chosen questionnaire.
- Mediated human (MH): Here, the human interviewer sits in a room different from the participant. The interaction between the interviewer and the participant takes place through a video conferencing link, where the interaction occurs solely through a screen.
- Virtual Human (VH): Here, the interviewer is a fully functional virtual human implemented using the Aria ValPusa framework.

Each of the sessions consisted of questions from a different questionnaire delivered using a different method of delivery. The pairing of the questionnaire and the delivery mode was done at random for each participant and therefore varies from one participant to another. The order in which these sessions are run was also set randomly. This was done to prevent any order effect. It should be noted that after the first round of our study, we invited some participants (those who did not complete the PHQ9 questionnaire with VH interaction earlier) to come back and do the study only for the PHQ9 questionnaire with VH interaction mode to increase the number of samples for the VH interaction mode.

Participants completed an electronic version of the questionnaires the day before the study, which could then be compared with the scores obtained from the different interactive sessions of humans / virtual humans completed during the study. In this way, the study explores the effect of mediation in questionnaire responses.

4.1.2 Data Analysis and Results

The VHQ study collected data from 55 participants where 49% of participants had no depression ($P_s \in [0,4]$), 33% had mild depression($P_s \in [5,9]$), 13% had moder-

ate depression($P_s \in [10, 14]$) and 5% belonged to the Moderately severe category ($P_s \in [15, 19]$). Participants with severe depression scores were excluded from the study to mitigate harm following the ethical mandates set out by the school of computer science ethics committee.

VHQ study used the Two One Sided t-Tests (TOST) procedure Schuirmann (1987)

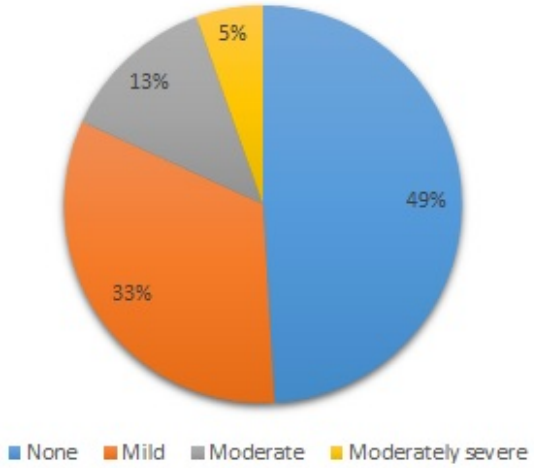


Figure 4.1: Depression severity distribution among participants, according to PHQ9 scores.

which is a popular method to test for equality between the means of 2 sets of samples. It is a statistical test that can be used to validate the hypothesis that the difference between 2 means is within a given interval. This interval is chosen to be the smallest effect size (mean difference) that can be tolerated and is specified in terms of an upper ($-\Delta_L$) and lower (Δ_U) equivalence bound. These bounds can be chosen in terms of raw differences or standardised differences such as Cohen's d (Cohen's d $= \Delta/\sigma$, where Δ denotes the raw difference and σ denotes the standard deviation). Using these equivalence bounds, two separate null hypotheses are defined: $H_{01} : \Delta \leq \Delta_L$ and $H_{02} : \Delta \geq \Delta_U$, where Δ denotes the observed effect. These hypotheses are tested using the t-test procedure. For each of these hypotheses, the p-values are calculated and compared with a threshold significance level (α) to determine whether they can be rejected or not. If both hypotheses can be rejected ($p < \alpha$ for each hypothesis), it implies that the observed effect (Δ) lies within the equivalence bounds ($-\Delta_L, \Delta_U$) and is statistically small enough to imply the equivalence of the two means.

Table 4.1, 4.2 shows the mean questionnaire scores, the number of samples, and

the p values of the TOST equivalence test for face-to-face (FF) and mediated human interaction (MH), respectively. In Table 4.1 it can be observed that for the face-to-face interaction mode, the effect size was found to be within its respective equivalence bounds (at significance level $\alpha = 0.05$) for PHQ-9 and GAD-7. Similar results were observed for the case of human mediated interaction (Table 4.2), where the effect size was found to be within the equivalence bounds.

The above results show that the differences in scores between self-administered questionnaires and real-human administered questionnaires were less than the minimum difference which could be regarded practically relevant (for PHQ-9 and GAD-7) indicating that self-administration of these questionnaires can be considered equivalent to the administration of these questionnaires by real human as well as virtual human agents. It can also be noticed that for the PHQ9 and GAD7 questionnaires, the mean scores for the self-administered case are always slightly higher compared to the mean scores from the questionnaires administered through interviews for each interaction mode (FF, MH and VH). This indicates that some participants might be suppressing their scores while being interviewed. However, this effect is too small to have clinical significance and scores can be considered practically equivalent in each case.

Questionnaire	Self-admin mean score	FF mean score	Mean of difference (std)	Number of samples	TOST p-value
PHQ9	5.05	4.95	0.10 (2.6)	20	<0.05
GAD7	5.20	4.72	0.50 (1.8)	18	<0.05

Table 4.1: Comparison of mean scores from self-administered questionnaire and questionnaire responses from face to face (FF) interaction.

Questionnaire	Self-admin mean score	MH mean score	Mean of difference (std)	Number of samples	TOST p-value
PHQ9	5.42	5.14	0.28 (2.3)	14	<0.05
GAD7	6.00	5.56	0.43 (2.7)	23	<0.05

Table 4.2: Comparison of mean scores from self-administered questionnaire and questionnaire responses from mediated human (MH) interaction

4.2 Mood-Seasons App Study

The study involves the design and development of a simplistic app that can capture videos from the user and allow them to complete the PHQ-8 questionnaire in order to collect audio, visual, and language data in-the-wild. The design of the application took the advice of experts in clinical psychiatry. The main reason for proposing an app for data collection is to achieve EMI, i.e., sample user experiences at any time and place chosen by the user. Due to increased accessibility, we can maximise the participation of the general population.

4.2.1 Definition of Mood

Mood has various definitions in the digital mental health literature and following section highlights those definitions and relates them to the mood assessment terminology used in the context of this thesis. Different definitions of mood include the following:

- Mood refers to a pervasive and sustained emotional state that influences one's perception of the world. Moods tend to be less intense than emotions and often lack a contextual stimulus. (Van de Leemput et al., 2014)

- Mood represents a predominant internal feeling state that persists over time and influences one's perceptions and behaviours. Moods tend to be less intense than emotions and often occur without a specific trigger. (Torous et al., 2018)

- A mood is a relatively long-lasting emotional state which influences an individual's perception of the world. Moods tend to be less intense than emotions and often emerge without a specific event acting as a stimulus. (Hollis et al., 2017)

In summary, mood refers to an internal, pervasive feeling state that is not tied to a specific trigger, is less intense than discrete emotions, and persists over an extended period of time. The mood influences one's overall outlook and response tendencies. This conception aligns with the thesis' use of mood assessments to characterise depression severity over a 1-2 week timeframe. Specifically, mood refers to a predominant affective state that reflects an individual's stable tendencies over the assessed time interval rather than momentary emotional reactions. The mood assessments used in this thesis provide a snapshot of this persistent internal state to

indicate the current severity of depression symptoms. Just as the literature describes mood on a continuum from positive to negative valence, the mood assessments here capture the degree of negative mood as a proxy for depression severity. In summary, mood is operationalised in this thesis as a persistent affective state that provides a window into the severity of depression symptoms over an extended period of time.

4.2.2 Specification Of The App

The app 'Mood-Seasons' has a simple interface that collects data from the user in the form of a self-reported PHQ 8 score and a short video clip and provides a mood feedback to the user. Figure 4.3 shows the current screenshots of the mobile app. The app interface starts with the PHQ 8 questionnaire and follows with a generic prompt for the user's video recording. The user can record the video by answering the generic prompt, facilitating the capture of facial and voice behaviours as appearance, including head pose, gaze, facial expressions, and speech.

The video prompts available on the app draws from generic topics, so that the user's

Figure 4.2: Illustrations of mood feedback, 'moodicons' assigned to the user at the end of a session. Winter indicates severe depression, rainy indicates moderate severe depression, autumn indicates moderate depression, spring indicates mild depression and summer indicates no depression.

response is not too emotive i.e. does not alter his/her current mood and at the same time interesting enough to discuss in front of the camera. The app provides mood feedback to the user based on the calculated PHQ 8 score illustrated as a 'moodicon', which was exclusively designed and illustrated by the author of the thesis for the app. A 'moodicon' represents different mood states mapped to seasons – summer representing the happy state, to winter corresponding to extremely low mood. Mood icons are illustrations, shown in Figure 4.2, so they provide an identifiable and engaging feedback to the participant regarding their mood state instead of a direct score. They are shown in Figure 3. The app also allows the user to keep track of

this mood feedback on a calendar, which they can then reflect on during the current month. This can be used as an incentive to use the app more often, if desired.

4.2.3 Implementation And Deployment

The implementation of the app was carried out using Cordova, JavaScript, PHP, and MySQL. Using Cordova for app development makes the app platform-agnostic, i.e., enabling it to run on browser, android, and iOS. Currently, the implementation runs on Android and Web browser and can run on multiple devices such as desktops, laptops, mobile phones, and tablets. Screenshots of the app are shown in Figure 4.4 and Figure 4.3.

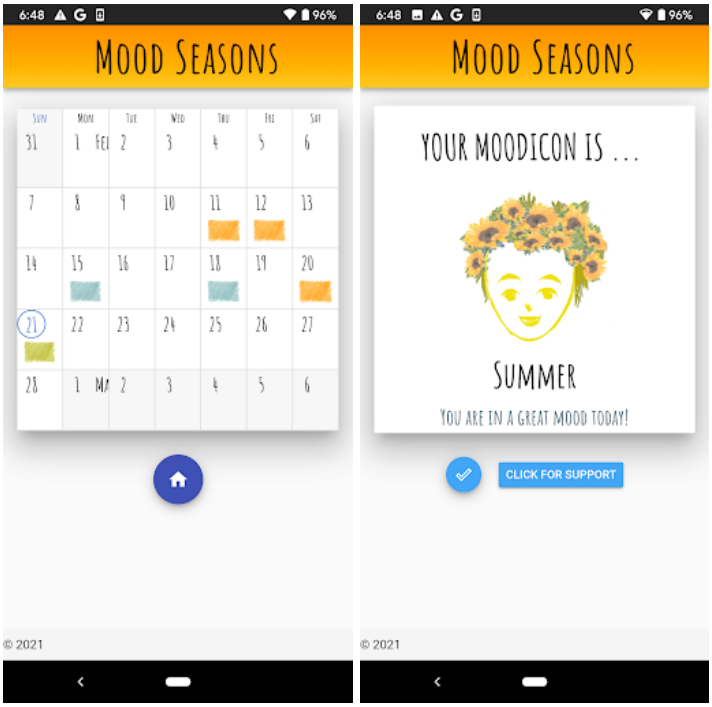


Figure 4.4: Demonstration of the smartphone-app, showing screenshots from the Mood-Seasons app interface.

No personal information that identifies an individual user is collected by the app. The app uses a device's unique ID to identify the user and saves the data in an encrypted manner in a School of Computer Science server. The session details of the user interaction, including the unique device ID, time stamp, depression severity score, and the location of the encrypted video file on the drive, is logged in a database

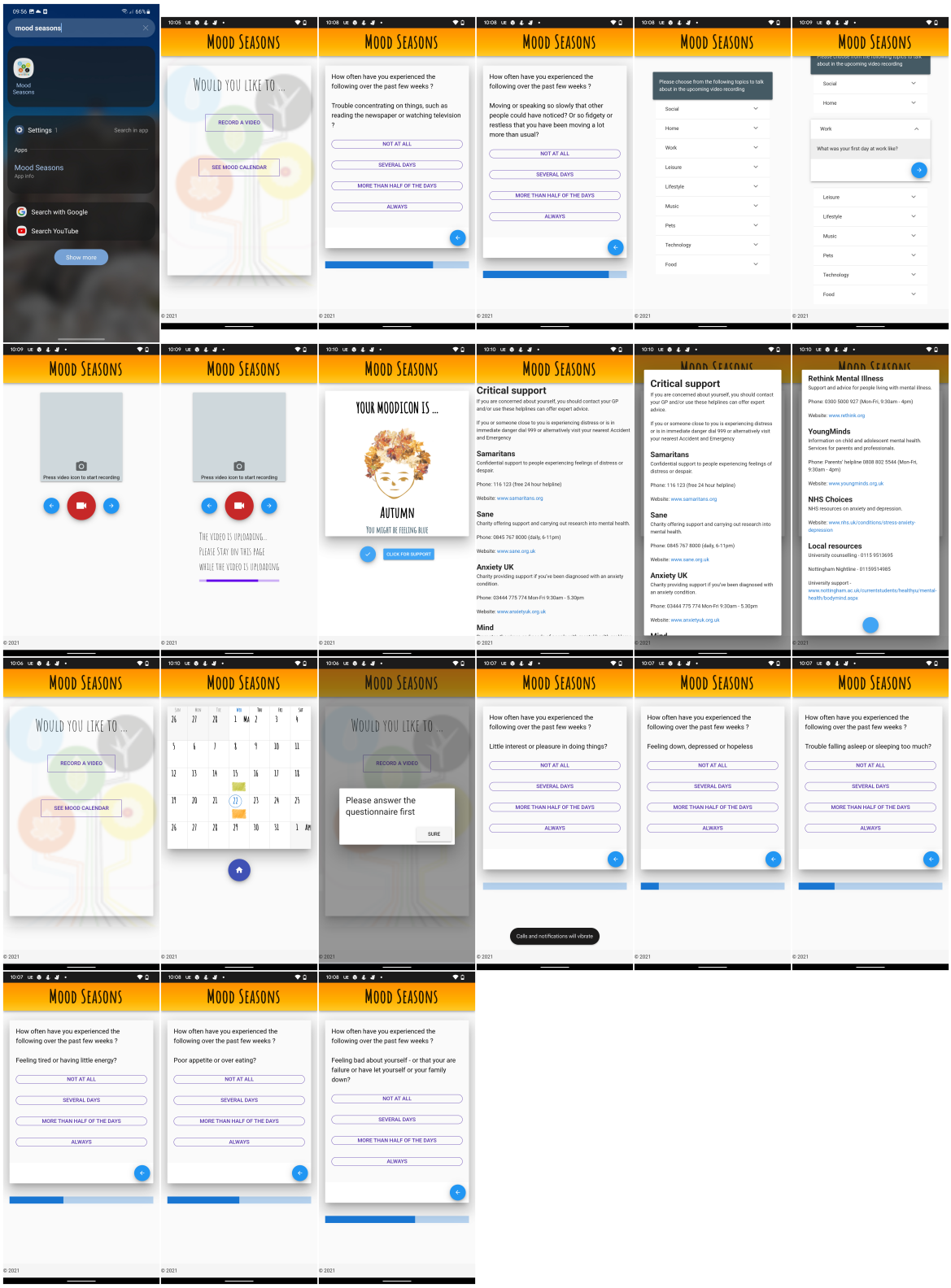


Figure 4.3: Mood-Seasons App screenshots

hosted by the School of Computer Science. Sending and receiving data to and from the server is handled by a server-side PHP script that follows security guidelines to ensure safe transmission of sensitive user data.

The app development also involved a PPI group with lived experiences to evaluate the usability of the app and make necessary updates to the design and execution. The ethics application for data collection was submitted and then approved by the ethics committee of the school of medicine. The study was carried out from March 2021 to June 2021. The initial protocol for deployment included face-to-face sessions with the user, which, however, was adapted to conduct the study entirely online, due to the Covid-19 pandemic without ever having face-to-face involvement from participants.

4.3 Mood-Seasons App Study Protocol

This section details the protocol laid out for the data collection study, approved by the ethics committee of the School of Medicine at the University of Nottingham. A detailed diagram of the protocol is provided in Figure 4.5.

4.3.1 Participant recruitment:

Participants were recruited from around the University of Nottingham, using poster advertisements sent to several faculty mailing lists and publicity through social media and contacting mental health organizations like MQresearch and Callforparticipants. The eligibility criteria for participant inclusion were:

1. 18 years and over
2. English fluency
3. Able to provide written informed consent
4. Has access to any Android device or any device with a web browser, camera, and microphone
5. Not currently undergoing treatment for depression or anxiety.

PHQ-9 scores should not be too high to ensure participant safety, which was assessed in the pre-enrolment part of the recruitment process. Any person who scored above

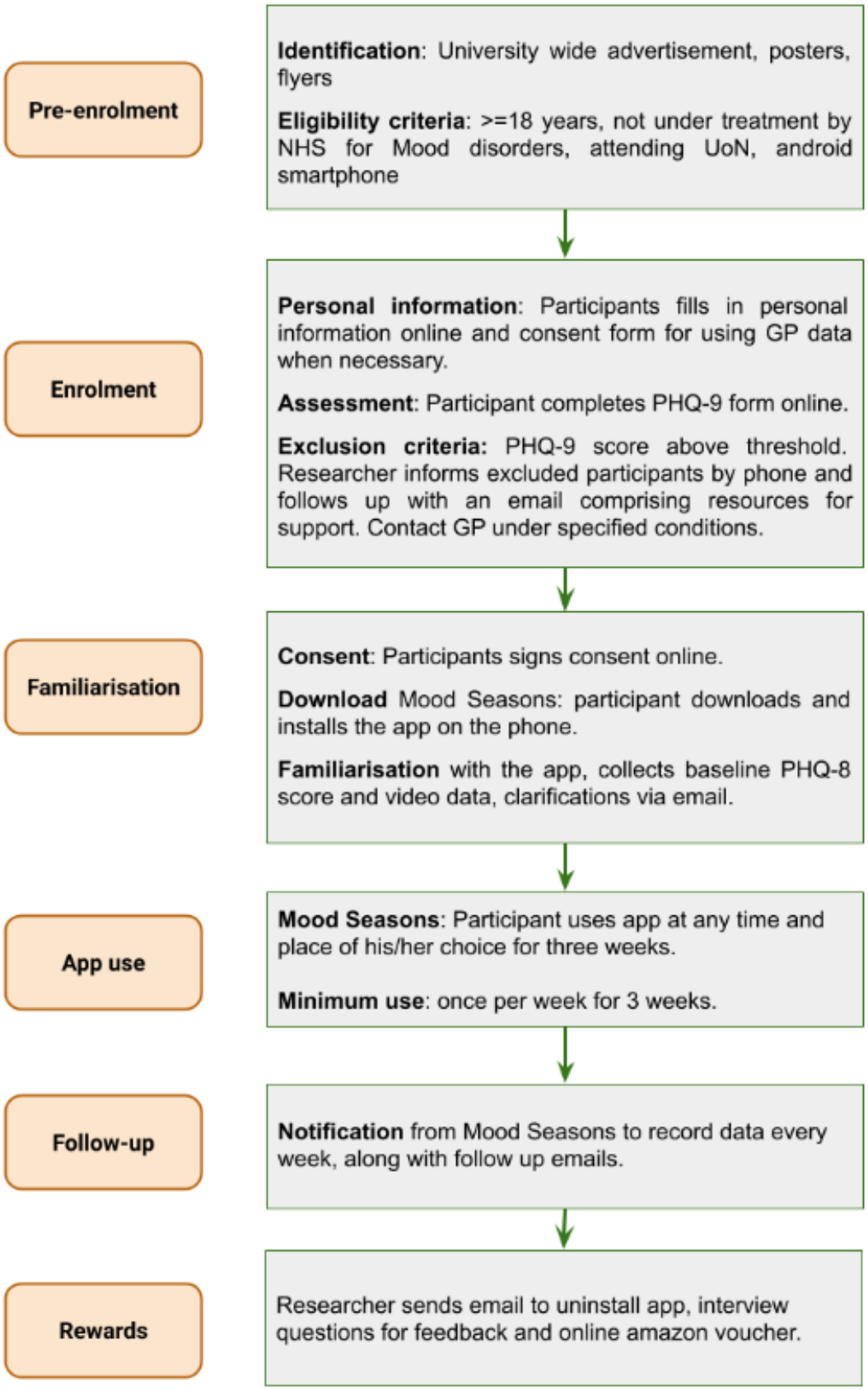


Figure 4.5: Mood-Seasons data collection study protocol

a certain threshold greater than 20 on the PHQ-9 questionnaire and greater than 2 on the self-harm question was not eligible for the study; instead, those who scored more than 20 were signed up to receive professional support.

4.3.2 Enrollment

Every participant who volunteers to participate in the study was first contacted by email and asked to complete an electronic version of the PHQ-9 questionnaire. The Patient Health Questionnaire – 9 (PHQ-9), is a self-administered questionnaire used to score 9 DSM-IV criteria for depression. It is widely used as a tool to monitor the severity of depression, as seen in Chapter 2.

A thorough review of the responses to the questionnaire established the eligibility of the person for the study in terms of the severity of depression. The responses of the selected participant to PHQ-8 are set as the 'ground truth' or 'gold standard' of their mental state. People who were undergoing treatment by the NHS for depression or those who scored high on the PHQ-9 questionnaire were not eligible for the study. These people were excluded from the study because there is a higher risk associated with such participants due to their elevated levels of depression, and they may require additional levels of support such as professional help.

For PHQ-9, a total score of > 20 or > 2 for the self harm question was considered high. These threshold scores have been determined from the results of a number of studies [18]. The threshold scores have not been explicitly mentioned in the participant information sheet and consent form because doing so will unnecessarily influence the participant and might affect his/her responses to the questionnaire.

Protocol for excluded participants

The excluded participants were to be individually contacted by the researcher, providing the reason they cannot participate, as well as giving standard details of how to seek help with their depression. The researcher ought to reach out to the participant via phone and then send a follow-up email, using a template phone call procedure and email approved by the ethics committee.

The participant information sheet specified that if the participant was unavailable via

phone for over a period of seven days, the researcher may contact the GP to provide additional support, as part of duty of care. Therefore, the study sought separate consent to contact the GP of the participant when they initially submitted personal information.

The study therefore collected the date of birth and the GP address information specifically for this kind of scenario in which the information needs to be passed on to the GP. The privacy safeguarding guidelines followed by the study also ensured that the participant's questionnaire data were destroyed unless they are found eligible for the study and sign the full consent form.

In case the participant was not eligible for the study (due to high PHQ-9 score), the data of their questionnaire were destroyed after making contact with the participant and ensuring that they have understood the advice given to them. If the participant could not be reached, their data ought to be passed onto the GP as mentioned above.

The main study

The eligible candidates on-boarded to the study with filling in a consent form sent along with a participant information sheet explaining their rights, detailed video instructions on how to download and install the app and the study protocol. The app collects self-reported PHQ-8 scores based on the past one week for three weeks and records a short video clip of the participant talking about a generic topic. The PHQ-8 will not include the last question pertaining to self-harm to minimise any potential triggering effects.

Each time a participant uses the app, they first fill in the PHQ-8 questionnaire displayed on the screen based on their experience over the past two weeks. Then the app navigates to a video prompt selection screen that contains a set of general questions which are arranged categorically, such as work, leisure, etc. The participant then chooses one of the questions from the categories, which takes them to a video recording screen. The participant then starts recording a video facing the camera with his face visible and talking about the prompt they chose previously.

The video duration is between 1 minute to 5 minutes. After the recording, the participant receives a mood feedback in the form of a 'moodicon', based on the

current PHQ-8 score. A mood calendar helps them track their mood during the current month. The participant can use the app as many times as preferred until the end of the study. The study required the participant to use the app three times at least, once after the app installation, and then follow it up every week for three weeks.The researcher sent follow-up emails every week to complete the sessions.

All participants who score high (>10) on the PHQ-8 questionnaires were notified by the app advising them to seek additional advice from their physician or the University Health Service (if not done already) should they feel their high scores are causing them any difficulties or if they believe they are getting worse. The app provides a list of resources available for mental health help and support.

After the end of three weeks, the researcher contacted the participants with the reward and asked for feedback on the participant's experience with the app. It also provided instructions on how to remove the app from the participant's phone. The study was incentivised where the participants received Amazon vouchers worth 10.00 GBP.

All data captured during the study (personal information, questionnaires, self-reported scores, audio, video) will be stored in accordance with the Data Protection Act and University of Nottingham's policies. At the start of the study, personally identifiable information about participants, such as name, date of birth, gender, email, phone number and GP address, was collected using online forms. A unique identification number was assigned to the participant, was used to store name and contact information (phone, email, GP address) separately from the date of birth, gender information, and the data sent from the app. Therefore, the participant cannot be personally identified. All data was stored in accordance with General Data Protection Regulation (GDPR).

The app itself does not collect any information such as name, age or gender, but uses the anonymous reference number and device's unique ID in order to identify the user sessions. Videos recorded in.mp4 format and responses to the PHQ-8 questionnaire in text format are encrypted and sent to a secure server. Before any kind of analysis, the video data was checked and pruned to remove identifiable information such as name or address. Personally identifiable information such as name, phone number,

GP address will be stored separately from demographic information such as gender, date of birth, linked by unique reference number. These data will be connected to the app data with the same identification number assigned to the participant. Therefore, the personal information is anonymised using this identification number. In other words, the participant's data will never be directly linked to their names within the app and an anonymous reference number will be used to establish the link. The Mood-Seasons data collection study and analysis follows the standard ethical procedures of the Faculty of Medicine and Health Sciences and the University of Nottingham.

Ethical implications

There are several ethical concerns regarding the nature of the data collected for this study, mainly because identifiable audiovisual recordings with mental health assessment information.The following section brings forth some ethical issues concerning this study into light.

The study recruits from the general population, which may comprise people with different levels of severity of depression. Some participants (or potential participants) may be considered vulnerable due to higher depression score and/or self-harm tendency. Data collection does not provide any treatment or mechanisms to address the needs of people with severe depression. The ethics of collecting audiovisual recordings from the severe depression category from the general population is not advisable, as opposed to a clinical population where clinical safeguards are put in place.

Therefore, people scoring very high (>=20 total or >=2 on suicide item) on the PHQ9 scale were not eligible to participate in this study and therefore were protected from any kind of possible exploitation or harm. In addition, these participants were encouraged to seek medical advice . The interaction with such participants took place only during the screening stage of recruitment (using online forms, phone, email) and any data collected till that point (name, questionnaire scores) was destroyed after informing them about the decision of the research team and ensuring that the participant has understood the advice given to them. In the unlikely case that the researcher was not able to get any response from such participants (for more than 7 days), their details ought to be passed on to their GP to ensure appropriate follow

up. If the participant scores high during the use of the app, then the app will notify the participant to seek help and provide links to additional resources for support. The app also lists an additional support information tab available for reference to all its users in search of help.

It is possible that a seemingly neutral topic could inadvertently trigger some psychological stress or anxiety beyond the risks encountered in everyday life. Participants were encouraged to discuss only generic topics that are comfortable to discuss while using the app. They are also provided with resources within the app, from which they can seek support.

This study requires facial recordings to be captured, and thus, by its very nature, the raw data will not be completely anonymous. However, all types of data were assigned an anonymous reference number, and thus the data will never be directly linked with their names. With the participant's consent (if they chose to "become part of a Mood-Seasons database to be shared amongst researchers who have signed up to the same ethical guidelines as the designers of this study" in the Consent Form), their data and restricted personal information (their age, gender, questionnaire scores) were part of the Mood-Seasons database that could be shared with other researchers who have signed up and agreed to follow the same ethical considerations. However, these researchers will not be allowed to further redistribute the data.

4.3.3 Perception of the App and Engagement

The study collected participant feedback informally asking how the participant felt during the study and the usability of the app. Participants also raised crucial points that would help enhance user engagement and, therefore, adherence. The following are the key points from their experience of using the app:

- The prompts for video recording could be improved, providing a wider array of choices for the participant based on their interests.
- Feedback in the form of Moodicons was seen as appealing and relatable to the participants.
- Several participants benefited from having a calendar that colour-coded their mood scores over the last 30 days.

- A popular request was the functionality to see all the months in the mood calendar, not only the current month.
- Since the study lasted three weeks, many participants 'simply forgot' to record the videos, unless prompted by the researcher. So, a functionality to remind them of the video recordings using notifications would be useful.
- Some participants preferred to hide their self-view during the recording, while others preferred to see themselves on the screen.
- Most of the participants felt that they were contributing to important research and were willing to record the videos as many times as needed.
- Some participants found the app extremely helpful because it served as a video diary for them, even if the study actively discouraged participants to engage in emotional triggers, if any.
- The convenience of the app – the flexibility to record the videos anywhere and any time of the participant's choice played a major role in their adherence.

4.4 The Mood-Seasons Data Set

The participants can be grouped into four categories of severity of depression based on the range of their PHQ-8 scores. A PHQ-8 score ranging from 0-4 indicates no depression, 5-9 indicates mild depression, 10-14 moderate depression, 15-19 moderately severe depression, and any score greater than 20 shows severe depression. The study excludes the severe depression category during the screening procedure. It uses the PHQ-8 version of the questionnaire PHQ-9, which removes the 9th item about suicidal thoughts in an attempt to enforce safeguards to avoid potential harm.

Of the 148 participants who completed the study, 134 were chosen for the Mood-Seasons data set. The study excluded information from any participants who did not provide their consent to share their data. The distribution of participants in the above four categories is shown in the Figure: 4.6. A cutoff score of PHQ-8 of 10 determines if a subject has depression. We can see that around 14% of the general population suffers from depression in the collected data set, of which around 10% falls into the moderate depression category and 4% shows signs of moderately severe depression.

Figure 4.7, 4.8 and 4.9 depict the gender, race, and age distributions of the participants, respectively. Figure 4.7 shows that the majority of participants are females,

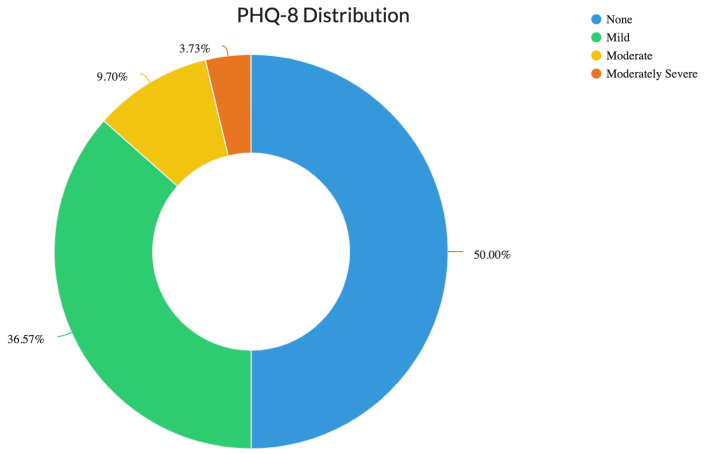


Figure 4.6: Distribution of PHQ-8 scores among participants for different categories of depression in the Mood-Seasons data set.

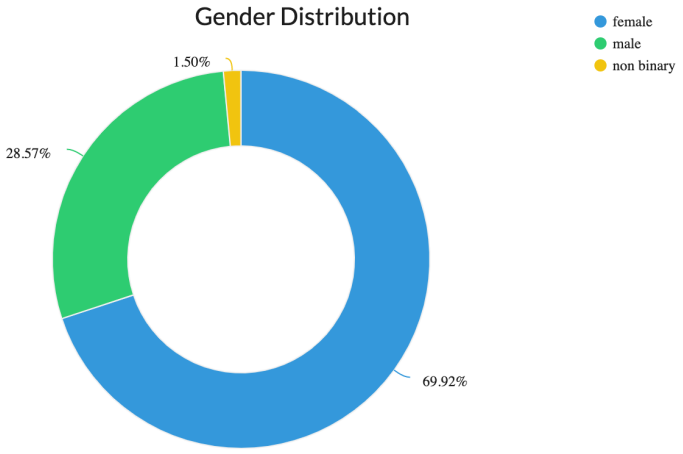


Figure 4.7: Distribution of gender in the Mood-Seasons data set.

with around 28.5 per cent male and 69.2
In Figure 4.8, 58% of the participants identified as Caucasian, 21 percent as Asian, 12 percent as African, 5 percent as East Asian, and 4 percent as Latino. The majority age group of the participants is 18-35, as shown in Figure 4.9 where at least two participants are present in all ranges. 77 per cent of the participants were aged between 18 and 35 years old, 20 per cent between 35 and 45 years old, and 3 per cent between 45 and 55 years old.

The study collected PHQ scores from participants over a three-week period, when

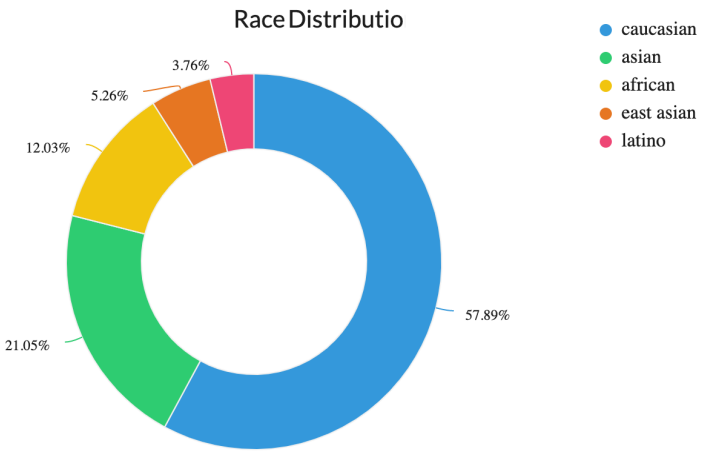


Figure 4.8: Distribution of race of the participants present in the Mood-Seasons data set.

participants were required to record videos every week. Most of the participants completed the study successfully by uploading one video per week for three weeks, some added more videos, while some recorded only one video.

There were several pre-processing operations for refining the data set, such as excluding low-quality, empty or invalid videos and removing participants who did not consent to sharing their data. The Mood-Seasons data set comprises 375 audiovisual recordings and transcripts from 134 unique participants. The average number of clips recorded by a participant is 2.79. The total duration of the recordings in the data set is about 9.8 hours or 556 minutes. The average duration of a recording is approx. 90 seconds or 1.5 minutes. The duration of the shortest clip is 60 milliseconds and longest clip has a duration of 6.13 minutes.

Figure 4.10 shows the distribution of the time span of the video recordings present in the Mood-Seasons data set. The high number of recordings that are above 60 seconds reflect the participants' adherence to the advice given during the study, i.e., to record a video of that lasts between 1 and 5 minutes.

A longitudinal analysis of the PHQ score is shown in Figure 4.11.It shows the PHQ-8 scores for each participant for all their recordings. The main observation is that the relative PHQ-8 scores over the course of three weeks remained stable, that is, within the category brackets, showing the longitudinal persistence of mood. This is

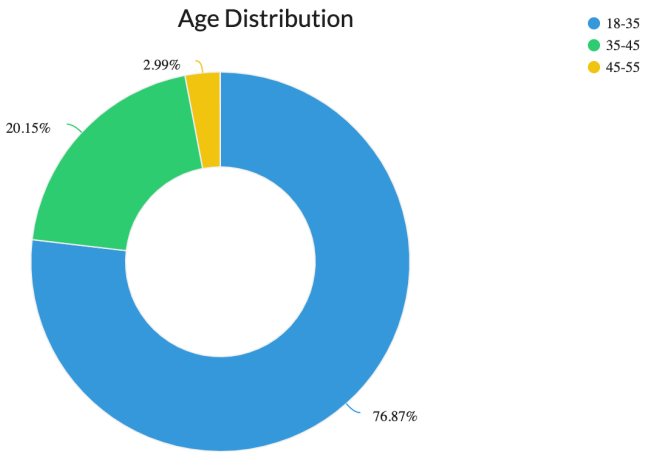


Figure 4.9: Distribution of age of the participants in the different age-groups present in the Mood-Seasons data set.

especially true for those who surpass the clinical threshold for depression, while the mild and no depression categories may alternate, showing that it is more natural to shift from a high mood to a lower energy phase or vice versa.

The study had two main limitations. First, the participants were not representative of the clinical population because people with a clinical diagnosis of depression were excluded. As a result, the study sample had a lower percentage of participants with severe depression symptoms, as it represented the general population. Second, the majority of the participants were young women. This means that the study results are more generalisable to young women than to the general population.

4.5 Conclusion

The fourth chapter described a large-scale data collection study that generated a novel, multimodal (audio-video-text) and longitudinal Mood-Seasons data set, which was acquired in natural, in-the-wild environments using a smartphone. The data set includes video and audio recordings, as well as textual transcriptions, from the general population, with depression severity measured from responses to a PHQ-8 questionnaire. The Virtual Human Questionnaire Study revealed that self-administered questionnaire responses are comparable to human- or virtual-human-mediated questionnaire responses.

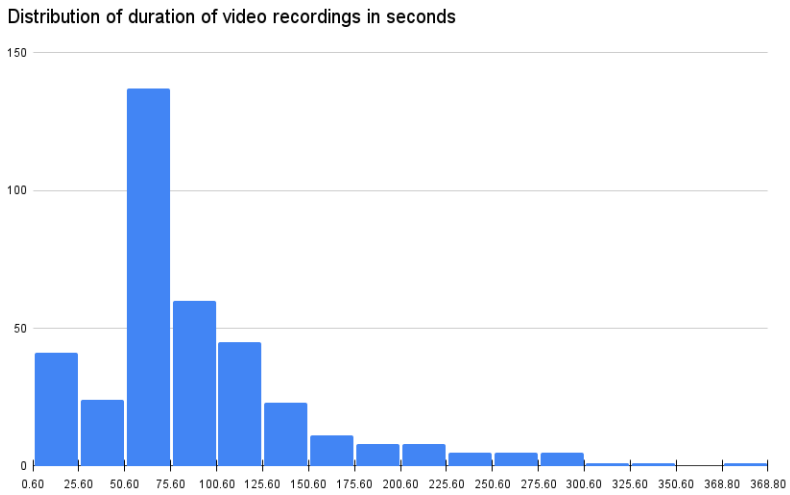


Figure 4.10: Distribution of the duration of video recordings in seconds of the participants present in the Mood-Seasons data set.

The chapter discussed the thorough and ethical design of the data gathering methods, as well as the development and implementation of the app. It also examined the lessons learnt by launching such an app, as well as the public's perspective of it. It also describes strategies to increase engagement and use of smartphone-based mental health data collection techniques.

The mood season data set was analysed in terms of the distribution of depression severity scores across numerous factors such as depression categories, age, gender, and race. A key limitation of the study is that the participants were from the general population and did not represent people with moderate-severe clinical depression. In Chapter 5, we will discuss a detailed approach for automated video analysis for depression assessment.

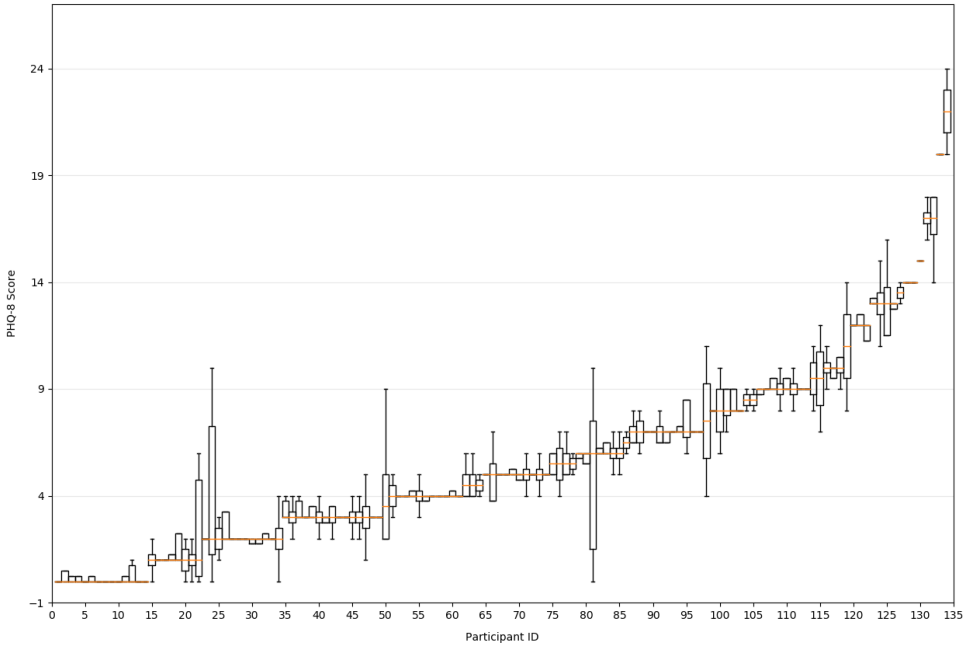


Figure 4.11: PHQ-8 scores for each of the participants for all their sessions. The orange lines on the box plot indicate the median score per participant in the Mood-Seasons data set.

Chapter 5

Depression Recognition

Observable traits of depression as stated by the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (APA) (Association et al., 2013) include both visual (facial expression and demeanour) and speech (increased pauses, muteness) indicators. Facial expressions, along with speech, are prominent behavioural observations that are strong indicators of mood disorders (American Psychiatric Association et al., 2013) including depression. Many approaches in the past have used different cues for detecting depression and other mood disorders, such as facial expression, gaze, head movement, body-pose (Pampouchidou et al., 2017a). Voice and speech analysis is used as a reliable means of estimating and tracking mood disorders (Cummins et al., 2015b), with studies having established accuracy through clinical trials.

An accurate characterisation of facial behaviour that can assess mood in real-time can be used as a reliable sensor in mental health technologies for managing mood disorders. This would open more opportunities to deliver behavioural interventions based on multimodal, audio, vision, and language data, prompting seamless user engagement during video sessions. This chapter presents approaches to the behavioural analysis of video recordings for detecting depression from video, audio, and text data in natural environments. The proposed research may be applied to the delivery of mental health care on automated patient monitoring or therapy administration platforms.

Many published studies addressing the problem of mood analysis for mental health disorders point out the difficulty in attaining labelled data at a large scale, (He et al.,

2018) mainly attributed to the clinical expertise needed and its sensitive nature. This makes most studies resort to collecting their own data sets in the laboratory, where most of the available data are in the form of clinical interviews with a limited number of subjects and in restricted clinical settings. It limits addressing the detection of mood disorders using facial behaviour in previously unseen environments or in-the-wild. To develop a system with real-world impact, it must be able to assess depression in natural environments. Chapter 4 presented a novel large data set that was collected using a smart phone in unconstrained real-world environments. This chapter will analyse the real-world data set using state-of-the-art and novel techniques to characterise the severity of depression.

This chapter is divided into two parts. The first part provides a comprehensive benchmark of state-of-the-art video analysis techniques on the newly curated MoodSeasons and publicly available AVEC 2014 data sets. The second part presents a multimodal transformer-based framework for automated depression severity prediction and includes extensive experiments to validate the effectiveness of the approach on both the Mood Seasons and AVEC 2014 data sets.

After presenting a thorough benchmark on the MoodSeasons data set, this chapter introduces a multimodal framework based on the multimodal transformer architecture for automated depression severity estimation. This section describes the motivation behind the use of multimodal transformers to learn strong representations from different sources of data, such as video, audio, and language. The section also discusses the two-stage design of the framework, which includes short-range and long-range analysis of multimodal sequences. It is followed by the introduction of a novel loss, termed differential loss, which helps leverage multiple videos from the same person to improve their prediction score for depression.

5.1 Bench-marking Automated Depression Analysis

There are several state-of-the-art video analysis methods for automated video prediction tasks, such as video activity recognition and predicting future states (Contributors, 2020). This section describes the main baselines employed for bench-marking automated depression recognition tasks on visual data or videos.

5.1.1 Datasets

The experiments in this study were conducted on the newly introduced Mood-Seasons dataset described in the Chapter 4 and a publicly available dataset for depression analysis, the AVEC 2013 dataset (Valstar et al., 2013). This section gives an overview of the two datasets and the pre-processing techniques applied for bench-marking. The AVEC 2013 dataset is included to show how the benchmark approaches compare with a well-studied dataset that is widely used in research on depression analysis.

The AVEC 2013 dataset is a widely used dataset for research on depression analysis and was part of the depression recognition challenge in AVEC 2013 (Valstar et al., 2014, 2013) that included 150 videos with depression labels. These videos were recorded in a controlled lab environment during task-based experiments, and the subjects' audio and video were recorded while they performed various tasks, such as speaking out loud while solving a task or sustained vowel phonation. The duration of the recordings ranged from 20 to 50 minutes, with a mean duration of 25 minutes, and the total duration of the clips was 240 hours. The subjects' mean age was 31.5 years and the original resolution of the videos was 640×480 pixels.

The experiments involving multimodal transformers employ part of the AVEC 2014 (Valstar et al., 2014) dataset. The AVEC 2014 dataset is a subset of the AVEC 2013 dataset that focusses on two tasks, the Northwind and Freeform tasks only. The North-Wind task involved reading a passage in German out loud, while the Freeform task involved answering open-ended questions in German, such as "What is your favourite dish?" or "What was your best gift?". To make the nature of the text modalities comparable to that of the Mood-Seasons dataset, the Freeform partition of the AVEC 2014 dataset was chosen for the experiments. This partition also includes 150 audiovisual video clips, but the duration of the clips is shorter than in the AVEC 2013 dataset, ranging from 6 seconds to 4 minutes and 8 seconds. For both datasets, each clip is labelled with a Beck Depression Inventory (BDI II) score, indicating the severity of depression, which ranges from 0 to 63.

As described in the previous chapter, the Mood-Seasons dataset comprises real-world videos collected from 134 participants. The training set of the Mood-Seasons

dataset for benchmarking experiments consists of 226 videos, while the validation set has 69 videos, and the testing set contains 79 videos. There are 87 unique subjects in the training set, 22 and 25 in the validation and testing sets, respectively. The dataset was split using a stratified mode where the distributions of dataset attributes such as gender, race, age, as well as the distribution of PHQ score were balanced. Figure 5.1 depicts the distribution of these factors in different dataset splits.

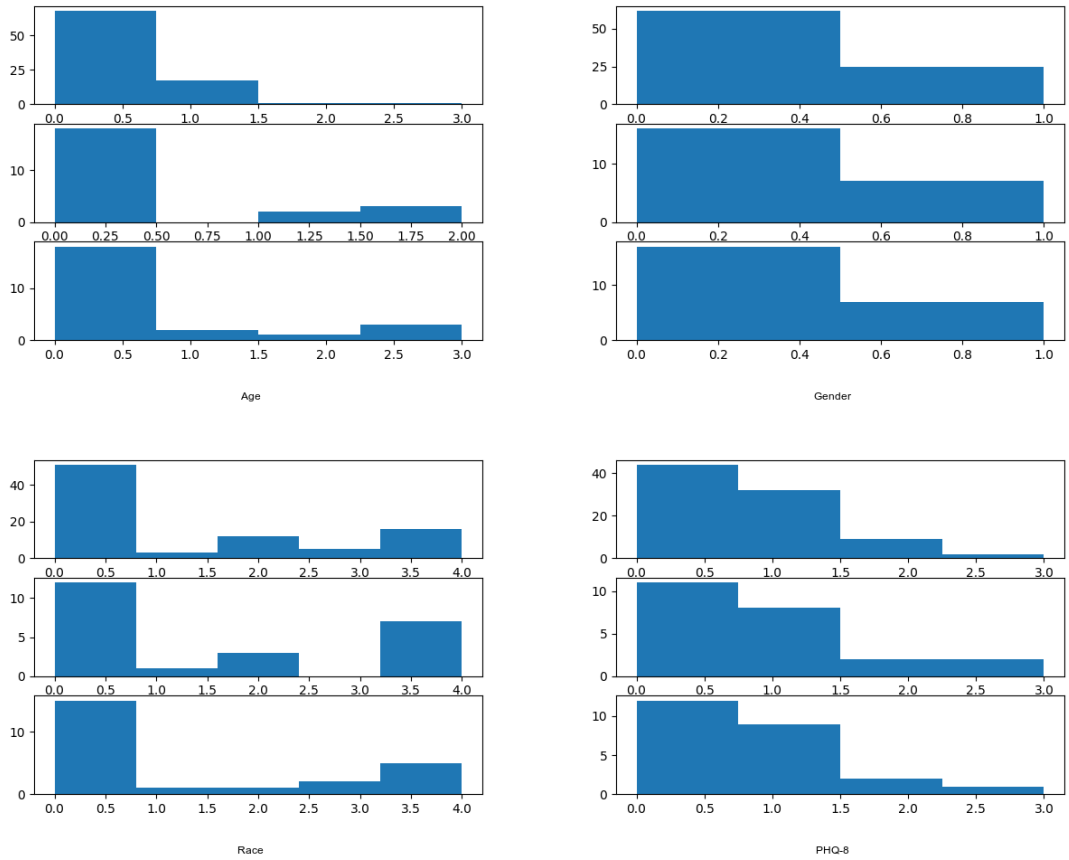


Figure 5.1: Distribution of different attribute categories, age, gender, race and PHQ scores in the three different dataset splits, training, testing, and validation (top to bottom rows in each plot) respectively

5.1.2 Evalutation Metrics

The objective of the depression recognition model developed in this work is to predict a continuous value representing the PhQ-8 depression severity score for

a given input video. In the literature, specifically in the renowned Audio Visual Emotion Challenge (AVEC) for depression recognition (AVEC 2013-19) Valstar et al. (2016), the approaches are evaluated using two measures: the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). These metrics are useful for the estimation of depression severity due to their ability to quantify the deviation between the predicted and true depression severity scores Valstar et al. (2013). Some recent studies Song et al. (2020) have also utilised Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) as evaluation metrics, which will also be included in this report. The mean absolute error or MAE is the average absolute error between the predicted and ground-truth PhQ-8 scores. It is given by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

The Root Mean Squared Error (RMSE) is a measure of the difference between the predicted and ground-truth depression severity scores. It is calculated by taking the square root of the average squared errors between the two. The formula for RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

5.1.3 Benchmarks

Given an input video $X_i \in V$, with a segment X_k comprising a set of frames $X_k = x_1, x_2, x_3, ..., x_n$ where N is the number of frames in the segment X_k and K is the number of segments in X_i, the goal is to predict a depression severity label y_i for the video X_i where $y_i \in [0, 24]$. The number of frames is $N = 16$ and the resolution of each frame is set to $x_{i,k} = 224 \times 224 \times 3$ for the benchmarking experiments. The value of K varies as the lengths of the videos are different. Only the segment length, that is, sequence of frames, is kept constant with a stride, $s = 2$ covering 1.07 seconds per segment. The batch size was set to $B = 8$.

The same settings were used on both the Mood-Seasons dataset and AVEC 2014. The video analysis techniques utilised for bench-marking compute either frame-level or segment-level predictions, and the final video-level prediction is computed as an

average of the predictions following a late-fusion approach.

The baseline models are classified into three types based on how they process these video data (i) 2D CNN approaches operate on the 2D frames only; (ii) 2D CNN + temporal methods additionally process temporal information from the 2D CNN features computed over temporal frames using sequential or attention mechanisms; (iii) the 3D CNN methods operate on 3D data directly, i.e., process spatial and temporal information simultaneously and includes established spatio-temporal frameworks such as 3D Resnet, I3D, C3D, TSM. The benchmark also includes a state-of-the-art video recognition network, slow-fast that processes video frames in separate slow and fast pathways. Here, we describe each of these baselines and its components in detail.

Depression prediction network: The depression recognition component of the benchmark comprises two parallel heads, one for the depression severity score regression and the other for classification. The first MLP head has two layers, one linear layer that takes an input 2048 and outputs 128 followed by a ReLu activation layer, then a linear layer that takes a 128-dimensional vector and outputs a depression severity score. The classification head has a similar architecture, except for the sigmoid activation at the end for BCE loss for classification.

Resnet 50 (He et al., 2016a): The Resnet 50 architecture (He et al., 2016a) is the chosen backbone architecture for extracting visual features. Deep convolutional neural networks are extremely successful in solving hard problems such as object category recognition in the wild, and the Resnet architecture helps scale CNNs, where the skip connections allow deep networks. Here, the Resnet 50 that is pre-trained on ImageNet dataset is further fine-tuned on VGG-FACE2 and FERA dataset for expression recognition to extract visual features relevant to facial behaviour. As a simple baseline, the Resnet 50 features are forwarded to the depression prediction network detailed above.

A set of (B × N) × C × H × W frames are fed into the Resnet network where a feature vector of (B × N) × 2048 dimensions is reshaped into a B X (2048N) vector and passed to the Depression network. All frames are used in this dense modelling approach but without modelling any temporal dynamics.

Resnet 50 GRU (Chung et al., 2014): This model allows for explicit comprehension of temporal information using a variant of a class of neural networks known as the recurrent neural network (RNNs), the Gated Recurrent Unit. RNNs are state-of-the-art in time-series modelling and prediction tasks and are popular in audio processing, speech recognition, weather prediction, etc. Originally designed for sequence-to-sequence modelling (an important NLP problem), RNNs are able to model the temporal context provided by the input frames, processing them sequentially using intermediate context representations to update the current temporal state (Chung et al., 2014). Gated Recurrent Unit is a type of RNN that uses gating mechanisms to manage the temporal information flow between units or cells that are similar to LSTMs but more memory efficient (Chung et al., 2014). This baseline takes the B × N × D dimensional visual features from the pre-trained Resenet-50 network from (B × N) × C × H × W input frames and forwards it to two GRU layers that aggregates temporal context sequentially over N frames and provides the output of the last hidden state as the aggregated features of dimension B × D_{GRU} for the input video clip, where $D_{GRU} = 256$. This temporal aggregate feature is then passed as input to the depression prediction network for the final prediction.

Resnet Attention The attention layer is an aggregation method that is permutation invariant, does not take into account the sequential ordering of frames as opposed to the Recurrent layer but computes a weighted attention between them. The Attention layer gets the input feature vector of size $B \times N \times D$ from the input video clip and forwards it to an attention layer, (Vaswani et al., 2017) which performs the attention operation on them to get an attention feature map of size $B \times D$. This is then passed to the depression recognition network detailed above.

We have seen that 2D CNNs are powerful in representing spatial information from individual video frames but lack the ability to model temporal dynamics or motion patterns from a set of video frames. Several architectures have been developed to address this gap for processing temporal information in CNN which are known widely as 3D CNNs, capable of incorporating spatio-temporal data from an input video clip. These 3D-CNN architectures lack recurrent layers, instead rely on 3D convolution (3D-Conv) and 3D pooling operations to retain temporal information of input sequences that would otherwise be lost in canonical 2D convolutions (Tran et al., 2015). This is especially beneficial as they can facilitate learning spatio-temporal

information early in the layers (Tran et al., 2015) as opposed to plugging in a temporal context learner (GRU/Attention as above) on top of the 2D feature maps. Here we consider several 3D CNN architectures that are widely employed in state-of-the-art video analysis tasks.

3D Resnet (Tran et al., 2018) - R(2+1)D: This spatio-temporal variant of a 3D-Resnet was proposed in (Tran et al., 2018). The main feature of this architecture is that it is composed of "(2+1)D" residual convolution blocks which factorises a 3D convolution into two separate and consecutive operations, a 2D spatial convolution and a 1D temporal convolution. The same 2+1 D convolutional blocks are reused across the network where the spatial and temporal convolutions alternate. Two advantages of this method were shown to be (i) ability to represent more complex functions through added nonlinearities and (ii) facilitates easier optimisation (validated by lower training errors). Resnet-50 variant of the R(2+1)D network architecture is used to process $B \times N \times C \times H \times W$ where $N = 16$, $C = 3$, $H = 224$,$W = 224$. The output of the network is a feature vector of dimension $B \times 1 \times D \times D$ which is then reshaped to a 1-D feature and forwarded to the depression prediction network for predicting the depression severity label and group.

I3D (Carreira and Zisserman, 2017): The I3D model architecture was introduced by Carreira and Zisserman (2017) for video action classification. The main feature of this model is that it takes a 2D CNN architecture and simply transforms it into a 3D CNN by inflating all the filters and pooling kernels. For instance, a 2D convolutional filter of size $N \times N$ becomes $N \times N \times N$. The advantage of this method is that a pretrained 2D CNN such as Resnet-50 can be converted into a 3D CNN by repeating the weights of the network modules by N times along the temporal dimension and rescaling by $1/t$ where t is the number of input frames. This initialisation produces the same output as the 2D pre-trained model run on a single frame (Carreira and Zisserman, 2017). The I3D variant used for the experiments is initialised from the pre-trained Resnet-50. The features of I3D are passed to the depression prediction network.

C3D (Tran et al., 2015): C3Ds are deep three-dimensional convolutional neural networks with a uniform design that consists of $3 \times 3 \times 3$ convolutional kernels followed by a 2 x 2 x 2 pooling at each layer. All video frames are resized to the

size 224×224. The input dimensions are $B \times N \times C \times H \times W$. The network has 5 convolution layers and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), the output of the final pooling layer of dimension $B \times 256$ is passed to the depression prediction network for getting the final labels.

TSM (Lin et al., 2019): the Temporal Shift Module (TSM) was developed for high accuracy and low computation cost video understanding (Lin et al., 2019). TSM is a generic and effective module that can be inserted into 2D CNNs to achieve temporal modelling, with no additional computation cost or parameters. TSM enables learning of temporal concepts, which 2D networks cannot model. Inspired by traditional convolution operations, the TSM shifts activation in a video model along the temporal dimension for information fusion from neighbouring frames. Similarly to a 2D CNN TSM takes an input clip of shape $B \times C \times N \times H \times W$ and returns features of size $Bx256$ which are then forwarded to the depression prediction network.

SlowFast (Feichtenhofer et al., 2019): The SlowFast network architecture for video recognition combines spatial and temporal streams by providing each path with the raw video, but at different temporal rates. The slow pathway captures spatial semantics, while the fast pathway captures fast and fine motion. The SlowFast network achieves strong performance for both in many video recognition tasks. The input to the SlowFast network is the same input clip that is sampled at different frequencies. As in the other architectures above, the output features are passed to the depression prediction network. The default parameters were retained for training SlowFast network.

The results presented in Table 5.1 and demonstrate the performance of the benchmark models in the validation and test sets of Mood-Seasons, respectively. A range of metrics, including mean absolute error (MAE), mean squared error (MSE), concordance correlation coefficient (CCC), and Pearson correlation coefficient (PCC) were used to evaluate the performance of the models. It is noteworthy that models incorporating temporal understanding, such as those using gated recurrent units (GRUs), attention mechanisms, or 3D convolutions, outperformed the baseline model, which lacks an understanding of temporal dependencies in the videos.

In general, the C3D and TSM models exhibit competitive performance on the Mood-

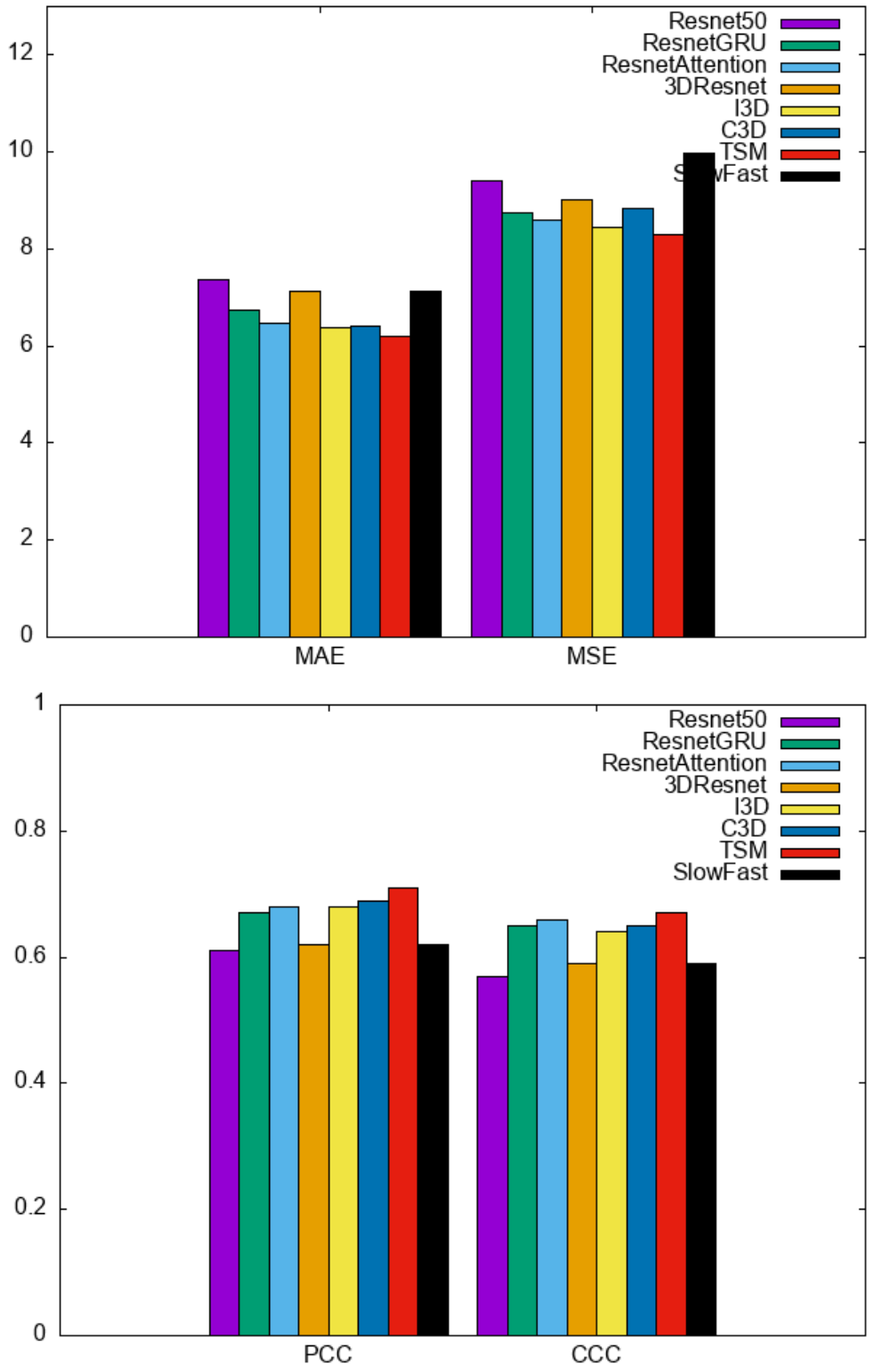


Figure 5.2: Performance benchmark of different video analysis approaches on AVEC2013 validation set. Note that the range of BDI scores in the AVEC 2013 dataset is from 0-63.

Method	MAE↓	RMSE↓	PCC↑	CCC↑
Chance-level	4.4	16.3	-	-
Resnet-50	3.68	4.60	0.29	0.25
Resnet-GRU	3.37	4.46	0.41	0.33
Resnet-Attention	3.35	4.28	0.42	0.32
3D-Resnet	3.53	4.57	0.35	0.30
I3D	3.55	4.45	0.32	0.27
C3D	3.35	4.28	0.43	0.32
TSM	3.44	4.39	0.30	0.25
Slow-Fast	3.69	4.63	0.22	0.18
Resnet-50	3.92	4.83	0.29	0.27
Resnet-GRU	3.62	4.46	0.32	0.30
Resnet-Attention	3.60	4.41	0.30	0.28
3D-Resnet	3.82	4.85	0.28	0.23
I3D	3.86	4.54	0.14	0.13
C3D	3.65	4.60	0.27	0.23
TSM	3.64	4.47	0.30	0.25
Slow-Fast	3.72	4.87	0.22	0.18

Table 5.1: Quantitative comparison of different baselines and ablation studies for Mood-Seasons validation set (top) and test set (bottom).Note that the range of PhQ scores in the Mood Seasons dataset is from 0-24.

Seasons dataset. Similar trends were also observed in experiments on the AVEC 2013 dataset. The results are shown in Table 5.2 and Figure 5.2 However, the SlowFast model (Feichtenhofer et al., 2019), which is a state-of-the-art approach for video activity recognition, did not perform as well compared to other 3D or temporal models. Among the 3D models, some, such as I3D and 3D Resnet, demonstrated relatively lower performance, while others, such as those utilizing aggregation methods like GRUs and attention mechanisms, performed comparably.

As shown in Table 5.2, TSM has the best overall performance, followed by C3D, I3D, and Resnet-Attention. The other methods perform slightly worse. To test whether the difference in performance between the methods is statistically significant a paired t-test can be used. The following table shows the results of the paired

Method	MAE↓	RMSE↓	PCC↑	CCC↑
Resnet-50	7.38	9.4	0.61	0.57
Resnet-GRU	6.75	8.75	0.67	0.65
Resnet-Attention	6.46	8.6	0.67	**0.68**
3D-Resnet	7.12	9.02	0.62	0.59
I3D	6.39	8.46	0.68	0.64
C3D	6.42	8.33	0.69	0.65
TSM	**6.2**	**8.31**	**0.70**	0.67
Slow-Fast	7.12	9.98	0.62	0.59

Table 5.2: Quantitative comparison of different baselines for AVEC 2013 Validation set.Note that the range of BDI scores in the AVEC 2013 dataset is from 0-63.

t-test comparing TSM to the other methods on the AVEC 2013 data in Table 5.3:
The p-value is less than 0.05 for all of the methods except for C3D and Resnet-

Table 5.3: Statistical significance of the difference in performance between TSM and the other methods in AVEC 201.

Method	p-value
C3D	0.59
I3D	0.23
Resnet-Attention	0.72
3D-Resnet	0.03
Resnet-GRU	0.001
Resnet-50	0.0001
Slow-Fast	0.00001

Attention. This means that the difference in performance between TSM and these methods is statistically significant. Based on the above statistical analysis, we can conclude that TSM has the best overall performance on the AVEC 2013 validation set. The difference in performance between TSM and the other methods is statistically significant, except for C3D and Resnet-Attention.
For the Mood-Seasons dataset, a pairwise t-test between TSM and the rest of the methods in the testing set is performed and results presented in 5.4. The p-value is less than 0.05 for all methods except for C3D and Resnet-Attention. This means that the difference in performance between TSM and these methods is statistically

significant. The TSM, C3D, and Resnet-Attention methods have shown promising

Table 5.4: Statistical significance of the difference in performance between TSM and the other methods on the Mood-Seasons test set.

Method	p-value
C3D	0.93
I3D	0.01
Resnet-Attention	0.70
3D-Resnet	0.005
Resnet-GRU	0.02
Resnet-50	0.0001
Slow-Fast	0.00001

benchmarks for depression recognition. In the Mood-Seasons test set, TSM achieved the best overall performance, followed by C3D and Resnet-Attention. The difference in performance between TSM and the other methods was statistically significant, except for C3D and Resnet-Attention. These results suggest that these methods may be useful for developing new tools for depression detection and monitoring.

5.2 Multi-modal Transformers For Audio, Visual And Language Fusion

This section presents a comprehensive description of the multimodal transformer framework designed for the task of depression estimation, including its motivation, architecture, design methodology, and objective functions. Most state-of-the-art methods for fusing audio-visual-language features rely on word- or utterance-level alignment. However, the estimation of the patient's health questionnaire (PHQ-8) score, a measure of the severity of depression, requires long-range inference of the subject's mood state, which is better captured by analysing their entire mood diary video.

MULT (Tsai et al., 2019) is an extensive applied audio-vision language latent representation fusion architecture based on transformers where the model is able to attend to different modalities of data simultaneously, allowing it to learn rich multi-dimensional representations of the input. It forgoes the need for precise temporal

alignment of individual modalities and instead exploits attention mechanisms for implicit intermodality alignment. The self-attention mechanism allows the model to attend to different parts of the input data simultaneously, allowing it to learn which parts of the input data are relevant to each other. This allows the model to align the different modalities of the data and to learn a coherent representation of the input.

The main attraction of this method is the flexibility it offers the modalities to attend to relevant and complementary information from freely available segments of other modalities.This results in the learning of multimodal behavioural windows, which capture the complex interplay between the different modalities. The ability to learn such multimodal representations is crucial for tasks that involve understanding the behaviour of individuals, such as the estimation of the severity of depression.

The transformer-based method facilitates the discovery of long-range dependencies in the modality-specific latent representations (Tsai et al., 2019) that are pivotal for comprehending depressive markers. Unlike continuous affect labels, such as valence and arousal, the characteristics of depression may occur at varying time steps within a video, and the temporal consistency of depressive markers cannot be guaranteed throughout the video. For instance, low mood markers such as the passive face, low pitch, long pauses between words, or even a subtle expression of depressive the mood may be spread across the video sequence. Using explicit word-level alignment may discard useful information, such as pauses or sighs, in an audio segment corresponding to a video sequence.

The ability of transformer architecture to tackle unaligned sequences makes it particularly well-suited for the task of recognising depression severity, compared to archetypal baseline models such as CNN+RNN that require explicit temporal alignment of modalities. Conventional utterance/word-aligned models often rely on extensive feature engineering to align modalities using time-steps, whereas the transformer approach addresses the problem more holistically, allowing the model to attend to information from multiple modalities in an unaligned manner and infer depression severity from long-range, multimodal behaviour-aligned windows, rather than pre-defined, narrow multimodal time-aligned windows.

This property of the transformer architecture is particularly important when analysing

depression, as context can play a significant role in the elicited facial expressions or mood states of participants. As stated in Chapter 4, the study's video prompts were designed to exclude potential emotional triggers to avoid obscuring the audiovisual markers with emotional responses to specific prompts. However, it was observed that the prompt "tell us about your pet" elicited higher arousal and positive valence in subjects with high depression scores. This highlights the importance of long-range modelling and the ability to parse unordered sequences in identifying sporadic occurrences of depressive markers in such scenarios.

5.2.1 Architecture

The transformers are capable of capturing dependencies between distant parts of the sequence and can process information embedded in long videos. The multimodal attention component of transformer architectures is crucial for learning from the different modalities we use in this paper: video, audio, and text.

Attention Mechanism. The transformer block, originally proposed in (Vaswani et al., 2017), consists of several multi-head attention modules that calculate a weighted representation of all other tokens in a given input sequence for each token embedding. This weighted representation is combined with the input representation of the given token and passed to the next layer.

Attention mechanisms enable the model to learn how to assign weights to different parts of the input based on their importance or relevance to the task at hand. This is typically achieved by computing the dot products between the different parts of the input and using these dot products to compute the weights for each part. These weights can be used to weight the corresponding parts of the input, resulting in a weighted sum that represents the output of the attention mechanism.

The idea behind this approach is that the dot products capture relationships between the different parts of the input, and the weights reflect the importance or relevance of those parts to the task. By learning these weights, the model can focus on the most relevant parts of the input and learn complex relationships between them, improving its performance on the task.

There are several methods to implement multimodal attention (Hendricks et al., 2021), including merged attention, modality-specific attention and co-attention. In merged attention, the tokens representing queries, keys, and values can come from any input modalities, i.e., the tokens belong to a joint pool of modalities. In the co-attention module that we use in this work following (Tsai et al., 2019), given the queries from one modality, for example, language, we compute the keys and values from another modality.

The overall multimodal transformer architecture consists of four components, a projection layer, multimodal transformer module followed by a self-attention transformer, and finally by a classification layer. The multimodal transformer block, shown in Figure 5.3 is the original transformer attention block proposed in (Vaswani et al., 2017) redesigned to include the attention mechanism for multimodal information fusion by Tsai et al. (2019).

Each multimodal transformer block has several co-attention layers that generate intermediate merged features for feed-forward fusion in the subsequent layers. The self-attention transformer consists of several feed-forward multi-head attention layers, as shown in the Figure 5.4.

After introducing the main building blocks of the architecture, we can now look at the different steps of using the transformers in the current work. Given two input modalities α and β, with respective sequences $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$,the queries, keys, and values are defined as follows:

$$Q_\alpha = X_\alpha W_{Q_\alpha}$$

$$K_\beta = X_\beta W_{K_\beta}$$

$$V_\beta = X_\beta W_{V_\beta}$$

where the weights $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$, and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$.

The cross-modal attention Y_α represents the fused latent representation or adaptation from modality β to modality α and is defined as follows:

$$Y_\alpha = CM_{\beta \to \alpha}(X_\alpha, X_\beta) = \text{softmax}\left(\frac{Q_\alpha (K_\beta)^T}{\sqrt{d_k}}\right) V_\beta$$

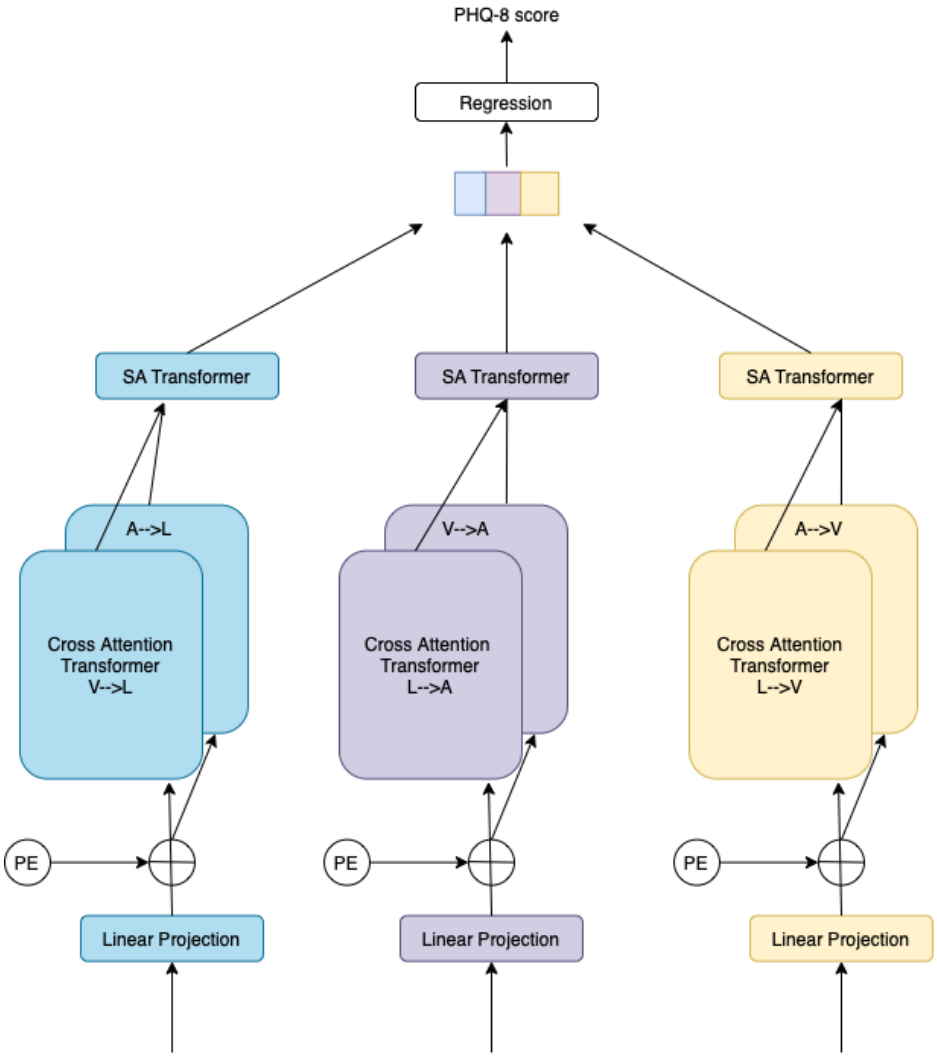


Figure 5.3: Architecture of the multimodal transformer network for audio, video, and language feature fusion.

where cross-modal attention Y_α represents a sequence of length T_α in the feature space of V_β. For each time step i in this sequence, the attention mechanism computes a score for each time step j in the sequence of modality β, indicating how much attention should be given to information in time step j when generating the representation at time step i. The score for each time step j is determined by the dot product of the query and key at time steps i and j, respectively, and is normalised by the length of the key d_k. The final representation at time step i is obtained by taking a weighted sum of the values at all time steps j in the sequence of modality β, where the weights are determined by the softmax of the scores.

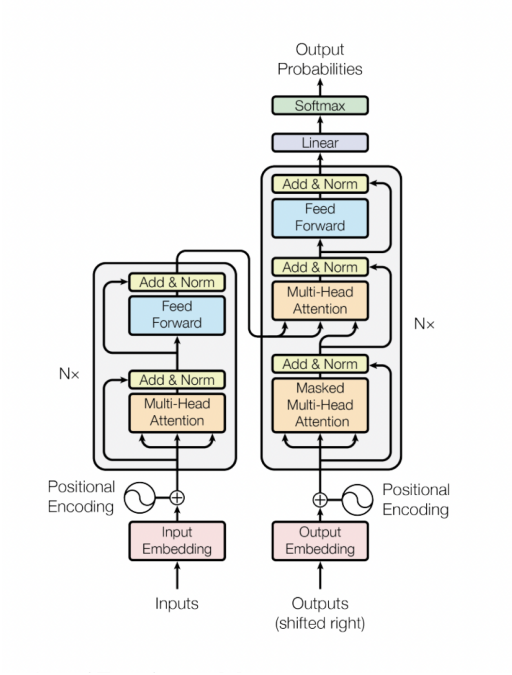


Figure 5.4: Architecture of the multi-head attention layers in the transformer block (Vaswani et al., 2017).

Co-attention Transformer Block. Several research problems, such as visual question answering, video captioning, and modality alignment involve an amalgamation of information from different input modalities. A well-studied solution is to employ co-attention modules in the highly influential transformer architecture in such multimodal settings (Hendricks et al., 2021). Refer to Figure 5.5.

A pairwise attention module for a given modality aims at generating multimodal abstractions from another input domain. For instance, in order to incorporate information from the language stream with respect to the visual stream, the approach takes query latent Q_v from visual features and keys K_l and values V_l from the language stream. Such an approach conditions the language attention on the visual stream, i.e. it attends to features from the language domain that are most relevant to the given sequence of visual features. In the case of three modalities, namely, audio, language, and vision, the multimodal transformer block also includes an attention module that generates language attention features conditioned on the audio features from its keys K_l and valuesV_l and the query matrix Q_a from the audio domain. Since we consider three modalities for identifying the markers of depression, there are six

co-attention transformer blocks to incorporate possible permutations of pairwise co-attention.

Self-attention transformer block. The self-attention transformer block is a stack of multihead attention layers, as shown in Figure 5.5. The multimodal transformer block provides the merged attention features from the three modalities to the self-attention block. Each self-attention block attends to each of the combined attention pooled features from the target domain.

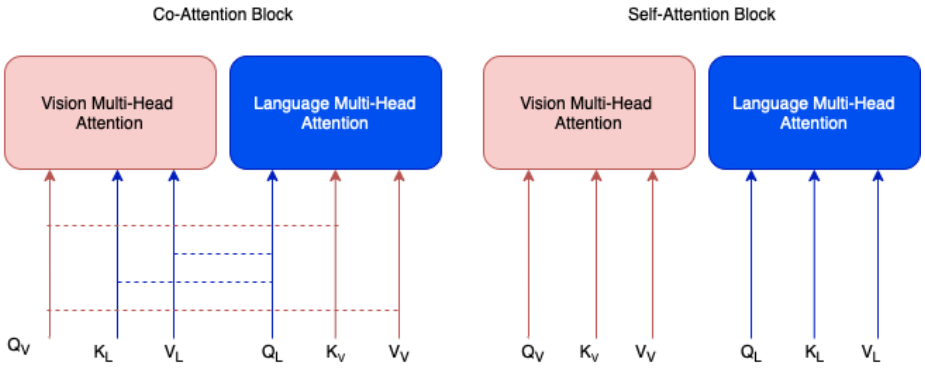


Figure 5.5: A co-attention and self attention block visualised.

An extensive study (Hendricks et al., 2021) on several methods of attention mechanisms used in transformers for multimodal fusion reports that co-attention and merged attention provide meaningful multimodal representations that lead to the success of multimodal transformers compared to using only modality-specific multi-head self-attention modules. We report our findings from the experiments comparing the co-attention and modality-specific self-attention modules in the experiments section.

The multimodal depression analysis framework makes two architectural design choices: sentence-level modelling of depression and video-level modelling of depression. The decision to use sentence-level modelling as a short-range modelling approach was based on the desire to leverage the implicit alignment property of transformers over unaligned sequences (Tsai et al., 2019). This allows for a more comprehensive analysis, as a sentence is self-contained and typically refers to a single context and includes non-verbal cues such as pauses, transitions, etc. which can then be part of the model's input.The framework then employs a light-weight transformer

to aggregate video-level understanding based on the learnt representations from the sentence-level modelling. The video level multimodal transformer architecture follows a light-weight structure with a single multi-head, self-attention head and aggregates sentence level multimodal embeddings. This allowed a more comprehensive understanding of the relationship between depression and the multimodal data.

The following section describes the multimodal data processing pipeline, the feature extraction steps for each of the modalities, such as video, audio, and text, and the training objectives of the proposed framework.

5.2.2 Dataset Preprocessing

The Mood-Seasons dataset consists of videos that include images and audio from the mood diaries recorded by the participants. To prepare the dataset for multimodal learning, the following pre-processing steps were undertaken.

1. The first step was to transcribe the audio files in the dataset using the Google automatic-speech-recognition (ASR) API. This API returns word-level transcripts with time stamps for each word, allowing us to determine when each word was spoken in the audio. For the AVEC 2014 Freeform dataset, the transcription was done in German. However, some discrepancies were noted in the transcripts, particularly for videos featuring participants with accents that were difficult to understand and had to be manually corrected.

2. The second step involved pre-processing the transcripts to retrieve sentence-level transcripts with time stamps. Word-level transcripts were combined into sentences using timestamps from punctuation.

3. The third step encompassed cropping the audio and video files according to the sentence-level time stamps obtained from the Google ASR. This allowed us to create sentence-level clips that were aligned according to the derived time stamps for audio, video, and text. These sentence-level clips were then used for further analysis.

The Mood-Seasons dataset was pre-processed to generate a total of 1806 sentence-level, multimodal clips after removing any invalid sequences. The clips were split into training. validation and testing splits with 1183, 287 and 333 clips, respectively. For

the AVEC2014 Freeform dataset, there were 106 clips for training and 106 sequences for testing. From these clips, various multimodal features were extracted from the video, audio, and text modalities using techniques described in the subsequent sections.

5.2.3 Video Feature Extraction

Visual features for the multimodal transformer framework were encoded by the dense frame-level Resnet50 baseline model described in the benchmark section 5.1.The Resnet50 model extracts dense visual features corresponding to facial attributes from all the available frames from the input video clip of the shape $B \times N \times C \times H \times W$. The idea is to provide the multimodal fusion transformer with dense low-level information where all video frames are included in the representation without information loss.

For feature extraction, each input video clip representing a sentence is forwarded to the pre-trained Resnet50 model for depression severity estimation which generated video features of the size $B \times (N_{video} \times N_f) \times D_{video}$ where N_{video} is the number of sequences in the video with the number of frames $N_f = 16$ each and has feature dimension $D_{video} = 2048$. The N_{video} varies according to the length of the sentence clip, and was set to have a maximum of 135 clips for the Mood-Seasons dataset and 150 for the AVEC2014 Freeform dataset. When the N_{video} was less than the maximum set value, the sequences were zero-padded. For instance, the input feature representation for the visual modality will be $8 \times 2160 \times 2048$.

5.2.4 Audio Feature Extraction

Audio features for the multimodal framework were extracted using a popular audio classification network known as VGGish (Hershey et al., 2017). VGGish is a widely used pre-trained CNN released by Google (Hershey et al., 2017) for audio processing tasks such as recognition and classification, whose architecture was inspired by VGG networks designed for image classification. The architecture consists of a series of 17 convolution layers and activations, followed by max-pooling. The VGGish network operates on log-mel spectrogram representations of the audio clips.

Once the audio clippings corresponding to the sentences were obtained, it was

pre-processed to generate the spectrograms required for the audio model. The audio wav file was normalised to the range $[-1, 1]$. If the audio data has two channels, the channels are averaged to produce a single channel. The data are then resampled to a lower sample rate for computational efficiency. It is cropped into overlapping windows and a Hamming window is applied to reduce spectral leakage and smooth the data. A Fast Fourier Transformation is applied to calculate the power spectrum of the data, and frequencies outside a specified range are filtered out. Mel Frequency Filter Banks are applied to transform the data into the Mel frequency scale, which is more closely aligned with human perception of sound. Finally, the natural logarithm of the resulting values is taken to reduce the dynamic range of the data.

The data is then divided into nonoverlapping frames and decomposed using a short-time Fourier transform with a window size of 25ms and a frame shift of 10ms. The resulting spectrogram is transformed into 64 Mel-spaced frequency bins and log-transformed to add a small offset and avoid numerical issues. These log-mel spectrogram patches are used as input for VGGish. This process produces an array of shape (96, 64) from an input of 975 ms of audio data. The dimension of the features returned by VGGish is $B \times N \times D_{audio}$ where $D_{audio} = 128$. The value of N is variable depending on the length of the audio clip corresponding to the sentence. The maximum number of frames for VGGish audio features for each sentence is set to 75. The size of audio features per sentence-level clip is $B \times N_{audio} \times D_{audio}$, where $D_{audio} = 128$ and $N_{audio} = 75$. As mentioned above, the size of N varies according to the length of each audio clip corresponding to a sentence, which is then zero padded to produce a vector of features of N frames.

5.2.5 Language Feature Extraction

The sentence level transcripts obtained from the automatic-speech-recognition module are encoded into language features for the multimodal transformers using the well-known BERT (Bidirectional Encoder Representations from Transformers) model. BERT is a state-of-the-art NLP model that achieves high performance on several language downstream tasks such as translation, sentiment analysis, generation, etc. The BERT model attends to bidirectional context, that is, it combines both the left and right context of each word for its self-supervision task and was shown to have excellent performance in context modelling tasks (Devlin et al., 2018). BERT was

pre-trained on large amounts of text data using the self-supervised objective of masked-language modelling, where the words in a sentence are masked out randomly and the model was tasked to predict the right word based on the context in the sentence. Unlike other language models that were trained unidirectional, i.e., only using the previous words to predict the next, BERT generated high quality contextualised representations. Therefore, it is an excellent choice for embedding the sentences for multimodal fusion in the context of depression analysis.

First, each sentence transcript is converted into a format compatible to BERT model. This involves converting the words in the sentence to a set of tokens, with a special [CLS] token appended to the beginning and [SEP] token appended to the end of the sentence. The tokens should correspond to the specific vocabulary used by BERT. Once the sentence is converted to tokens from the BERT vocabulary, the BERT model was used to extract the features from the sentence. There are several ways in which the features can be computed from the output of the BERT model, which is detailed in (Devlin et al., 2018). The output dimension of the BERT model is $B \times N_{text} \times L \times Dtext$ where $B = 1$, N_{text} correspond to the words or tokens in the sentence $L = 13$ represent the output layers and $D_{text} = 768$ is the size of the feature embedding. The feature vector representing the input sentence is computed by concatenating the latent vector from the last four layers of the model for each of the token. This gives a final output of $B \times N_{text} \times D_l$ where $D_l = 3072$. This approach has been shown to perform best for many downstream tasks since the initial layers of the BERT model does not have any context information learnt, but the final layer may be too specific to the pre-training task (Devlin et al., 2018). The maximum number of word tokens N_{text} is set to 220 for the Mood-Seasons dataset and 200 for the AVEC 2014 FreeForm dataset.

5.2.6 Training Methodology

Training of the proposed multimodal depression recognition framework has two stages, corresponding to different temporal granularities, that is, sentence level and video level. Once the input videos are pre-processed into shorter clips based on the sentence transcripts, unimodal networks are trained for the modalities audio and video separately. These uni-modal networks are used as feature extractors for the respective input modalities. The pre-trained BERT model is used directly for lan-

guage feature extraction. The sentence level features of the three modalities, namely, audio, video, and text from pre-trained unimodal models, are then combined using the multimodal transformer architecture described in section 5.2. Then a video-level light-weight transformer discussed in section 5.2 is used to aggregate the learnt multimodal sentence level representations and derive the final depression severity score.

Data augmentation is a popular technique in deep learning that can improve model performance by artificially increasing the size and diversity of the training dataset. It works by applying random transformations to existing data, such as cropping, scaling, rotating, and flipping. This helps to prevent overfitting, which is a problem that can occur when a model learns the training data too well and is unable to generalise to new data. Data augmentation is particularly important for video-based deep learning models, as video data can be very expensive and time-consuming to collect and label.

Several geometric and photometric augmentations were used in training the visual feature extractor model, Resnet50, for data augmentation and regularisation. These augmentations include random blur, crop, scale, rotation, masking, cut-mix, and colour jitter applied over the video frames. One observation was that Pytorch transforms applied different transformations to the frames in the same clip, which led to suboptimal training. This was resolved by using the Kornia library (Riba et al., 2020) for video augmentation resulting in uniform transformations being applied to a clip of 16 frames. These augmentations were also implemented in the GPU which sped up the training time considerably.

The proposed framework incorporated several hyperparameters to optimise the model's performance. For the visual feature extraction model, the learning rate, optimiser, and depth of the GRU units were impactful hyperparameters. A grid search was used to evaluate various learning rate values within the range of [1e-5, 1e-1], and the best results were achieved with a learning rate of 1e-3. The batch size was set to 8, the number of frames was 16. The model was trained for 60 epochs using a cosine annealing learning rate scheduler with 5 warm-up epochs.

In the unimodal audio model, the VGGish (Hershey et al., 2017) architecture was

utilised and the training hyperparameters were optimised. A learning rate of 1×10^{-2} was found to be optimal, in conjunction with the Adam optimiser, which was fine-tuned to 25 epochs. Two variants of fine-tuning or transfer learning were explored: fine-tuning the entire network and fine-tuning only the last 5 layers of the network (2 convolutional and 3 fully connected). The results of the experiment indicated that training all layers was necessary to achieve satisfactory performance on the depression recognition task.

For language feature extraction, the pre-trained BERT model was leveraged directly, rather than being fine-tuned for the task. This decision was made due to the observed poor performance of the fine-tuned version, which was likely due to the limited size of the available data relative to the capacity of BERT, a large language model. For the 2014 Freeform AVEC subset, a BERT model was used that was pre-trained in the German language.

The sentence level multimodal transformer architecture features 8 layers, with 5 multi-head attention heads and 512 hidden unit dimension. The batch size is set to 8 and the learning rate is set to 0.01 with Adam as the optimiser. The Multimodal transformer is trained to 35 epochs. The video-level multimodal transformer architecture follows a light-weight structure with only 1 layer, 1 multi-head attention head. The batch size is set to 1, with a learning rate of 0.01 and trained to 20 epochs.

Several minimisation objectives or loss functions were used to train the uni-modal and multimodal transformer models. Loss functions are discussed in detail in the following section.

MSE Loss

The Mean-Squared Error (MSE) loss function is commonly used in regression tasks and measures the average squared difference between the predicted value and the true value. In the context of estimating the severity of depression, the MSE loss function can be used to measure the difference between the predicted severity score of depression, y_{pred}, and the actual severity score of depression, y_{true}. Given a batch

of N samples, the MSE loss, $\mathcal{L}_{MSE}$, is calculated as follows:

$$\mathcal{L}_{MSE} = \frac{1}{N} * \sum (y_{pred} - y_{true})^2 \tag{5.1}$$

MAE Loss

The second loss function used is Mean absolute Error, which measures the absolute difference or error between the estimated depression severtity score y_{pred} and the ground truth severity score y_{true}. Given a batch of N samples or video clips, the loss $\mathcal{L}_{MAE}$ is computed as:

$$\mathcal{L}_{MAE} = \frac{1}{N} * \sum |y_{pred} - y_{true}| \tag{5.2}$$

Differential Loss

Different individuals may experience different symptoms of depression to varying degrees (Cohn et al., 2009b). For the videos recorded by an individual, any dissimilarity in its feature representations should arise from the difference in the corresponding PHQ scores and not from external conditions present in the video due to noisy surroundings which may include background images, sounds, lighting, etc. Here, the framework introduces the differential loss, where any difference in the PHQ score of an individual is also proportional to the dissimilarity in their feature representation. In this way there are no constraints on how the features should be represented across different individuals; however, for the same individual, the loss ensures that the differences in the features should only encode the differences in depression severity scores.

The loss can be used for training the Mood-Seasons dataset because it consists of longitudinal data from the participants to improve the prediction of depression. In other words, differential loss is simply a measure that enforces coherence between the multimodal features of videos, including facial data, voice, and speech content, of the same person in the Mood-Seasons dataset. By using multiple video samples from the same person, the analysis of their expression of depression can be more holistic and improve the prediction of depression in those videos.

To calculate differential loss, the framework compares the difference in the sever-

ity score labels of the Patient Health Questionnaire (PHQ), y_{ins} and y_{ref}, of two video sequences of the same identity, X_{ref} and X_{ins}, with the dissimilarity of the characteristics between the corresponding multimodal features of those sequences. This accounts for a differential signal for the network. The multimodal features are generated by the multimodal transformer G, which takes the input sequences X_{ref} and X_{ins} and produces multimodal feature representations $G(X_{ref})$ and $G(X_{ins})$, respectively.

The dissimilarity of features, $\mathcal{D}_{feat}$, is calculated using the cosine similarity between the multimodal representations of features, scaled by a factor of $\alpha = max(1, 2, ..24)$, which represents the maximum value of the PHQ-8 score range. Cosine similarity is a measure of similarity between two vectors and is defined as the cosine of the angle between them.

The differential loss is then calculated as the mean squared error between the difference in the PHQ severity score labels and the feature dissimilarity, given by:

$$\begin{aligned}
\mathcal{L}_{\text{diff}} &= \text{MSE}(\text{D}_{\text{feat}},\ \text{D}_{\text{score}}) \\
\mathcal{D}_{\text{feat}} &= (1 - \ \text{S}_{\text{C}}(\text{G}(\text{X}_{\text{ref}}),\ \text{G}(\text{X}_{\text{ins}})) * \alpha \\
\text{D}_{\text{score}} &= |\text{y}_{\text{ref}} - \text{y}_{\text{ins}}|
\end{aligned}$$

where MSE is the mean-squared error, where $\mathcal{S}_C$ is the cosine similarity. The mean squared error is a common loss function used in regression tasks that measures the average squared difference between the predicted values and the true values.

Metric learning Bellet et al. (2013) is a general approach to learning similarity or measures. Metric learning algorithms typically work by optimising a loss function that encourages the algorithm to produce embeddings for the data points such that the embeddings of similar data points are close together in Euclidean space, and the embeddings of dissimilar data points are far apart. Siamese networks and contrastive loss functions are popular metric learning approaches and may also be a suitable alternative to the differential loss proposed in this section.

5.3 Experiments

This section details the performance evaluation the proposed method. Several experiments were conducted to validate the efficacy of the proposed methods. The following sections present the datasets, evaluation metrics along with the experimental results and discusses each in detail. These results include comparisons with state-of-the-art approaches and several ablation studies.

5.3.1 Multi-modal Transformer Results

The results from the proposed multimodal transformer framework on the MoodSeasons dataset are presented in Table 5.5.

In comparison to the benchmark models outlined in section 5.1.3, the multimodal

Method	MAE↓	RMSE↓	PCC↑
Resnet-GRU	3.37	4.46	0.41
C3D	3.35	4.28	0.43
MMT	**2.62**	**3.43**	**0.56**
Resnet-GRU	3.62	4.46	0.32
TSM	3.64	4.47	0.30
MMT	**2.89**	**3.65**	**0.52**

Table 5.5: A comparison of different methods for depression severity estimation. The first block shows the results for the baseline and proposed approaches for the validation set of Mood Seasons dataset, and the second block shows the results for the testing set. Bold numbers indicate the best performance for each measure.

transformer architecture exhibits a significantly higher performance with a Mean Absolute Error (MAE) of 2.62, representing a 21% improvement from the best-performing spatio-temporal model, C3D, and a Root Mean Squared Error (RMSE) of 3.36, demonstrating a 15% improvement on the validation set. On the testing set, which appears to be slightly more challenging compared to the validation set, the multimodal transformer model achieved an MAE of 2.89 and an RMSE of 3.94,

showing a 19% and 11% improvement, respectively, compared to the top-performing benchmark, TSM, on the MoodSeasons dataset. These results demonstrate the superior performance of the proposed multimodal framework in predicting depression severity. Refer to Figure 5.6.

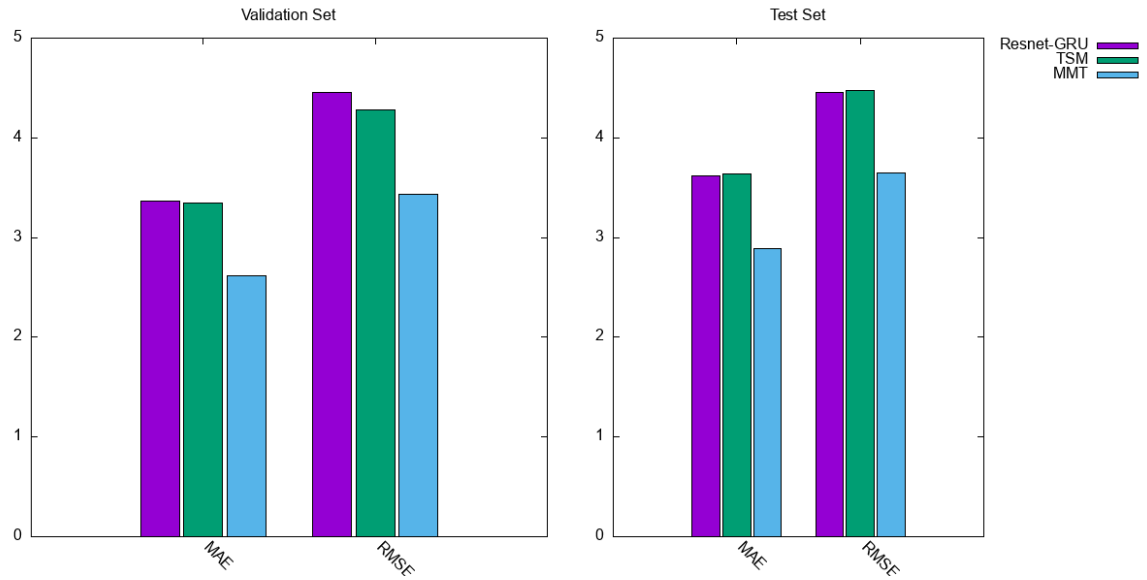


Figure 5.6: Performance benchmark of different video analysis approaches on MoodSeasons dataset

5.3.2 Comparison of Uni-Modal and Multi-Modal Approaches

This set of experiments compare the performance of various uni-modal and multi-modal methods for depression estimation. The following models are evaluated:

1. The Resnet-50 model, as unimodal approach for the visual modality
2. The VGGish model, as unimodal approach for the audio modality
3. Multimodal approaches that combine two modalities, including audiovisual, visual-language, and audio-language using (i)late fusion of two unimodal model combinations. Note that Language modality is excluded, as pretrained language features were not fine-tuned for depression recognition. (ii)variants of the proposed multimodal transformer approach that operate on only two modalities instead of three. This was achieved by modifying the multimodal transformer architecture to include only two modalities.

4. Multimodal approaches that combine three modalities, using either late fusion or the proposed multi-modal transformer approach.

Table 5.6 reports the results of comparing the above-mentioned unimodal (single modality) and multimodal (multiple modalities) models for the task of predicting depression severity scores on the Mood Seasons validation dataset. Performance evaluation includes results from three different evaluation metrics: mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (PCC).The modality of each model is indicated in the first column, with (V) representing visual modality, (A) representing audio modality, and (L) representing language modality. The second column lists the specific method used for each model. Note that the results are reported on sentence level clips, where the final prediction per video is an average of the sentence level predictions. Note that the evaluation of the above models are conducted on sentence-level clips, with the final prediction for each video being derived as the mean of the predictions made at the sentence level.

Overall, the multimodal transformer model (A-V-L) performs the best, with the lowest MAE of 2.81, RMSE of 3.78, and highest PCC value of 0.54. The unimodal visual model, Resnet-50, which is based on dense visual frames, performs better than the unimodal audio model, which is based on VGGish. This suggests the effectiveness of visual signals at estimating depression. A multimodal model that combines predictions from both visual and audio modalities using a late-fusion approach performs better than either modality applied alone, indicating that the fusion of the two modalities leads to improved performance.

Among the multimodal models using transformer architecture, the audiovisual transformer shows the best performance compared to the models using language modality. However, when the language modality is used in conjunction with video and audio modalities, using the proposed multimodal approach it shows the best performance. This suggests that language features are effective for predicting depression severity in the Mood Seasons dataset when used in conjunction with both audio and video data.

The above results suggest that multimodal models, especially those that combine visual, audio, and language modalities, are more effective for predicting depression

than unimodal models, and that the models based on transformer architecture may be particularly well-suited for the task of depression severity estimation.

Modality	Method	MAE↓	RMSE↓	PCC↑
(V)	Resnet-50	3.68	4.60	0.29
(A)	VGGish	3.89	4.74	0.32
(A-V)	Late fusion (VGGish+Resnet-50)	3.42	4.48	0.39
(A-V)	Audio-Visual transformer	3.06	3.97	0.48
(V-L)	Visual-Language transformer	3.17	4.05	0.43
(A-L)	Audio-Language transformer	3.24	4.25	0.38
(A-V-L)	Multimodal transformer	**2.81**	**3.78**	**0.54**

Table 5.6: A comparison of uni-modal and multi-modal approaches on the MoodSeasons validation set. Bold numbers show the best performance. A, V, L represented audio, video, and language modalities.

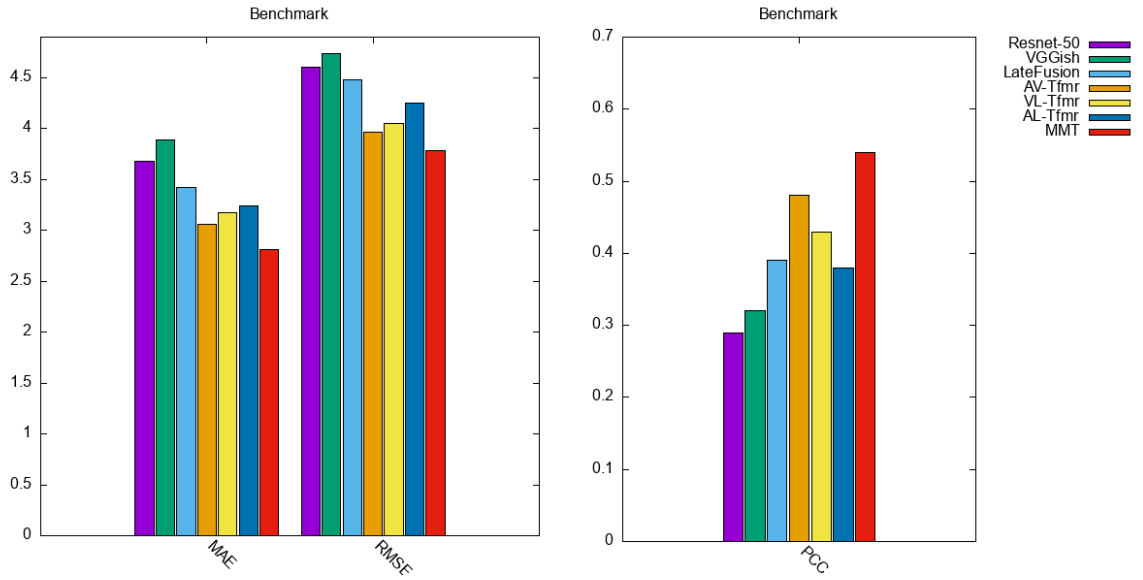


Figure 5.7: A comparison of uni-modal and multi-modal approaches on the MoodSeasons validation set. A, V, L represented audio, video, and language modalities.

5.3.3 Ablation Studies

The ablation studies in this section evaluate the various components used in the proposed framework. All experiments were conducted on the Mood Seasons dataset's validation partition. The following experiments are conducted:

- Sentence-based vs. Video-based: These experiments compare the performance of multimodal transformers that operate on different temporal granularities, namely sentence-based and video-based. The video-level multimodal transformer aggregates the multimodal features using its light-weight fusion architecture from all the sentence-level clips to estimate a depression severity score. The results are shown in Table 5.7.
- Self-attention vs. Cross Attention: The transformer architecture of the proposed framework uses cross-attention heads for multimodal fusion. This experiment replaces these blocks with self-attention heads and reports the results at the video level, where the predictions across videos are averaged for the final prediction. The results are shown in Table 5.9.
- Differential loss: The proposed framework introduces the differential loss desscribed in section 5.2.6. This experiment quantifies the influence of this loss function on predicting depression by comparing the performance of the model with and without this loss component. The results are shown in Table 5.8. These experiments aim to understand the contributions of different components in the proposed framework for depression estimation.

Table 5.7 shows a comparison of sentence level, Sentence MMT and video level, Video-MMT, multimodal transformer models for estimating depression severity. According to the results in the table, it can be seen that the Video-MMT method outperforms the Sentence-MMT method in terms of mean absolute error (MAE) and root mean squared error (RMSE). Specifically, the Video-MMT method has an MAE of 2.62, which improves the MAE of the Sentence-MMT method by 6.78%. Similarly, the Video-MMT method has an RMSE of 3.43, which is a significant relative improvement of 9.5% compared to the RMSE of the Sentence-MMT method. These substantial improvements in performance show that using video-level understanding and long-range modeling can lead to more accurate comprehension of depression by automated systems.

Method	MAE↓	RMSE↓	PCC↑
Sentence - MMT	2.81	3.78	0.54
Video - MMT	2.62	3.43	0.56

Table 5.7: A comparison of different temporal granularities in methods for depression severity estimation. An MMT model trained only using sentence level clips and an MMT model with a video level aggregator based on self attention is compared.

Table 5.8 shows the ablation study on the influence of the differential loss. It is clear that the addition of the differential loss function to the training objectives improved the performance of the proposed approach in terms of all three evaluation metrics. The MAE, RMSE were lower when using the differential loss function compared to the baseline model using only MSE and MAE losses.The MAE decreased by approximately 5.4%, the RMSE decreased by approximately 1.8%. It can be derived from these improvements that differential loss function was effective at improving the performance of depression severity estimation.

MMT	MAE↓	RMSE↓	PCC↑
+ MAE + MSE	2.98	3.85	0.49
+ MAE + MSE + DL	**2.81**	**3.78**	**0.54**

Table 5.8: A loss ablation study showing the influence of differential loss. A baseline MMT model that is only trained with MAE and MSE loss is compared to the MMT model with the additional differential loss

The table 5.9 compares the performance of the proposed approach using types of attention blocks, self attention and co-attention and presents the results in terms of MAE, RMSE, and PCC. Overall, the co-attention mechanism performs better than the self-attention blocks. It has a lower MAE and RMSE. In particular, the co-attention block has an MAE of 2.81 and an RMSE of 3.78, while the self attention variant has an MAE of 3.15 and an RMSE of 4.11. The relative improvement in MAE for co-attention is 10.7% and that of RMSE is 8.2%. These significant improvements reinforce the idea

that co-attention mechanisms are superior in generating multimodal representations that are effective in estimating depression severity.

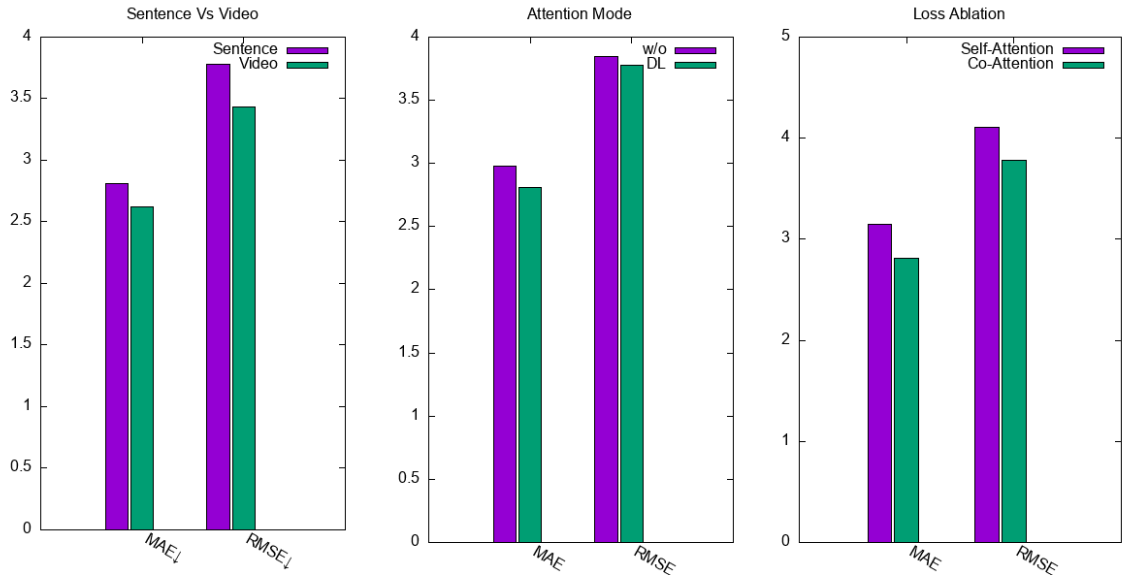


Figure 5.8: Ablation studies for (i) Sentence level and Video level (ii) differential loss and (iii) attention modes.

Method	MAE↓	RMSE↓	PCC↑
Self-Attention MMT	3.15	4.11	0.47
Co-Attention MMT	**2.81**	**3.78**	**0.54**

Table 5.9: A comparison different attention blocks, specifically, the type of attention blocks used in the MMT architecture, namely self-attention and Co-attention

5.3.4 Experiments on the AVEC 2014 Dataset

To further assess the effectiveness of the proposed multimodal approach, experiments were conducted to evaluate its performance on a public dataset and compare it to state-of-the-art depression recognition methods. Therefore, in addition to benchmarking the Mood Seasons dataset with the current state-of-the-art video analysis methods, the proposed approach was compared to other state-of-the-art methods on a publicly available dataset.

Table 5.10 reports the state-of-the-art performance comparison, where all the sota methods are evaluated on the testing partition of the AVEC 2014 dataset. The MMT approach results for both testing and validation sets are provided. The results of these experiments show that the proposed multimodal approach has very competitive performance and is on par with the state-of-the-art in terms of MAE and MSE. These results provide evidence of the efficacy of the proposed approach in tackling the challenge of depression recognition and highlight its potential for use both in-the-wild videos and in-the-lab settings.

	Method	MAE	RMSE
Unimodal	Baseline (Valstar et al., 2014)	8.86	10.86
	Zhu et al. (2017c)	7.47	9.55
	Al Jazaery and Guo (2018)	7.22	9.20
	Zhou et al. (2018)	**6.21**	8.39
	Zhou et al. (2020)	6.59	8.30
	Uddin et al. (2020)	6.86	8.78
	He et al. (2021)	6.59	8.39
	Song et al. (2020)	6.78	8.30
	de Melo et al. (2020)	6.59	8.31
	de Melo et al. (2021)	**6.06**	**7.65**
Multimodal	Sidorov and Minker (2014)	11.20	13.87
	Jan et al. (2017)	6.68	[8.01]
	Niu et al. (2020)	6.43	8.60
	MMT	6.54	8.20
	MMT (Validation)	5.84	7.64

Table 5.10: Comparison of MMT approach with state-of-the-art approaches on AVEC 2014 testing set. Unimodal and Multi-modal approaches are compared. Bottom row reports results of MMT on the AVEC 2014 validation set.

5.3.5 Limitations

The study's primary limitation is the limited size of the testing and validation sets, with 69 and 79 videos each. While larger than the AVEC datasets, that consists of

150 videos in total with 50 videos each for training, validation and testing, only 13% of this set comprises participants classified as depressed based on the PhQ-8 scale. A cross-validation approach could evaluate the proposed MMT approach's generalization and robustness.

Another limitation in the sentence based approach for short term modelling is that the pauses and gaps between the sentences would be ignored. This can be tackled by an enhanced speech recognition model to include pause duration between consecutive sentences.

5.4 Conclusion

This chapter has addressed the aims of the thesis by making two significant contributions to the field of automated depression analysis:

- An extensive benchmark of state-of-the-art video analysis techniques on the newly collected Mood Seasons and publicly available AVEC 2014 datasets. The benchmark showed that models with temporal context understanding performed the best, which is an important finding for future research in automated depression analysis.
- A two-staged multimodal transformer based approach to automated depression severity prediction. The approach achieved promising results on both the Mood Seasons and AVEC 2014 datasets, demonstrating the feasibility of using multimodal data for depression analysis.

The Mood Seasons dataset is a valuable new resource for the research community, and the benchmark provides insights into the performance of different video analysis techniques on depression data. The proposed two-staged multimodal transformer based approach is a novel and promising approach to automated depression severity prediction.

The work presented in this chapter is a promising step towards developing advanced methods for automated depression analysis. It makes significant contributions to the field of automated depression analysis and provides a foundation for future research in this area.

Chapter 6

Face Image Generation And Applications In Anonymisation

6.1 Introduction

Synthetic data generation plays a crucial role in computer vision, enabling training without the need for extensive real datasets. For tasks like facial recognition, augmenting training data with artificially generated variations in pose, expression, and lighting enhances generalization and mitigates overfitting on limited real examples. Moreover, synthetic faces offer attribute labels not available in real datasets. This chapter focuses on developing an innovative method for synthesizing facial images in arbitrary poses, with the aim of facilitating synthetic data generation for face analysis applications.

Facial data anonymization is equally vital, especially in sensitive domains like healthcare. Safeguarding patient privacy when sharing datasets for research necessitates the removal of identifying information. Face anonymization obscures identity while preserving other facial attributes and background, allowing the application of computer vision techniques to mental health data with confidential patient identities protected. The work presented demonstrates the anonymization of a depression video dataset, serving as an initial proof-of-concept of how the face manipulation approach can ensure privacy protection.

While the face generation and anonymization techniques explored in this chap-

ter hold significant potential, they constitute foundational work that requires further development. The results establish initial feasibility and provide a launching point for future research.

This research builds on previous chapters' efforts in automated depression analysis from a newly collected multimodal, longitudinal and real-world dataset. Synthetic data generation and anonymization applied to the in-the-wild Mood-Seasons dataset and evaluated for depression assessment using the techniques developed in Chapter 5 is a significant contribution for developing private and generalisable models for mental health applications.

The key contributions are as follows: 1) Introducing a novel method for manipulating facial pose and expression through attribute transfer from an exemplar image, 2) Incorporating an appearance transfer module to integrate features across domains, 3) Demonstrating the potential of this approach by anonymizing a depression dataset as an initial case study.

In summary, this chapter introduces innovative techniques for synthetic face generation and anonymization, addressing critical data and privacy challenges in applying computer vision to sensitive domains like healthcare. While further research is imperative, the results mark significant initial strides and proof-of-concept demonstrations in these directions.

The generation of face images with varying poses and expressions is a rapidly evolving area in computer vision, particularly in face analysis. Its applications span data-efficient learning through augmentation, adaptation, and few-shot learning, as well as virtual avatars, adversarial attacks, privacy preservation, and research in interpretability and explainability. Face manipulation encompasses techniques like face swapping and expression transfer, involving alterations to facial attributes such as head pose, landmarks, gaze, identity, gender, race, or emotions. These techniques find applications in both consumer industries and research.

In research, face image generation is increasingly employed for data-efficient learning and privacy preservation. Augmenting training datasets with a diverse set of images can enhance the performance of downstream tasks such as facial recognition or re-

identification, mitigating overfitting. For instance, Generative Adversarial Networks (GANs) have been utilized to generate synthetic face images with different poses, lighting conditions, and backgrounds, leading to substantial improvements in recognition accuracy with limited real data. Microsoft's synthetic dataset outperformed state-of-the-art face analysis methods on various benchmarks, offering labels at no additional cost.

Face anonymization is a pivotal aspect of safeguarding individual privacy and falls within the broader domain of face manipulation techniques. This is particularly pertinent in datasets where facial images carry risks of identity theft and unauthorized access, leading to potential misuse. By implementing suitable methods, it is possible to eliminate face identity while maintaining modeling accuracy.

This chapter introduces a fresh approach to manipulate facial images, altering their pose and expressions. The first section outlines the method, its motivation, and the steps for synthesizing novel views of a face image based on key points, along with experiments validating its efficacy in face pose transfer. The second section discusses how this face manipulation method can be employed to anonymize face images in the collected Mood-Seasons datasets. It covers the necessary adjustments for anonymization and evaluates its effectiveness. The final section addresses further considerations regarding the use of anonymized datasets in depression recognition and offers recommendations for integrating anonymized data into depression analysis.

6.2 Face Manipulation Via Hallucination

This research work tackles the following image generation problem: given a face image I_X of pose X (including rigid pose and expression) and a target pose Y, can we generate a new face image I_Y that preserves the general appearance and layout of I but depict the face in pose Y? Previous methods have tackled this problem using paired data during training. This means that the training set contains several instances of the same identity in different poses as well as capturing conditions (e.g. illumination). However, the paired setting is restrictive mainly because constructing such a diverse dataset with thousands of different subjects is by no means straightforward. In this work, the goal is the more general setting of unpaired generation

where a large number of training images is available but no effort has been made in terms of collecting and labelling images of the same identity.

The challenge in this assumed unpaired training is how to effectively combine information from the input image and the target pose. Specifically,one needs to ensure 2 key requirements: (a) only the general appearance, and layout from the source image are transferred; (b) the generated image is in the correct pose, and is of high visual quality.

To address (a) this research proposes for the first time to model the general appearance, layout, and background of the input image using a low-resolution version of it which is progressively passed through a hallucination network to generate features at higher resolutions. The experiments show that such a formulation is significantly simpler than previous approaches for appearance modelling based on auto-encoders which introduce an unnecessary complexity into the process, that of learning a latent representation which typically is not well disentangled in terms of pose and appearance.

Moreover, the framework uses a conditional pose-guided generator GAN framework to generate images in the target pose. To address (b), and inspired by (Park et al., 2019b), the work proposes a fully learnable and spatially-aware appearance transfer module which can cope with misalignment between the input source image and the target pose and can effectively combine the features from the hallucination network with the features produced by the generator. The end result is that the generator produces face images in the target pose by integrating, in the process, features from the hallucination network, capturing the appearance of the source image. In summary, the contributions of this research are:

- proposes an unpaired image-to-image translation method in which a face hallucination network guides a pose-synthesis network to manipulate the input low-resolution image according to the target pose.

- introduces the Appearance Transfer Module, a fully trainable spatially-aware module to deal with the misalignment between the hallucination features and those generated by the pose-synthesis network.

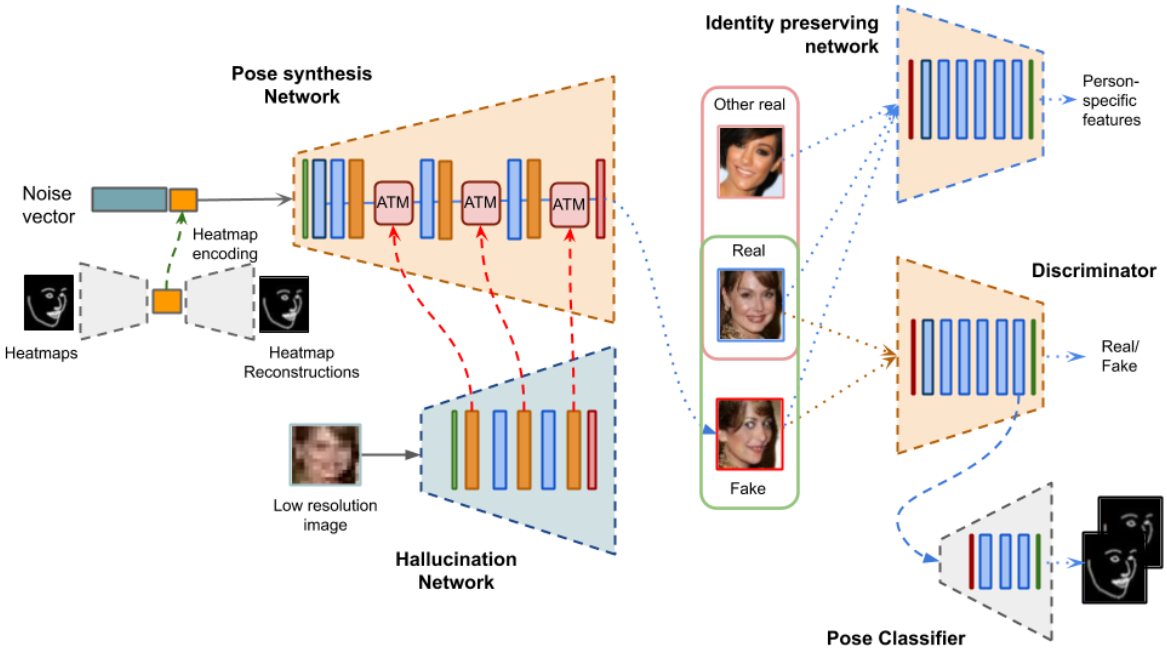


Figure 6.1: Image-to-image translation approach in an unpaired setting, where a low-resolution facial image is forwarded to a hallucination network (bottom), to produce appearance-specific features, that are used by a pose-synthesis network (top), through a newly introduced Appearance Transfer Module (ATM). The method is learned in a GAN setting, using a discriminator (bottom right) with an auxiliary pose classifier, and an identity preserving network, that is trained in a collaborative way with a contrastive loss (top right).

- proposes an auxiliary classifier network, which faciliates unsupervised conditional face image generation and enforces pose transfer without needing labels
- introduces an identity preserving method that is trained in an unsupervised way, by using an auxiliary feature extractor and a contrastive loss between the real and generated images.
- The experiments show that the method outperforms prior work on unpaired face generation by considerable margin. The sections also provide analysis showcasing the effect of the different components of the system.

6.3 Related Work

Face manipulation: One can distinguish works in face synthesis or manipulation according to whether they aim at random face generation (Karras et al., 2017, 2019; Kossaifi et al., 2018) or at modifying or re-enacting a given face image (Choi et al., 2018; Pumarola et al., 2018a; Zakharov et al., 2019). In both cases, it is often assumed an attribute or "style" driven approach, where the target is to generate faces that possess a specific attribute or follow a certain style. Works on face manipulation typically include frontalization (TP-GAN (Huang et al., 2017)), pose synthesis (CAPG-GAN (Hu et al., 2018)), or expression synthesis (CMN-Net (Wang et al., 2018), Animation (Pumarola et al., 2018a)). Other works aim at generating attributes to the target images, such as StarGAN (Choi et al., 2018) which synthesizes faces according to target facial attributes, such as "Blonde Hair", or facial expressions. Some methods use 3DMM to control pose, expression and lighting settings for synthesizing images (Gecer et al., 2018; Shen et al., 2018). Finally, some works aim at geometry-driven face manipulation, often referred to as face reenactment. In this line, some works propose to modify an input image according to a set of target expressions (Thies et al., 2016), and landmarks (Sanchez and Valstar, 2018).

It is worth noting that most methods for face manipulation fall into the paired training setting. As such, from the methods mentioned above, the method is somewhat related to StarGAN (Choi et al., 2018) and GANimation (Pumarola et al., 2018a) in a sense that a facial attribute (hair color, expression, etc. in StarGAN, and AUs in GANimation) is used to manipulate an input face image under an unpaired training

setting.

Human pose manipulation: Although the work aims at faces, there is also a large body of work on human synthesis and manipulation. Some of most recent works build on a paired setting, and try to enforce a correct feature transfer, similarly to the proposed Appearance Transfer Module (ATM). For instance, (Zhu et al., 2019) proposes a Progressive Pose Attention, with modules that mix features coming from the input image with those coming from a pose encoder. However, there is no explicit mechanism to deal with spatial misalignment between the input image and the target pose, and thus it is not clear whether the transfer blocks proposed in (Zhu et al., 2019) can be used effective for the unpaired case too. Alternatively, the works of (Dong et al., 2018; Li et al., 2019) do explicitly deal with misalignment through a geometric warping modelling, which may not always be valid (depends on the motion model assumed and the dataset) or might be hard to implement. the proposed ATM, inspired by (Park et al., 2019b), removes the need of geometric warping and is completely learnable without assuming any motion model.

Related human pose manipulation works assuming an unpaired setting are (Lorenz et al., 2019; Ma et al., 2018; Pumarola et al., 2018b; Song et al., 2019). The work of (Ma et al., 2018) is one of the first ones to attempt unpaired training introducing a multi-branch reconstruction network for disentangling and manipulating foreground, background and pose information which are then combined to reconstruct the input image itself. The method of (Pumarola et al., 2018b) extends Cycle-GAN (Zhu et al., 2017b) for multi-view synthesis using a conditional pose loss and an identity loss. This method has been shown less capable of preserving the appearance of the reference image. More recently, (Song et al., 2019) uses a module for firstly generating a semantic map under the target pose, and guided by the that map and the reference image, an appearance generation module synthesizes the final output image. the method by-passes the step of predicting a dense semantic map, which is on its own a difficult problem. Finally, the work of (Lorenz et al., 2019) aims at disentangling pose and appearance in an unsupervised way, by using an image-to-image translation approach, thus being the latter not the ultimate goal.

6.4 Proposed Approach

The overall description of the proposed approach is depicted in Fig. 6.1, and consists of five blocks: a pose encoder, a hallucination network, a pose-synthesis network, a discriminator with two heads – one to distinguish real and fake images and one containing a pose classifier, an appearance transfer module and an identity preserving network.

The goal is to transfer the appearance of an input face to a target pose, without the use of paired training data. To do so, the method relies on a pre-trained hallucination network, that provides features capturing the appearance of the input face at multiple spatial resolutions. These features are integrated gradually with the ones generated by the pose-synthesis network – the main network in the pipeline – the goal of which is to generate the face in the target pose. The integration is done through the proposed appearance transfer module. The pose encoder simply provides an embedding of the target pose, which is fed as input into the pose-synthesis network. Moreover, the identity preserving network ensures that the generated face has the same identity as the input face. The discriminator networks are used to train the pose-synthesis under an unpaired training setting.

Notation: Images are represented as I. The input and target pose, defined as an edge map, connecting the facial landmarks corresponding to different facial parts (i.e., face boundary, eyebrows, eyes, nose, mouth), are represented by X and Y, respectively. I_X represents the image corresponding to pose X. The subscript LR refers to low-resolution images (i.e., 16×16 images), and SR to refer to images or features coming from the hallucination network (often referred to as super-resolution). The subscript PS refers to features computed within the Pose-Synthesis network.

6.4.1 Framework components

Pose encoder: Rather than feeding the network with the target edge-maps, the framework first encodes them into a low-dimensional representation of $64 \times 4 \times 4$ learned in an auto-encoding framework that aims at reconstructing the maps from the low-dimensional representation. This pose encoder is trained beforehand and kept frozen during the synthesis training. The *encoding* of the input pose is denoted

as $E(Y)$.

Hallucination network: The hallucination network is based on the face super-resolution network of (Bulat and Tzimiropoulos, 2018), which is driven by a facial landmark localization network that enforces both high and low resolution images to have the same facial structure. The network takes as input a 16×16 low-resolution image and generates its high resolution counterpart, which in this paper is set to 128×128[1]. In order to seamlessly transfer the features from the hallucination network to the pose-synthesis pipeline, the proposed approach slightly modified the architecture of the former to make it similar to that of the latter. In particular, the hallucination network consists of a 3×3 convolution layer, 5 residual blocks and a final 3×3 convolution layer as shown in Fig. 6.1. It consists of up-sampling residual layers that act across different spatial resolutions. Rather than using transposed convolutions, the residual blocks are composed of a pixel shuffle module followed by a convolution. This design is chosen to match the architecture of the hallucination network with the pose-synthesis network, which borrows its design from (Arjovsky and Bottou, 2017b). The hallucination network is pre-trained following (Bulat and Tzimiropoulos, 2018), and kept frozen for the training of the pose-synthesis network.

Pose-synthesis network: The pose-synthesis network takes as input a 128-d noise vector z, and the low-dimensional representation of the target pose, $E(Y)$, and produces a face image the geometry of which is defined by Y. The noise z goes first through a linear layer that brings its resolution to $512 \times 4 \times 4$, which is then concatenated with the encoded pose. The architecture of the pose-synthesis network is depicted in Fig. 6.1, and consists of a fully connected linear layer, 7 residual blocks, and a final 3×3 convolutional layer. The pose-synthesis network can be trained with or without (i.e., on its own) integrating features from the hallucination network. The framework used the latter approach to pre-train a network to initialize the former. On its own, the goal of the pose-synthesis network is to generate a realistic face image that follows the geometry of Y. To accomplish this task without paired data, the network is firstly trained using a GAN approach, where a conditional discriminator is used to distinguish real and generated images, as well as to predict the pose in the input image.

[1]For ablation studies, the work also uses a 64×64 resolution

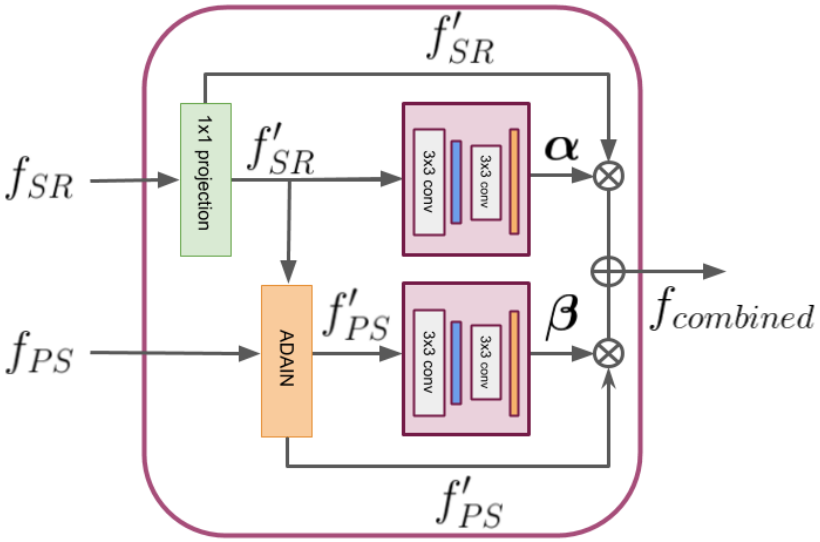


Figure 6.2: Appearance transfer module. (a) The spatial α, β based appearance transfer module

6.4.2 Appearance Transfer

The pose-synthesis network allows generating realistic face images that follow the geometry of the target pose Y, whereas the hallucination network allows capturing appearance features at different scales of the source input image. This section shows how one can use the latter to drive the former to translate the input image I_X into the target pose Y. It was observed that combining both networks enables the appearance transfer from the input image to the target pose, in an approach that can be trained in a fully unpaired setting.

Appearance Transfer Module: To allow the generated image to follow the style of the input one, the proposed approach borrows ideas from (Karras et al., 2019) and (Park et al., 2019b) and propose to inject features from the hallucination network into the different layers of the pose-synthesis network, in a style-transfer fashion. However, note that the features from the input image are not aligned with those of the pose-synthesis network, and therefore a spatial module is necessary to ensure a correct transfer. To this end, the method proposes a novel Appearance Transfer Module (ATM), which plays a key role in the method, allowing the pose-synthesis network to incorporate appearance features at different resolutions from the hallucination network.

For a given spatial resolution, the ATM combines the features from the halluci-

nation network, f_{SR}, with the features from the pose-synthesis network, f_{PS}. The ATM includes (i) a 1x1 convolution in order to align (channel-wise) f_{SR} with f_{PS}; The output of this operation is a new set of aligned features f'_{SR}. (ii) an Adaptive Instance normalization (AdaIn) layer (Huang and Belongie, 2017) that aligns the feature statistics (μ and σ) of f_{PS} with those of f'_{SR}, to produce a new set of features f'_{PS}; and (iii), inspired by (Park et al., 2019b), a "spatially aware" combination layer, that defines a spatial learnable weighted combination of the feature maps from f'_{PS} and f'_{SR}. The AdaIn layer is defined as:

$$f'_{PS} = \sigma(f'_{SR})\left(\frac{f_{PS} - \mu(f_{PS})}{\sigma(f_{PS})}\right) + \mu(f'_{SR}), \tag{6.1}$$

whereas the combined feature is given by

$$f_{combined} = \alpha \odot f'_{SR} + \beta \odot f'_{PS}. \tag{6.2}$$

In Eqn. 6.2, $\odot$ represents element-wise multiplication between features and weighting masks $\alpha = \phi(f'_{SR})$ and $\beta = \psi(f'_{PS})$ which have **the same spatial resolution as the features**. The masks α and β are produced by convolutional modules, ϕ and ψ, that operate on hallucination and pose synthesis features, respectively (pink modules in Fig. 6.2). Each module consists of a 3×3 convolution, ReLU and average pooling layers followed by another 3×3 convolution and a bi-linear upsampling. α and β control the contribution of the two feature maps during the combination. This will help the network not to learn to copy the features exclusively from the hallucination network.

Conditional Discriminator: The work uses a conditional discriminator based on the auxiliary classifier GAN to generate valid images conditioned on a given pose. The framework adds a pose edge map regressor head on top of the discriminator as the auxiliary classifier to condition on pose information. In conditional GAN training, the generator takes as input a noise vector z and a class condition label, c which is represented by the target pose in the method.The discriminator provides both the probability distribution over the real data (i.e. real/fake) and a probability distribution over the class label. However, unlike traditional methods, this novel conditional discriminator adds an auxiliary heatmap regression network which enforces landmark adherence in the generated images.

The objective function has two components,L_c representing the log-likelihood of the class labels and L_{adv}, the adversarial loss or the log-likelihood of the real data.

$$L_c = E[logP(C = c|X_{real})] + E[logP(C = c|X_{fake})] \tag{6.3}$$

L_c is estimated by a novel pose heatmap regression network, thereby forgoing the need to have supervision, i.e., real images in the target pose. The discriminator consists of residual blocks, with downsampling operations implemented using mean pooling within the residual blocks. The discriminator mirrors the layers of the generator. The discriminator also uses spectral normalization, as provided in (Miyato et al., 2018).

Identity preserving network: In addition to the aforementioned ATM module, an identity preserving network is introduced, which is targeted at producing features that are close for pairs of input/generated images so that identity is preserved, and far for other combinations of input images and/or generated images. The network has the form of a ResNet-18(He et al., 2016b), in which the last Average Pooling and FC layers are removed, and is trained along with the discriminator. The network produces a 8192-d feature vector for an input resolution of 128×128. It uses a contrastive loss (Hadsell et al., 2006) to train the network (see below). Note that the proposed approach do not make use of paired training data: when updating the feature extractor, use as positive pairs the input and generated images in a batch; to generate the negative samples, the approach pairs the input (or generated) images in a batch with a shuffled version of the batch.[2] When updating the generator, the negative pairs come from the generated images of a shuffled version of the input batch. This way, the generator will try to produce images that have similar features for the positive pairs, and dissimilar for the negative ones.

6.4.3 Training Methodology

The training of the pipeline is divided into two main stages. The first one comprises pre-training the pose encoder, and the face hallucination and pose-synthesis net-

[2]Due to the randomness in the batch sampling, there might be cases where the negative pairs are composed of two images corresponding to the same person, although the occurrence of this scenario is negligible.

works. The face hallucination network is trained as in (Bulat and Tzimiropoulos, 2018), whereas the pose-synthesis network is trained using a GAN approach with the conditional discriminator defined earlier.

Once the pose encoder, face hallucination, and pose-synthesis networks have been trained, they are integrated into the whole pipeline. While the encoder and face hallucination networks remain frozen, the pose-synthesis network is re-trained, with initial weights being the pre-trained ones. The corresponding discriminator is disregarded and trained from scratch along with the whole pipeline. Overall, the last stage comprises training the pose-synthesis network with the ATM modules, the discriminator (D), and the identity preserving network (IP). Following the GAN notation, the following will refer to the whole block consisting of the pose encoder, the pose-synthesis, the ATM modules, and the face hallucination network, as G. To simplify notation, the output of G will be denoted as $\hat{I}_Y = G(z, I_X, Y)$, illustrating the dependency on the noise z, the input image I_X with pose X, and the target pose Y. Recall that in G only the parameters of the ATM modules and the pose-synthesis network are learnable in this stage. The loss that G, D, and IP aim to optimize is decomposed into several terms, detailed below:
Adversarial loss: The approach adopts the hinge adversarial loss of (Lim and Ye, 2017b). The loss for the discriminator is defined as:

$$\begin{aligned} \mathcal{L}_{adv}^{D} &= E_{z,I_X,Y}[\min(0, -D(G(z, I_X, Y)) - 1)] \\ &+ E_{I_X}[\min(0, -1 + D(I_X)] \end{aligned} \tag{6.4}$$

whereas the loss for the generator is given by:

$$\mathcal{L}_{adv}^{G} = -E_{z,I_X,Y}[D(G(z, I_X, Y))]. \tag{6.5}$$

Pose loss: In order to ensure that the generated image follows the geometry defined by the target pose Y, the method proposes an an auxiliary classifier on top of the discriminator. The pose classifier is trained using heatmap regression as in (Bulat and Tzimiropoulos, 2018), and is used to update the generator weights according to the localization error estimated on the generated image. The pose loss is denoted as $\mathcal{L}_p$ The pose regressor network optimizes the L2 loss between the generated heatmaps

and corresponding ground truth heatmaps. The classification loss, L_c is given by,

$$L_p = E_{z,I_X,Y,X}\|D(G(z, I_X, Y)) - Y\|^2 + \|D(I_X) - X\|^2 \tag{6.6}$$

Reconstruction loss: In order for the network to preserve the input appearance when the source and target poses are the same, the reconstruction loss is used (Choi et al., 2018):

$$\mathcal{L}_r = E_{z,I_X,Y}\|I_X - G(z, \hat{I}_Y, X)\|^2 \tag{6.7}$$

Contrastive loss: The identity preserving network should produce features that help distinguish whether two images are from the same person at different poses, or are from different persons. The Contrastive Loss (Hadsell et al., 2006) typically used for face recognition is used:

$$\mathcal{L}_{con} = (1 - y_{ij})(\Delta f_a^{ij})^2 + y_{ij} \max(0, m - \Delta f_a^{ij})^2, \tag{6.8}$$

where Δf_a^{ij} corresponds to the L2 norm between the features corresponding to images i and j, and y_{ij} is 1 if i, j is a positive pair, and 0 otherwise. The margin m is set to 1 in the experiments. Given that the proposed approach builds on an unpaired scenario, we are not given pairs of images corresponding to the same person at different poses. Therefore, a collaborative training approach is implemented: the output of the generator is used to update the identity preserving network, and vice versa. The method defines the positive pairs as the tuples $\{I_X, \hat{I}_Y\}$. When updating the identity preserving network, the negative pairs are defined as $\{I_X, I'_{X'}\}$, where the images $I'_{X'}$ come from a shuffled version of the input batch. When updating the generator, the negative pairs are formed using the input images and the shuffled version of the input batch, as $\{\hat{I}_Y, I'_{X'}\}$.

Full objective: The training objectives for D and the G are:

$$L_D = \lambda_{adv} L_{adv}^D + \lambda_p L_p, \tag{6.9}$$

$$L_G = \lambda_{adv} L_{adv}^G + \lambda_p L_p + \lambda_r L_r + \lambda_{con} L_{con}, \tag{6.10}$$

where the λ values are the weights for each loss term.

6.4.4 Pre-training Procedure

Three main components of the method undergo pretraining, the hallucination network, the pose encoder and the conditional GAN. A description of the training process of these components are given below.

Hallucination Network The hallucination network of section **3.1**, G_{SR}, takes a low resolution input image, I_{LR} and generates a high resolution image, $\hat{I}_{HR}$.

I_{HR} represents the ground truth, high resolution image.The hallucination network was trained using the same loss functions as (Bulat and Tzimiropoulos, 2018) given by,

$$L_H = \alpha L_{pixel} + \beta L_{feat} + \gamma L_{heatmap} \tag{6.11}$$

where,

$$L_{pixel} = \|I_{HR} - \hat{I}_{HR}\|^2 \tag{6.12}$$

$$L_{feat} = \|\phi(I_{HR}) - \phi(\hat{I}_{HR})\|^2 \tag{6.13}$$

L_{feat} represents the perceptual loss between the super-resolved image and its ground truth. $\phi(I_{HR})$ and $\phi(\hat{I_{HR}})$ represents the low and mid level features extracted from a pre-trained Resnet that was trained on ImageNet dataset.

$$L_{heatmap} = \|\tilde{H} - \hat{H}\|^2 \tag{6.14}$$

$\tilde{H}$ represents the heatmaps given by a face alignment network denoting the landmark locations of the ground truth image I_{HR} and $\hat{H}$ denotes the heatmaps provided by

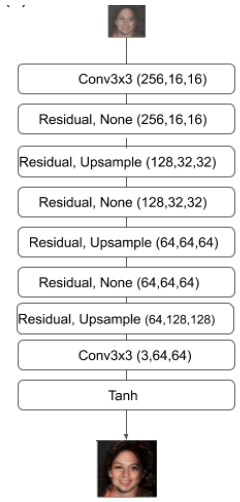
the face alignment network on the super-resolved image, $\hat{I_{HR}}$. During training, the method uses $\alpha = 0.5$, $\beta = 0.5$ and $\gamma = 0.5$.

Pose encoding network The auto-encoder for the pose encoder network in section **3.1** of the paper, is trained using a reconstruction loss given by,

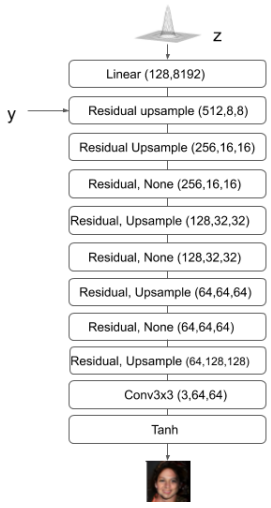
$$\mathcal{L}_{rec} = \|Y - \hat{Y}\|^2 \tag{6.15}$$

where $\tilde{Y}$ represents the input edge map and $\hat{Y}$ represents the corresponding reconstructed edge map from the decoder.

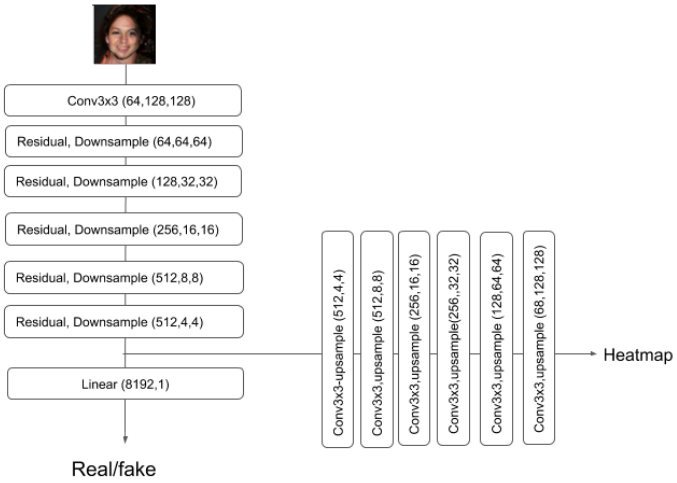
Conditional GAN The method first trains a conditional GAN that can synthesize images that are conditioned on a given target pose. The architecture of the generator network is the same as that of the pose synthesis network, but without the appearance transfer modules added. The discriminator network architecture is also the same as that is described in section **3.1.** of the paper. The framework uses the adversarial hinge loss, L_{adv} and pose regression loss L_p as given in equations **(3)**, **(4)** and **(5)** in section **3.3.** of the paper. The conditional GAN is trained using $\lambda_{adv} = 1.0$ and $\lambda_p = 1.0$.



(a) (a)



(b) (b)



(c) (c)

Figure 6.3: Network Architectures of our (a) Hallucination network (b) Pose Synthesis network, y denotes the encoding from the pose encoder network (c) Conditional discriminator with auxiliary pose classifier. The residual blocks follow the architecture of (Gulrajani et al., 2017a)

.

6.5 Experiments

This section presents the experiments that validate the effectiveness of the proposed approach. To this end, the proposed approach is compared to state-of-the-art methods, and also present several ablation studies. These experiments validate the results both qualitatively and quantitatively.

Datasets: CelebA dataset (Liu et al., 2015), which comprises $\sim$ 200k images of 10,177 celebrities portraying a wide array of pose and expressions, is used. A training split of 192,600 images is used out of which 6,400 images were chosen for testing, with no overlapping identities. Facial landmarks were extracted using (Bulat and Tzimiropoulos, 2017). The images are cropped and resized according to the landmarks to 128×128. In addition, the Multi-PIE dataset (Gross et al., 2010) is used, which contains images of 337 subjects, captured in a controlled setting condition under 15 viewpoints and 19 illuminations. From Multi-PIE, 60,000 images are used to train, and 5,500 to test, with no overlapping identities.

Training Details: The approach uses batch size of 64 and Adam optimizer with $\beta_1, \beta_2 = (0, 0.999)$ for training. The learning rate for the generator and discriminator is set to 0.0002. The models are trained for 160K iterations, taking 3 days on a single NVIDIA Titan X Pascal. The values of λ are set as $\lambda_{adv} = 1$, $\lambda_p = 10$, $\lambda_r = 5$, and $\lambda_{con} = 5$.

Pre-training phase: The framework trains a hallucination network for each of the aforementioned databases. The training of this network is done in advance, following the protocol of (Bulat and Tzimiropoulos, 2018). The LR images are generated from the input images by applying a spatial downsampling. The network is trained to recover the input images.

Performance metrics: Inception Score (IS, (Salimans et al., 2016)) and Fréchet Inception Distance (FID, (Heusel et al., 2017)) are used to evaluate the quality of each method: low FID and high IS values indicate better quality. The experiments also measure the capacity of the approach to preserve identity by computing the similarity (L2) between the features extracted using the publicly available face recog-

nition model LightCNN (Wu et al., 2018) from the input and generated images, as well as the percentage of images that yield a distance lower than 0.2.

6.5.1 Super-resolution Performance

The proposed approach first involves training the super-resolution network according to the specifications of (Bulat and Tzimiropoulos, 2018). The Super-resolution network was trained with pixel loss, perceptual loss and heatmap loss with contribution factors of 1.0,1.0 and 0.5 respectively. Hyperparameters include a learning rate of 2.5e-4 which was decreased to 1e-5 during the 100 epochs of training. The FAN was fine-tuned along with this network for another 5 epochs. The best NME, PSNR and SSIM for the faces dataset are given in Table 6.1.

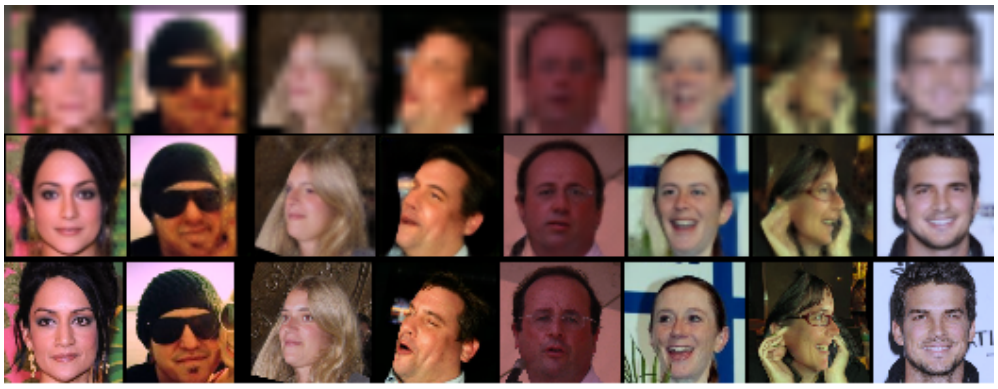


Figure 6.4: Super-resolution experiment results. The top row shows the low resolution 16x16 images, the middle row shows the generated high resolution images and the third row shows the high resolution ground truth images.

Table 6.1: Super-resolution Results

Measure	Pose 60	Pose 90	Pose 30
SSIM	0.79	0.77	0.79
PSNR	23.4	22.5	23.2
NME	60.8	53.8	61.4

6.5.2 Comparison with state-of-the-art

The experiments compare the proposed approach against the state-of-the-art methods Pix2PixHD(Wang et al., 2019) and StarGAN (Choi et al., 2018). The former is not intended to be used for image-to-image translation, and is primarily used to compare the quality of the generated images. StarGAN is the state-of-the-art method in unpaired image-to-image translation. Both methods are accompanied by publicly available implementations, which are used to train the corresponding models using the aforementioned databases. To the best of the knowledge, StarGAN has not been applied to landmark-guided image-to-image translation, and hence the code is modified to make the discriminator accept edge maps as attributes.

The proposed approach is compared against a very strong baseline: the *in-house conditional GAN*, i.e. the proposed GAN-based generator with a conditional discriminator that generates face images conditioned on a given pose and random noise. This network is actually the pose-synthesis network **without the integration of the features from the hallucination network**. The network is trained using adversarial and pose regression losses and serves as the pre-trained network used to initialize the final version of the method.

The results comparing the method on both CelebA and Multi-PIE against the three aforementioned methods are shown in Table 6.2. the method outperforms all other methods both in terms of FID and Inception Score and delivers the highest quality images in the experiments.

In addition, qualitative evaluation are provided in Figure. 6.5, Figure. 6.6, Figure 6.7.StarGAN fails to produce good quality results under a pose change setting. It struggles to generate high frequency details and fails, especially when attributes like glasses are present. The visual comparison shows evident superiority of the method in generating higher quality images that corresponds to a target pose, while keeping the source image features (e.g. hair color, skin color, makeup, glasses, texture, lighting, etc.) the method is also robust towards generating images in a wide range of poses, (e.g. profile to frontal) and can manipulate expressions.

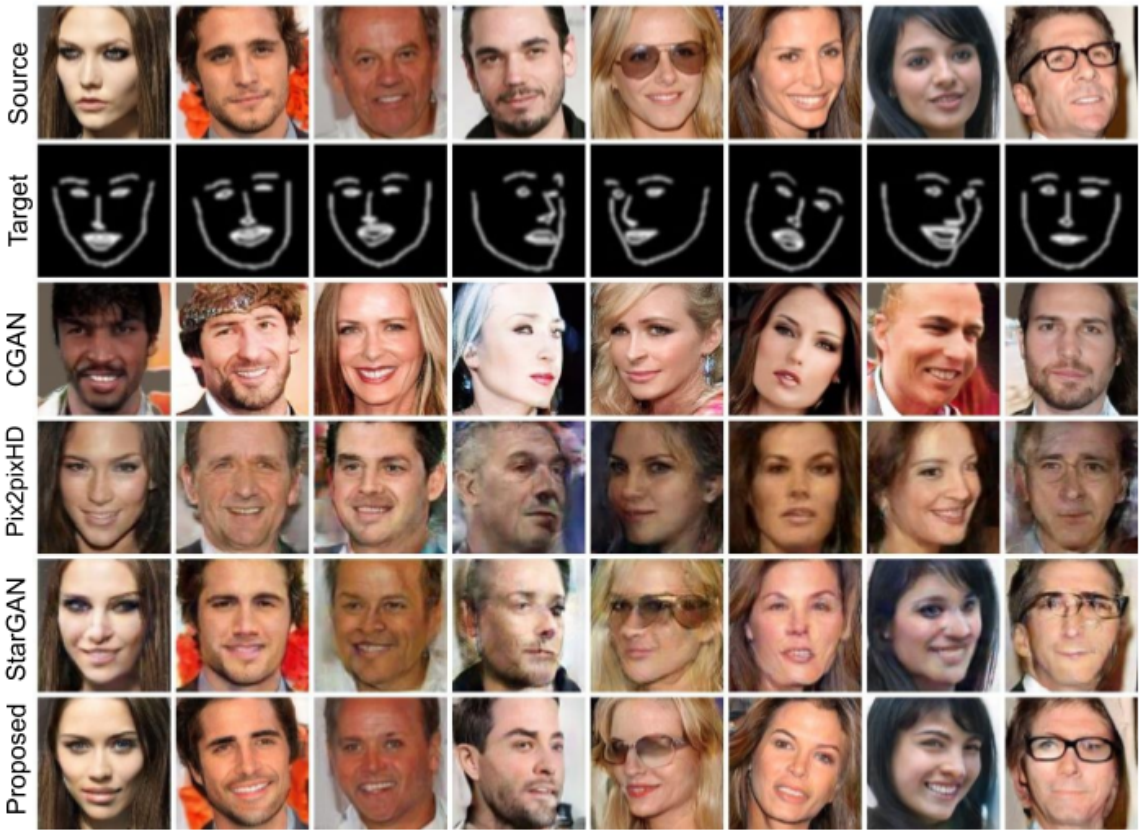


Figure 6.5: Qualitative comparison w.r.t to state-of-the-art (a) Source image, (b) Target pose, (c) Conditional GAN, (d) Pix2pixHD, (e) StarGAN and (f) Proposed method.

	Method	FID↓	IS↑		Method	FID↓	IS↑
MultiPIE	Real data	0.00	2.14	CelebA	Real data	0.01	3.49
	CGAN	22.9	1.79		CGAN	7.40	2.42
	Pix2pixHD	19.30	1.58		Pix2pixHD	41.68	2.62
	StarGAN	25.29	**1.81**		StarGAN	12.78	2.55
	Ours	**15.90**	1.78		Ours	**6.14**	**2.65**

Table 6.2: Comparison w.r.t state-of-the-art methods.Note: The method does not outperform StarGAN in Inception Score, although it offers competitive performance in FID.

Ablation studies

In this Section the experiments investigate the contribution of the different components of the proposed method. To reduce training times, a target 64×64 resolution is used.

Figure 6.6: CELEBA state-of-the-art comparison. Additional results from CELEBA dataset with respect to the baseline CGAN, and state-of-the-art methods, Pix2pixHD, StarGAN are shown above.

Settings

A.1 – Hallucination network vs Appearance autoencoder: A key feature of the approach is the use of the hallucination network for capturing the appearance of the input target image, and guiding the pose synthesis network. An obvious alternative to the approach would be to use an appearance autoencoder (trained to reconstruct facial images), and then try to transfer features from that network in order to guide the the pose synthesis network.

Two variants for the appearance autoencoder were explored. In the first one (A.1.1), the input source image in full resolution is fed first to an encoder. The output of

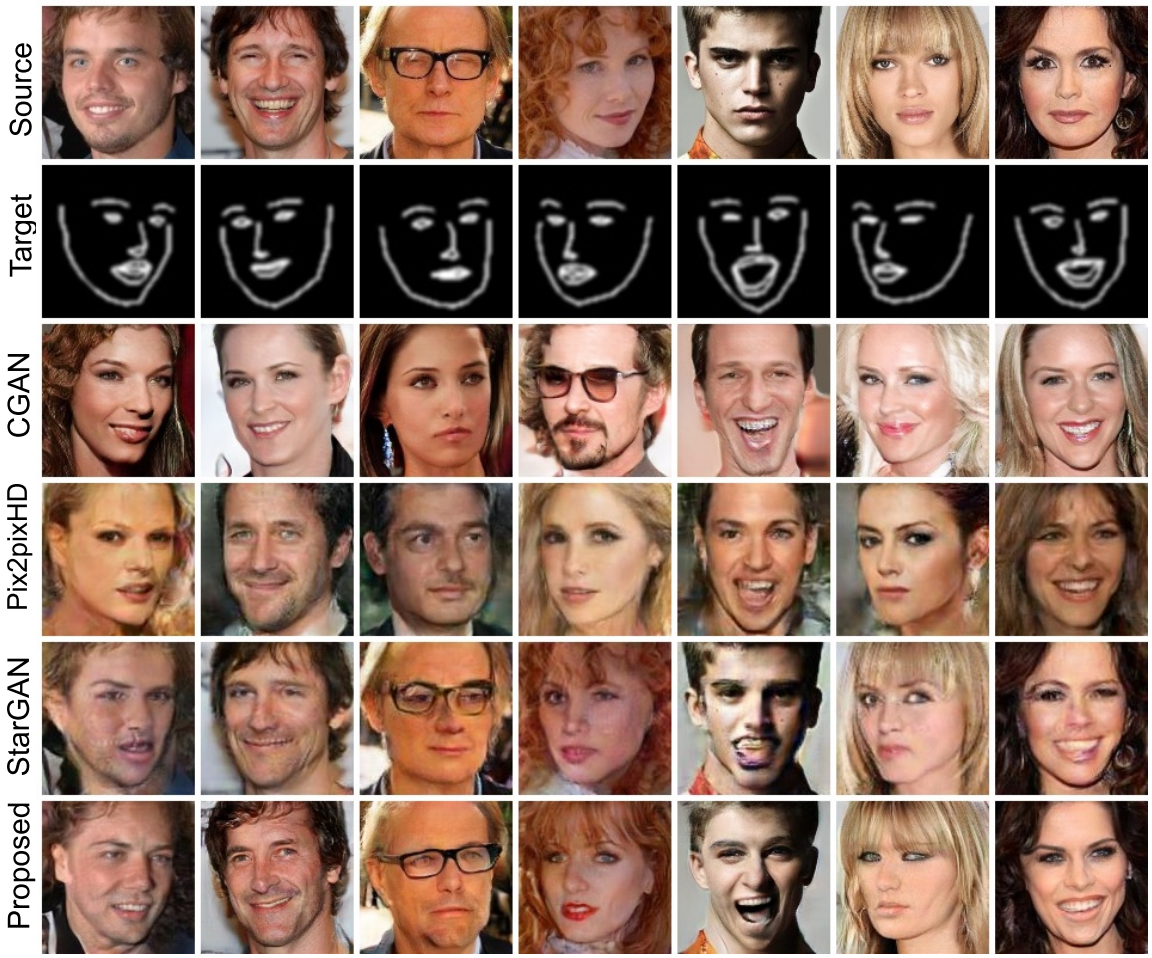


Figure 6.7: CELEBA state-of-the-art comparison. Additional results from CELEBA dataset with respect to the baseline CGAN, and state-of-the-art methods, Pix2pixHD, StarGAN are shown above.

the encoder (i.e. the appearance embedding) is then concatenated with the pose encoding and then sent to the generator. No other feature integration takes place in this variant. The encoder is trained along with the pose-synthesis network. The second variant (A.1.2) is even more similar to the method, as features from the decoder part of the auto-encoder are progressively integrated to the pose-synthesis network in a similar fashion to that of the proposed pipeline.

A.2 – Different variants of ATM: The second key feature of the method is the proposed ATM for effectively transferring the appearance of the input image while

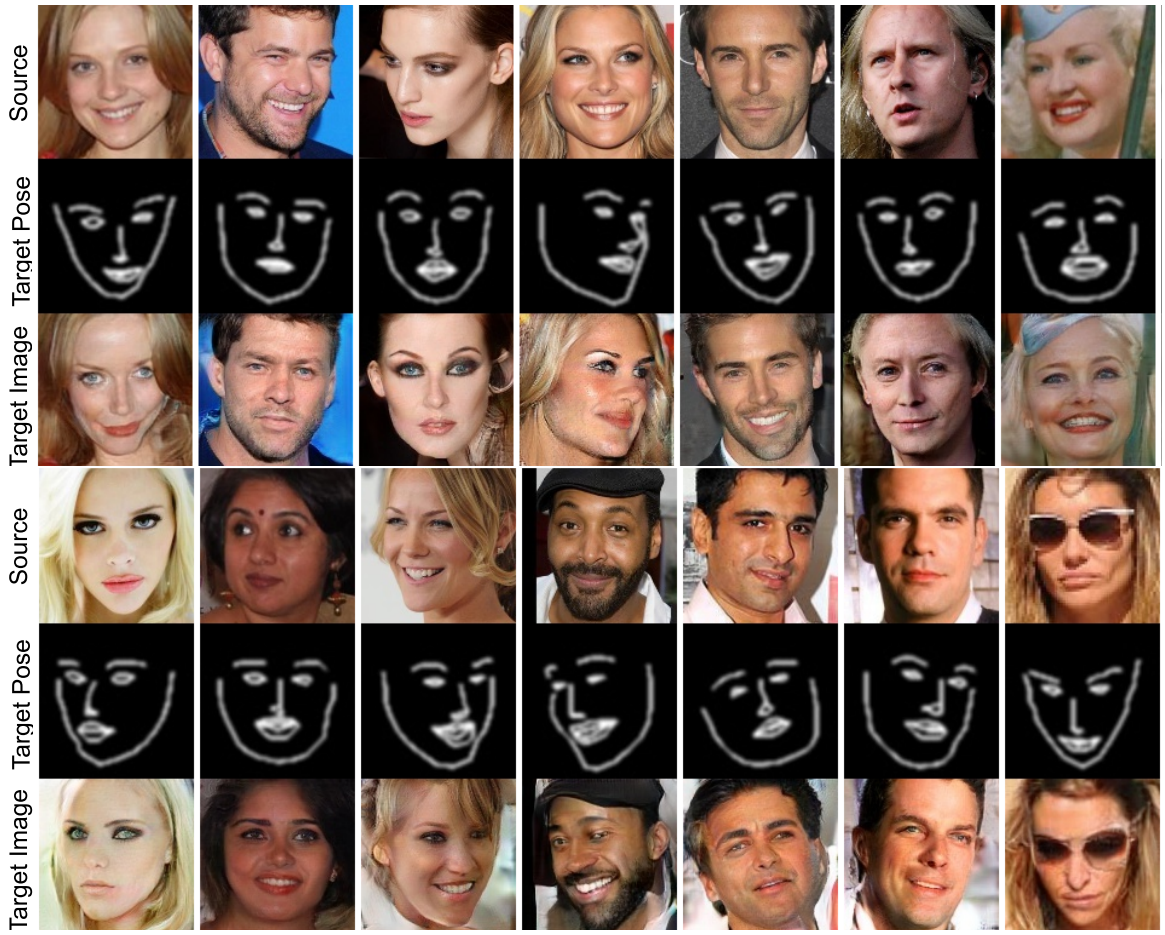


Figure 6.8: CELEBA results. Additional results from CELEBA dataset using the proposed method are shown above, featuring the source image, the target pose and the corresponding generated image.

generating a new facial image in the desired target pose. Two alternatives to the proposed ATM module are explored: In the first variant (A.2.1), the functions ϕ and ψ for producing the spatial masks α and β are materialized using an *Hourglass* (Newell et al., 2016b), reducing the spatial resolution down to 4×4. The second variant (A.2.2) also uses an Hourglass, but this time the module does not learn spatial masks α and β to combine hallucination and pose synthesis features as in Eqn. 6.2, but directly produces $f_{combined}$ as the output of the Hourglass. The input to the Hourglass is the concatenation of hallucination and pose synthesis after the AdaIN layer.

The decoder mirrors the encoder, with convolutions being replaced by transposed convolutions. The encoder-decoder block includes skip connections between the encoder and decoder layers. An alternative with no skip connections was investigated,

Figure 6.9: Qualitative comparison -identity preserving: (a) Source image, (b) Proposed method without identity, (c) Proposed method with identity.

Figure 6.10: MultiPIE pose and expression manipulation in extreme profile views. Row 1 shows the input image and row 2 shows the corresponding pose and expression transfer.

although the generated images were not preserving the input appearance. In this setting, the approach is to concatenate the SR features coming from the hallucination network with the PS features coming from the pose-synthesis network, after being first brought to the SR feature distribution through an AdaIN normalisation layer. The concatenated features are downsampled through a 1x1 convolution and then passed through a small hourglass (Newell et al., 2016b) module with skip connections.

Figure 6.11: MultiPIE face rotation. Column 1 shows the source image and columns 2-5 show face rotation results.

A.3 – Identity preserving network: Another important contribution of the work is the identity preserving network trained with the contrastive loss. To study its influence, the approach was also trained without it.

We study the influence of the reconstruction loss and the identity network in the produced results. To this end, the framework first trains (A.3.1) a network using only the hallucination network, the pose-synthesis with ATM blocks, and the discriminator, without the reconstruction loss and identity network. Then, the method trains a network using the same configuration and the reconstruction loss (A.3.2). The influence of adding the identity preserving network at the full 128×128 resolution.

A.4 – Expression and pose synthesis: The experiments show qualitatively the performance of the network at producing specific expression and pose images on MultiPie dataset.

6.5.3 Results and discussion

Quantitative results in terms of FID and IS for the A.1-A.3 are shown in Table 6.3. The results for A.3, in terms of average pairwise distance and percentage of images with distance under 0.2, are shown in Table 6.4. From the reported results it can be concluded that: (a) that replacing the hallucination network with an autoencoder has detrimental effect in quality, (b) the alternatives to the ATM module yield

poorer results than the ATM configuration, and (c) the identity preserving network largely improves identity preservation, despite not yielding the best FID and IS scores.

Moreover, **qualitative** results for A.4 are shown in Fig. 6.11 and Fig. 6.10. It can be seen that the network can successfully manipulate the pose or the facial expression, generating images of high quality.
Fig. 6.10.

	Method	FID↓	IS↑
	Real data	0.00	3.01
(A.1.1)	AE-v1	35.06	2.53
(A.1.2)	AE-v2	10.05	2.20
(A.2.1)	Hourglass-v1	5.53	2.49
(A.2.2)	Hourglass-v2	5.18	2.49
(A.3)	Proposed w/o identity	**4.31**	**[2.57]**
	Proposed	**[4.6]**	**2.61**

Table 6.3: Quantitative comparison of different baselines and ablation studies for CelebA. Bold numbers are best performance, and bold numbers in brackets indicate second best.

Method	Mean Dist.↓	TPR↑
StarGAN	0.21	56%
Proposed w/o identity	0.22	48%
Proposed	**0.18**	**69**%

Table 6.4: Identity preserving results on CelebA

Importance of hallucination network: The reported results show the importance of having a hallucination network to transfer the input appearance, where the network appears to capture the relevant features at multiple scales, which seems crucial for an accurate transfer. When this module is replaced by an autoencoder, as described in A.1, the quality of the images deteriorate drastically. This is attributed to the fact that the autoencoder formulation introduces an unnecessary complexity into the process, that of learning a latent representation which typically is not well disentangled in terms of pose and appearance. As a result the decoder part of the autoencoder processes entangled features which cannot be easily integrated into a pose synthesis

network trained under the challenging unpaired setting.

Identity preservation: Table 6.4 shows the average source/generated image distance in the feature space, as well as the percentage of pairs with a threshold under 0.2. the method with the identity preserving network outperforms StarGAN as well as the variant of the method that doesn't use this extra network (A.3.2).The qualitative results showing the effectiveness of the identity preservation network is shown in Fig. 6.9.

The encoder does not need to explicitly capture the relevant facial features for an image-to-image translation task. When replacing the hallucination network by an encoder-decoder, the generated images were of very poor quality.

Figure 6.12: Failure cases featuring extreme poses, occlusions.

Limitations and failure cases: Failure cases typically include difficult target poses, and occlusions in the source images as shown in Fig. 6.12.

6.6 Data Anonymisation For Privacy Preservation

From the previous section it is clear that the ability to generate images of faces in various poses and expressions is a challenging task but finds significant applications in face analysis research including several including data augmentation, face editing, expression transfer and face anonymization. Advancements in computer vision are continuously improving the quality and realism of the generated images. Privacy is a critical aspect in face analysis research, especially with sensitive datasets in mental health digital bio-marker analysis. Face anonymization enhances privacy by removing identity information from images and preventing unauthorized use of facial recognition software.

Traditionally, face anonymization has been accomplished through obfuscating pixels, such as pixel permutation, blurring, and pixelation. Previous studies have also revealed the vulnerability of traditional anonymization techniques, such as eye masking, blurring, pixelation and permutation, which are either fully or partially reversible and susceptible to deanonymization efforts. Conversely, anonymization achieved through image synthesis-based methods has been demonstrated to be resistant and could not be broken for deanonymization, as evidenced by existing research Todt et al. (2022). An additional advantage of image synthesis is its ability to generate realistic and high fidelity facial images while preserving facial style attributes, such as hair color, skin color, and eye color, without altering the background. This feature makes face anonymization utilizing generative methods a promising direction of research to achieve robust and realistic face anonymization.

The face-manipulation method proposed in section 6.1 can be applied to anonymize the facial identity information in the Mood-Seasons dataset, demonstrating a proof-of-concept technology for high-fidelity face anonymization. The original framework includes a component to preserve identity information, but this can be omitted to generate anonymous versions of the input image in any desired pose. The goal of face anonymization is to create realistic images that retain the overall appearance of the face while removing identifying features, enabling sharing of images without compromising privacy.

6.6.1 Anonymization Using The Proposed Method

The components of the proposed framework also provide additional benefits to the anonymization process, where it allows for manipulating pose and expression independently through the pose-conditioned GAN. The conditional discriminator enforces the pose synthesis network to adhere to a given pose and expression and be highly realistic. The identity is enforced through the identity preserving network and contrastive training loss components. These additional constraints enforcing identity preservation can be simply removed from the framework by removing the identity preserving network and the contrastive loss component L_{con} by essentially setting the λ_{con} to 0 in 6.10. The resulting architecture of the network, without the

identity preserving component, is provided in Figure 6.13.

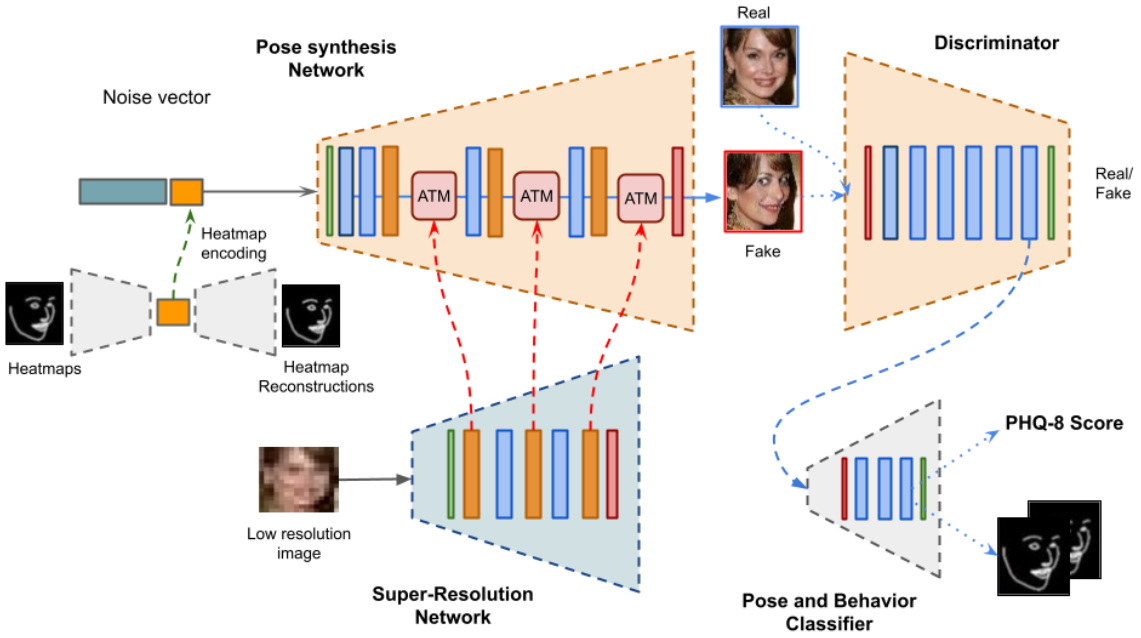


Figure 6.13: The modified architecture of the face manipulation framework where the identity preserving network is removed and a behavior regression head is added to the classifier network. Note that in the current set up, the behavior component is not included since the model is not trained on the Mood-Seasons dataset.

The network is only used for inference and no training is required for anonymizing face images. The Mood-Seasons dataset comprises a set of videos from 134 different identities. In order to test the data anonymization capacity of the approach, a subset of the videos from 50 identities were randomly chosen. For anonymizing the video X_i for an identity i, there are $X_i = x_{i1}, x_{i2}, x_{i3}, ..., x_{in}$ representing a set of face image frames and $Y_i = y_{i1}, y_{i2}, y_{i3}, ..., y_{in}$ representing edge maps from the landmarks from each of the frames. The approach takes as input a 16×16 low resolution, source image, x_s that represents the overall style and background of the target image, which is kept the same for all the frames. This is to ensure that the style remains uniform in the video and is chosen so that the face image is clearly visible without occlusions or blur and has adequate illumination. The edge maps are then iteratively passed through the network along with the fixed source image x_s. The pose synthesis generative network then generates an anonymized image in the desired pose provided by the edge map.

The evaluation of identity obfuscation can be quantified by assessing the performance of face recognition. This is accomplished by utilizing a pre-trained face recognition network to extract features from both the original image (containing the identity) and the anonymized image. The similarity between these features is then calculated, with the ideal outcome being that the anonymized features are dissimilar to the reference identity features.

In order to accomplish this, a video of a person with N frames is used, with a reference frame I_{ID} chosen to represent the identity and a randomly selected key point frame I_k used to provide target landmarks P_k and serve as a comparison frame for the face recognition network. The output of the anonymization framework is given by $I_{anonymized} = G(I_{ID}, P_k)$.

The identity similarity between the fixed source frame I_{ID} and another frame in the video (I_k) as well as its corresponding anonymized frame ($I_{anonymized}$) is calculated using Euclidean distance, as follows: $D_{anonymized} = \left|F(I_{ID}) - F(I_{anonymized})\right| 2$ and $Doriginal = |F(I_{ID}) - F(I_n)|_2$. Both the distances are calculated for a subset of 50 videos and provided in the Figure 6.15. The average distance between original identity frames is 0.67 and the average distance between the original and anonymized identity frames is 1.12. A threshold of 0.8 is used to decide if the features correspond to the same identity. The face recognition network used is FaceNet (Schroff et al., 2015).

Figure 6.14 illustrates sample outcomes from the anonymization process. It can be observed that the original identity information is successfully removed while preserving the background and style attributes such as gender, skin color, hair color and style, facial hair, head accessories etc. The generated image closely follows the edge map derived from the key points and as a consequence gaze direction is not preserved, as the 68 key points does not include the iris. This can be mitigated in future by incorporating landmarks for the eyes into the edge map. The same source frame is used to generate the style for all the images in the video, to enforce identity similarity within a video. Therefore, choosing a frame without occlusions like hands over the chin or poor lighting conditions is important.

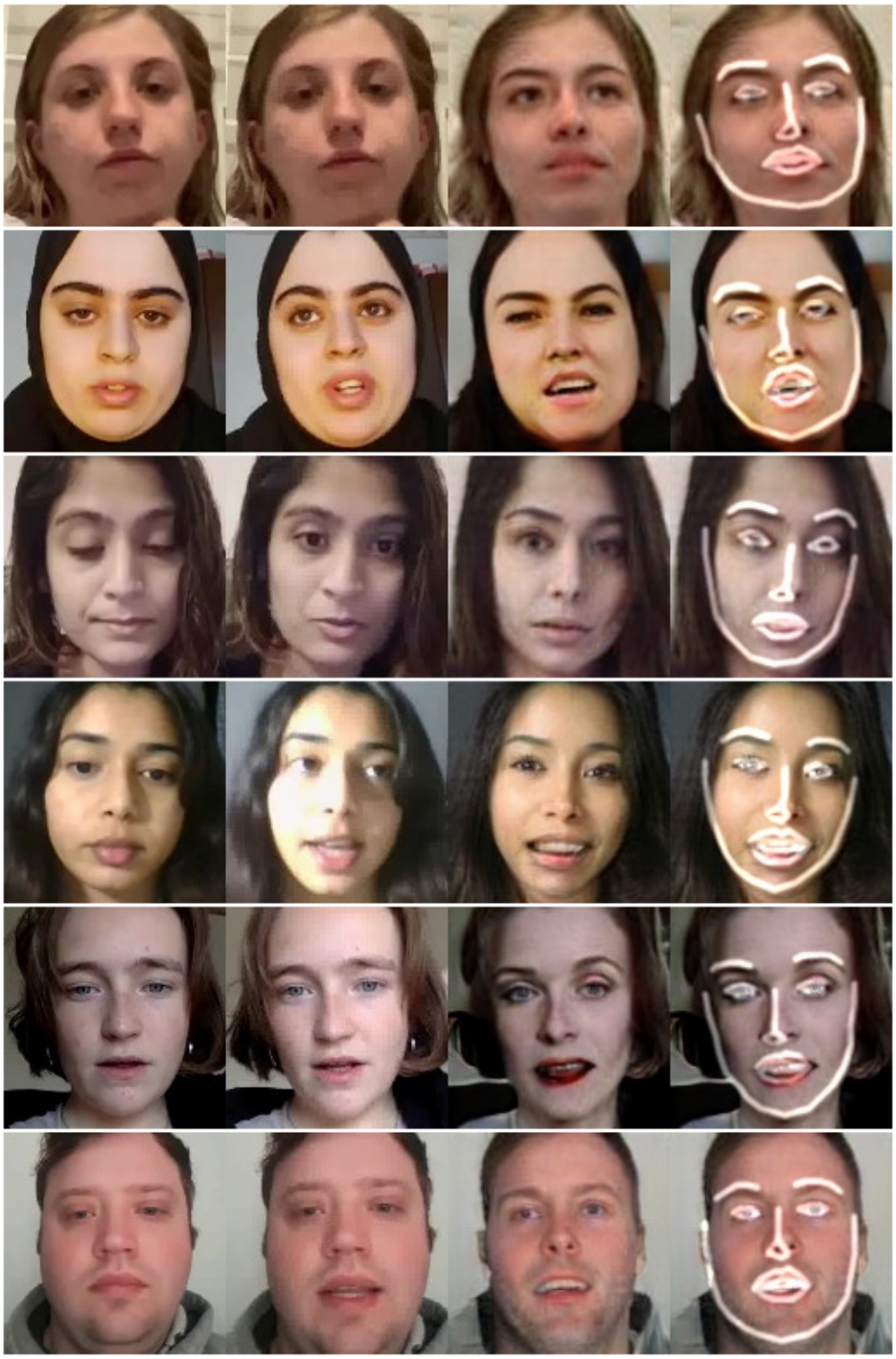


Figure 6.14: Qualitative Results – first column represents the reference image, second column shows the key point frame from the same video and the third and fourth column shows the anonymized face image and edge maps respectively

6.6.2 Evaluating Anonymisation Efficacy in Downstream Tasks

To evaluate the usefulness of the anonymisation method for downstream depression analysis, a baseline depression recognition model trained using the Resnet-50 archi-

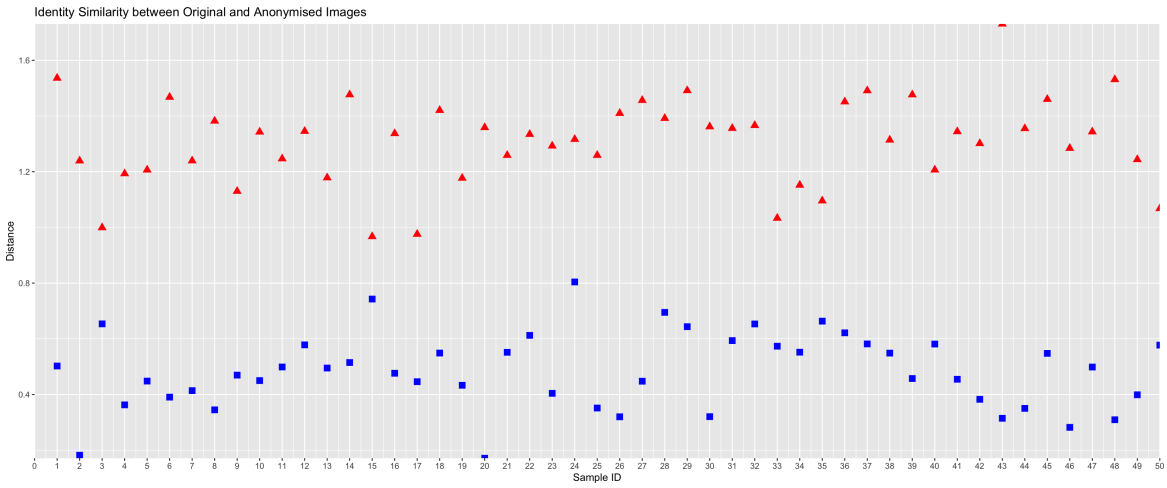


Figure 6.15: Quantitative comparison of face recognition similarity between original images and anonymized images. X-axis represents the individual identities and Y-axis represents the distances of the reference frame to that of the key point frame (in blue) and anonymized frame(in red).

tecture to predict frame level depression labels was used. The model was trained on the Mood Seasons dataset and for the purpose of the experiment the model was evaluated on 10 images selected from the testing set of Mood Seasons dataset. These 10 images were anonymised using the proposed anonymisation technique and the predictions of the model on the original and anonymised images are reported in Table 6.6. The prediction scores are compared and if the scores fall within a severity bin then it is considered as having high similarity.

Visualizing the results as a graph highlights the overall promising similarity in severity bins, while also calling attention to specific instances where anonymization altered the severity assessment. This can help guide efforts to improve anonymization methods. Figure 6.16 plotted the 10 image examples on the x-axis, with the original PHQ-8 score as a blue bar and the anonymized score as an orange bar.The y-axis shows the PHQ-8 score range from 0 to 24. The shaded regions divide this into the 4 severity bins, specifically, no depression, mild depression, moderate depression, moderately severe depression and severe depression. Looking at the severity bins, 6 of the 10 image pairs have high similarity, with the anonymized score falling in the same severity category as the original. All the other pairs have medium similarity, with the anonymized score shifting by one category.

Table 6.5: PHQ Score Similarity

Image	Original PhQ Score	Anonymized PhQ Score	Score Similarity
1	12	10	High
2	5	8	High
3	16	10	Medium
4	20	17	High
5	18	11	Medium
6	8	6	High
7	3	5	Medium
8	17	20	High
9	11	15	Medium
10	19	10	Medium

Table 6.6: Quantitative comparison of depression recognition scores between original images and anonymized images.

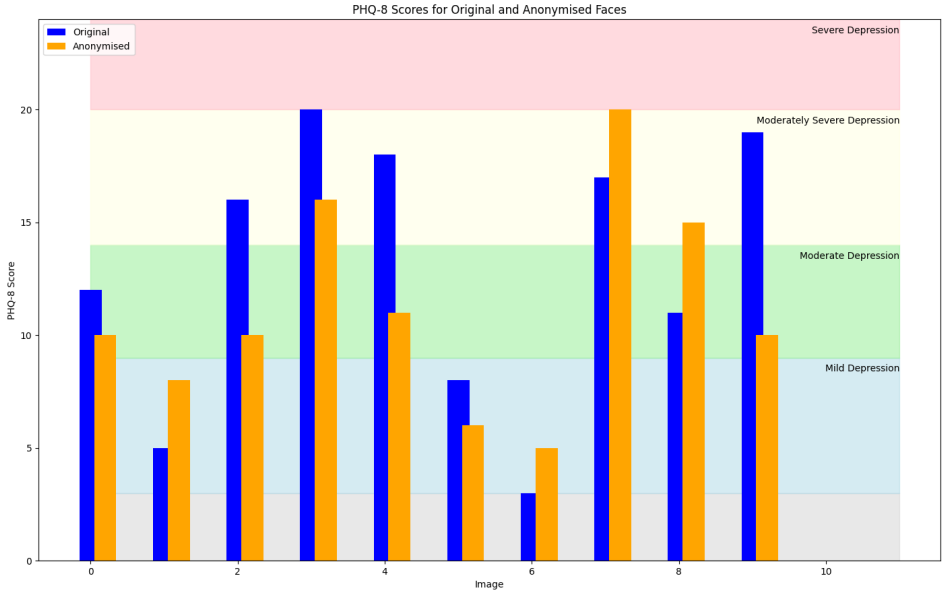


Figure 6.16: Quantitative comparison of depression recognition scores between original images and anonymized images. X-axis represents the individual identities and The y-axis shows the PHQ-8 score range from 0 to 24.

Overall, the anonymization preserved severity level 60% of the time. These mixed results illustrate that while anonymity-preserving methods have potential, there is significant room for improvement. Finding ways to better retain diagnostic facial

information after anonymization is an important direction for future work. Analyzing prediction consistency, as shown here, can quantitatively guide improvements to anonymization techniques. This method highlights areas for improvement to maintain severity bin consistency, especially for more severe depression levels. Analyzing score bin similarity on a larger dataset could further guide anonymization algorithm development.

6.6.3 Future Directions

The previous section provided a proof-of-concept demonstration of the newly introduced face manipulation method from section 6.1 in anonymizing the Mood-Seasons dataset. This method can be applied to securely store and share sensitive datasets used in depression recognition.

However, in order to further advance the proposed framework and make the anonymized version of the dataset suitable for behavior analysis, further investigation is necessary. The objective is to obscure identity characteristics while maintaining diagnostic information, so that privacy preservation does not negatively impact the performance of the downstream task and the digital markers remain intact.

To achieve this, the proposed framework should incorporate another behavior alignment component. A straightforward approach would be to retrain the network on the Mood-Seasons dataset, and add a PhQ-8 score regression head to the discriminator. This would task the discriminator with minimizing the discrepancy in behavior expression between the original and generated samples, thereby ensuring that the anonymized version of the dataset retains diagnostic information and that privacy preservation does not negatively impact the performance of the downstream task.

Additionally, the proposed framework should also consider other metrics to evaluate the preservation of diagnostic information in the anonymized version of the dataset. For instance, the performance of PhQ-8 regression on the dataset could be used as a metric for retention of diagnostic information. Another evaluation would be to find the correlation between objective behavioral analysis components like action units, facial expressions etc. between original and anonymized versions of the dataset. These metrics can ensure that the anonymized version of the dataset is still informa-

tive for behavior analysis.

In conclusion, the proposed framework should consider multiple aspects to ensure that privacy preservation does not negatively impact the performance of the downstream task and that the diagnostic information is retained in the anonymized version of the dataset.

6.7 Conclusion

This chapter introduced a framework for face generation that learns to synthesize novel face images that adhere to a given pose, whilst transferring appearance and style information from an exemplar image, in a semantically meaningful way. It includes an in-depth discussion of how to integrate appearance features of the exemplar image taken from a pre-trained hallucination network into the generation process of a conditional GAN using a novel appearance transfer module. The experiments then demonstrate both quantitatively and qualitatively the capability of the method to achieve high quality images that are both conditioned on target poses and source appearances.

The last section of the chapter applies the novel method in anonymizing the identities present in the Mood-Seasons dataset as a proof-of-concept. It discussed the modifications applied to the approach to accomplish face anonymization and measures the degree of anonymization achieved using the proposed method using quantitative and qualitative analysis. Further, the last section proposed suggestions and lays out a road map on how this can be further leveraged for automatic depression analysis.

Chapter 7

Conclusion

This thesis presented a new multi-modal, longitudinal dataset collected through a custom smartphone app for depression recognition in real-world settings, along with a state-of-the-art automated depression recognition system utilizing multimodal data. The thesis also introduced a unique privacy-preserving approach to anonymising face images in the dataset using generative methods. This chapter summarises the approaches presented in the thesis along with a discussion of limitations of the studies and suggests future directions for research in this area.

In the fourth chapter, a large-scale data collection study was conducted to generate a novel, multimodal (audio-video-text) and longitudinal dataset called Mood-Seasons. The dataset was collected using a smartphone in natural, in-the-wild environments and included video and audio recordings, as well as textual transcriptions, from the general public. The severity of depression was measured using responses to a PHQ-8 questionnaire.

The study also revealed that self-administered questionnaire responses are comparable to human or virtual-human mediated questionnaire responses. The data gathering methods were designed thoroughly and ethically, and the app was developed and implemented with lessons learnt from launching such an app and the general public's perspective of it. Strategies were discussed to boost engagement and use of smartphone-based mental health data gathering techniques.

The Mood-Seasons dataset was analyzed in terms of the distribution of depression severity scores across numerous factors such as depression categories, age, gender,

and race. A longitudinal analysis of the PHQ scores showed that the relative PHQ-8 scores over the course of three weeks remained steady, showing the longitudinal persistence of mood.

Despite successful data collection, the study had limitations. The app did not include a notification setting to automatically remind participants to complete the study and add a recording. Additionally, the database back-end used in the app only stored the final PHQ-8 score instead of scores against individual questionnaire elements. Collecting responses at each item level would provide a fine-grained insight into the participant's mood states.

Future studies can address these limitations by including a notification setting, collecting responses at each item level, and optimising data storage to eliminate a cap on video capturing time. Collecting a mood state indicator from participants would also enable further understanding of self-perceived mood against PHQ-8 scores, facilitating the development of systems that might be able to correlate personalised mood state with responses to the PHQ-8 score.

Chapter 5 presented an extensive benchmark of state-of-the-art video analysis techniques on the newly collected Mood-Seasons and publicly available AVEC 2014 datasets. The chapter presented a quantitative comparison between several methods, including spatial and temporal models, and showed that models with understanding of the temporal context showed the best performance.
The 3D temporal models, C3D and TSM models showed competitive performance on the Mood-Seasons dataset and AVEC 2013 datasets, whereas the state-of-the-art video recognition model, SlowFast, performed poorly when compared to other temporal models. Among the 3D models, some, such as I3D and 3D Resnet, demonstrated relatively lower performance, while others, such as those utilising temporal aggregation methods like GRUs and attention mechanisms showed competitive performance.

The second part of chapter 5 presented a two-staged multimodal transformer-based approach that addresses automated depression severity prediction and provided experiments to validate the effectiveness of the approach on both Mood-Seasons and AVEC 2014 datasets. Multimodal data was analysed using short- and long-range modelling transformer architectures. A novel loss called differential loss was in-

troduced to improve performance by leveraging multiple videos from one person. Various ablation studies were also conducted to assess the effectiveness of different components of the framework in predicting the severity of depression.

In comparison to the benchmark models, the multimodal transformer architecture outperformed benchmark methods by significant margins, especially the best-performing spatio-temporal model, C3D, in terms of the metrics MAE and RMSE respectively. On the testing set, which appears to be slightly more challenging compared to the validation set, the multimodal transformer model outperformed the top-performing benchmark, TSM, on the Mood-Seasons dataset, on MAE and RMSE respectively. These results demonstrate the superior performance of the proposed multimodal framework in predicting the severity of depression.

The video-level multimodal transformer showed considerable improvement in RMSE compared to the sentence-level multimodal transformer. These performance gains show that the use of video-level understanding and long-range modelling can lead to a more accurate automatic understanding of depression. Another ablation study showed that co-attention is a crucial component of multimodal feature fusion in the transformer architecture. The experiments also showed that the differential loss function was effective in improving the performance of the estimation of the severity of depression.

Interpretability is an important aspect of clinically relevant diagnostic tools. The methodology presented in the chapter focused on deep learning based techniques to analyse real-world, multimodal depression data. Investigating the interpretability of multimodal features learnt by the attention model would be helpful in understanding the decisions made by the model. A complementary route would be to use traditional hand-crafted features pertaining to different symptoms or human-interpretable traits, such as head movement, eye gaze directions, facial action units, voice features such as prosody, pitch, speech rate and language features. These would be a useful baseline for correlating automatically detected digital biomarkers to PHQ-8 symptoms.

Chapter 6 of the study introduced a framework for generating novel face images that adhere to a given pose while transferring appearance and style information from an exemplar image. The framework includes a discussion of how to integrate

appearance features from a pre-trained hallucination network into the generation process of a conditional GAN using a novel appearance transfer module. The experiments demonstrate the method's ability to achieve high-quality images that are both conditioned on target poses and source appearances.

The last section of the chapter applied the method to anonymise identities in the Mood-Seasons dataset as a proof-of-concept. Modifications applied to the approach to achieve face anonymisation were discussed and the degree of anonymisation achieved was measured using quantitative and qualitative analysis. The section proposed suggestions and laid out a roadmap on how the method could be leveraged for automatic depression analysis.

However, the proposed framework needs further investigation to make the anonymised version of the dataset suitable for behaviour analysis. To achieve this, the proposed framework should incorporate another behaviour alignment component. This would task the discriminator with minimising the discrepancy in behaviour expression between the original and generated samples, ensuring that the anonymised version of the dataset retains diagnostic information. Other metrics, such as PhQ-8 regression performance and correlation between objective behavioural analysis components, should also be considered to evaluate the preservation of diagnostic information in the anonymised dataset.

In conclusion, the proposed framework should consider multiple aspects to ensure that the preservation of privacy does not negatively impact the performance of the downstream task and that diagnostic information is retained in the anonymised version of the dataset. The framework could be applied to securely store and share sensitive datasets used in depression recognition.

One of the main limitations of the study is that the method is evaluated only for the degree of anonymisation and not extensively for the degree of behaviour preservation. As mentioned earlier, the proposed framework needs further investigation to make the anonymised version of the dataset suitable for behaviour analysis. The architecture of the network needs to be modified to include a behaviour preservation loss, as detailed in Chapter 6. This loss would task the discriminator with minimizing the discrepancy in behaviour expression between the original and generated samples,

ensuring that the anonymised version of the dataset retains diagnostic information.

Another limitation of the study is that the model is pre-trained on the CELEBA dataset, which is biased towards certain race and gender. This bias could result in lower quality results on race or gender outside the training distribution. This limitation can be mitigated by fine-tuning the model on the Mood-Seasons dataset, which includes a more diverse population.

One major limitation of the collected data was the sample population. Data were collected from the general public, and the ethical policies for the data collection study did not allow people with a valid clinical diagnosis of depression to participate. This means that a smaller percentage of participants showed symptoms of higher depression severity.

Another limitation was the demographics of the study sample. The average participant was a 27-year-old female. Although all efforts were made to recruit participants from various backgrounds, a large number of young women enrolled in the study. Therefore, the results are more representative of that population.

In summary, while the proposed framework shows promise in achieving high-quality image generation and face anonymisation, further investigation is necessary to ensure that behaviour preservation is also achieved. Additionally, bias in the pre-trained model can limit the quality of results for certain race or gender groups, but this can be addressed by fine-tuning the model on the target dataset.

7.1 Future Work

The proposed framework needs further investigation to make the anonymised version of the dataset suitable for behaviour analysis related to depression. Evaluating the anonymised dataset just on degree of anonymisation is insufficient. Metrics to measure how well mood, emotions, and behaviours are preserved after anonymisation should be included. This could include evaluating performance of pretrained models for depression recognition, emotion classification, and facial action unit detection on the original vs. anonymised dataset.

Measurement of the correlation of objective behavioural markers before and after anonymisation would help quantify information retention. Computing correlations between low-level behavioural cues, such as facial action units, head movements, and speech patterns, on the original and anonymised data is crucial. The high correlation suggests that useful signals are retained. Metrics like k-anonymity and l-diversity could be used to quantify the anonymity level achieved. Fine-tuning the anonymisation model on the Mood-Seasons dataset rather than just using the pretrained model on the CELEB-A dataset would help make it more robust to the diverse gender and racial groups represented in the new data.

Beyond quantitative metrics, a qualitative study to get feedback from clinicians on the utility of the anonymised dataset for diagnosing mood disorders would be helpful. Conducting manual ratings of perceived depression by clinicians viewing anonymised videos compared to originals would provide insights into effective mood preservation. The gathering of direct feedback from study participants on their perceived anonymity after viewing anonymised videos of themselves should also be considered. High self-reported anonymity suggests effective identity masking.

Incorporating co-design with patients in developing anonymisation techniques can help ensure they balance privacy protection and preservation of diagnostic signals relevant for depression assessment. Getting input from the users themselves on what they consider private information and what behavioural cues are acceptable to alter would help guide the development.

Interviews/focus groups could help understand patient priorities and concerns around the use of their visual data. Participatory design sessions to actively involve patients in generating and critiquing anonymisation approaches should be incorporated. Future studies should gather feedback from patients on anonymised versions of their own videos to guide refinements. Collaboration to develop appropriate consent processes that give patients control over anonymisation procedures performed on their data should be carried out. Recent literature provides examples of successful co-design in mental health contexts, such as Rennick-Egglestone et al. (2019) and Torous et al. (2019). Adopting human-centred design practises will be key to developing ethically and socially acceptable solutions for privacy protection in mental health research.

Exploring different levels of anonymisation, from minimal changes to complete facial replacement, could reveal the trade-offs between privacy and behaviour preservation. Participants could opt in to their preferred level. Developing privacy-preserving techniques in partnership with patients directly addresses important ethical concerns and helps build trust. This is an exciting area for future work.
Stepping back, the pandemic has accelerated the adoption of virtual mental health solutions. While machine learning and computer vision technologies offers tantalizing possibilities, it also raises complex ethical questions around privacy, consent, bias and transparency. Developing such sensitive technologies responsibly requires cross-disciplinary collaboration between engineers, clinicians and social scientists.

Overall, AI should act as an enhancer, not a replacement, for human understanding - combining the strengths of both to expand access to mental healthcare. The proposed approach involving multimodal transformers demonstrate the potential of automated depression analysis for augmenting clinicians' capabilities for objective mental health measurement. However, transparent and interpretable automated depression analysis is crucial for clinicians to trust and adopt these tools. Advances in explainable AI tailored to mental health data analysis would enable physician-automated system collaboration with clinicians leveraging AI as a supportive tool rather than a black box.

On the data collection side, smartphones offer an invaluable platform for gathering rich longitudinal mental health data at scale. However, creative solutions like gamification, active notifications and clear consent processes are imperative to drive user engagement while respecting privacy. As the anonymisation methods highlighted, true privacy-preserving data sharing for research remains an open challenge. Co-designing solutions with patients and using formal privacy measures are important ways forward.

Ultimately, realizing the full potential of automated depression analysis in mental healthcare requires just as much ethical foresight as technical innovation. With collaborative, human-centric design, computer vision and machine learning technologies can widen access to quality mental health support and create a more psychologically flourishing society. But developing such sensitive technologies responsibly demands

cross-disciplinary collaboration between engineers, clinicians and social scientists. Overall, a symbiotic human-AI approach is needed - combining the empathetic, contextual understanding of people with the vast data insights of machines.

This thesis contributed to methodological advances in multimodal depression detection and privacy-preserving data sharing. But these are just the first steps on the path to integrate AI ethically and responsibly into mental healthcare. With an eye towards the human impact, computer vision and machine learning can help democratise access to mental health support and move us toward a society where everyone has the tools to tend to their psychological wellbeing.

Bibliography

Al Jazaery, M. and Guo, G. (2018). Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., et al. (2012). From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, volume 19.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M. (2013a). Eye movement analysis for depression detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 4220–4224. IEEE.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M. (2013b). Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*, pages 4220–4224. IEEE.

Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., and Breakspear, M. (2013c). Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 283–288. IEEE.

Alpert, M., Pouget, E. R., and Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

American Psychiatric Association, A., Association, A. P., et al. (2013). Diagnostic and statistical manual of mental disorders: Dsm-5.

Ameringen, M., Turna, J., Khalesi, Z., Pullia, K., and Patterson, B. (2017). There is an app for that! the current state of mobile applications (apps) for dsm-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. *Depression and anxiety*.

Arjovsky, M. and Bottou, L. (2017a). Towards principled methods for training generative adversarial networks. *ArXiv*, abs/1701.04862.

Arjovsky, M. and Bottou, L. (2017b). Towards principled methods for training generative adversarial networks.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *ArXiv*, abs/1701.07875.

Association, A. P. et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Baer, L. and Blais, M. A. (2010). *Handbook of clinical rating scales and assessment in psychiatry and mental health*. Springer.

Bakker, D., Kazantzis, N., Rickwood, D., and Rickard, N. (2016). Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR mental health*, 3(1).

Bellet, A., Habrard, A., and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.

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

Bulat, A. and Tzimiropoulos, G. (2016). Convolutional aggregation of local evidence for large pose face alignment.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks).

Bulat, A. and Tzimiropoulos, G. (2018). Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117.

Burns, N. M., Begale, M., Duffecy, J., Gergle, D., Karr, J. C., Giangrande, E., and Mohr, C. D. (2011). Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res*, 13(3):e55.

Cai, H., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., Li, J., Yang, Z., Li, X., Zhao, Q., et al. (2020). Modma dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Chao, L., Tao, J., Yang, M., and Li, Y. (2015). Multi task sequence learning for depression scale prediction from video. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 526–531. IEEE.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Çiftçi, E., Kaya, H., Güleç, H., and Salah, A. A. (2018). The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F. (2009a). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F. (2009b). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., Cipresso, P., Riva, G., and Botella, C. (2019). Current state and future directions of technology-based ecological momentary assessment and

intervention for major depressive disorder: a systematic review. *Journal of clinical medicine*, 8(4):465.

Contributors, M. (2020). Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015a). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015b). A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49.

De Melo, W. C., Granger, E., and Hadid, A. (2019). Depression detection based on deep distribution learning. In *2019 IEEE international conference on image processing (ICIP)*, pages 4544–4548. IEEE.

de Melo, W. C., Granger, E., and Hadid, A. (2020). A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Transactions on Affective Computing*.

De Melo, W. C., Granger, E., and Lopez, M. B. (2020). Encoding temporal information for automatic depression recognition from facial analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1080–1084. IEEE.

de Melo, W. C., Granger, E., and Lopez, M. B. (2021). Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing*.

Depp, C. A., Mausbach, B., Granholm, E., Cardenas, V., Ben-Zeev, D., Patterson, T. L., Lebowitz, B. D., and Jeste, D. V. (2010). Mobile interventions for severe mental illness: design and preliminary data from three approaches. *The Journal of nervous and mental disease*, 198(10):715.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhall, A. and Goecke, R. (2015). A temporally piece-wise fisher vector approach for depression analysis. In *2015 International conference on affective computing and intelligent interaction (ACII)*, pages 255–259. IEEE.

Dibeklioglu, H., Hammal, Z., and Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*.

Dibeklioğlu, H., Hammal, Z., Yang, Y., and Cohn, J. F. (2015). Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM.

Dogan, E., Sander, C., Wagner, X., Hegerl, U., and Kohls, E. (2017). Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? systematic review. *Journal of medical Internet research*, 19(7):e262.

Dogrucu, A., Perucic, A., Isaro, A., Ball, D., Toto, E., Rundensteiner, E. A., Agu, E., Davis-Martin, R., and Boudreaux, E. (2020). Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health*, 17:100118.

Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., and Yin, J. (2018). Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*.

Dong, Y. and Yang, X. (2021). A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing*, 441:279–290.

Du, Z., Li, W., Huang, D., and Wang, Y. (2018). Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 23–30.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., and Wang, B. (2016). Behavior vs. introspection: refining prediction of clinical

depression via smartphone sensing data. In *2016 IEEE wireless health (WH)*, pages 1–8. IEEE.

Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E., Winther, O., Bardram, J. E., and Kessing, L. (2016). Voice analysis as an objective state marker in bipolar disorder. *Translational psychiatry*, 6(7):e856.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2).

Gecer, B., Bhattarai, B., Kittler, J., and Kim, T.-K. (2018). Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model.

Gibbons, R. D., Clark, D. C., and Kupfer, D. J. (1993). Exactly what does the hamilton depression rating scale measure? *Journal of psychiatric research*, 27(3):259–273.

Girard, J. M. and Cohn, J. F. (2015a). Automated audiovisual depression analysis. *Current opinion in psychology*, 4:75–79.

Girard, J. M. and Cohn, J. F. (2015b). Automated audiovisual depression analysis. *Current Opinion in Psychology*, 4:75 – 79. Depression.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Gong, Y. and Poellabauer, C. (2017). Topic modeling based multi-modal depression detection.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gratch, I., Choo, T.-H., Galfalvy, H., Keilp, J. G., Itzhaky, L., Mann, J. J., Oquendo, M. A., and Stanley, B. (2021). Detecting suicidal thoughts: The power of ecological momentary assessment. *Depression and anxiety*, 38(1):8–16.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews. Technical report, University of Southern California Los Angeles.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. 28(5):807–813.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017a). Improved training of wasserstein gans.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017b). Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777.

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., and Narayanan, S. (2014). Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 33–40.

Haddon-Hill, G., Kusumam, K., and Valstar, M. (2021). A simple baseline for evaluating expression transfer and anonymisation in video transfer. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–08. IEEE.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition.

He, L. and Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111.

He, L., Chan, J. C.-W., and Wang, Z. (2021). Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, 422:165–175.

He, L., Guo, C., Tiwari, P., Pandey, H. M., and Dang, W. (2022). Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *International Journal of Intelligent Systems*, 37(12):10140–10156.

He, L., Jiang, D., and Sahli, H. (2018). Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. *IEEE Transactions on Multimedia*.

Hendricks, L. A., Mellor, J., Schneider, R., Alayrac, J.-B., and Nematzadeh, A. (2021). Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*.

Hollis, C., Morriss, R., Martin, J., Amani, S., Cotton, R., Denis, M., and Lewis, S. (2015). Technological innovations in mental healthcare: harnessing the digital revolution.

Hong, R. H., Murphy, J. K., Michalak, E. E., Chakrabarty, T., Wang, Z., Parikh, S. V., Culpepper, L., Yatham, L. N., Lam, R. W., and Chen, J. (2021). Implementing

measurement-based care for depression: practical solutions for psychiatrists and primary care physicians. *Neuropsychiatric Disease and Treatment*, pages 79–90.

Hong, Y., Hwang, U., Yoo, J., and Yoon, S. (2019). How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)*, 52(1):10.

Hu, Y., Wu, X., Yu, B., He, R., and Sun, Z. (2018). Pose-guided photorealistic face rotation.

Huang, R., Zhang, S., Li, T., He, R., et al. (2017). Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510.

Huang, Z., Epps, J., Joachim, D., and Chen, M. (2018). Depression detection from short utterances via diverse smartphones in natural environmental conditions. In *INTERSPEECH*, pages 3393–3397.

Huckvale, K., Venkatesh, S., and Christensen, H. (2019). Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ digital medicine*, 2(1):1–11.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jain, V., Crowley, J. L., Dey, A. K., and Lux, A. (2014). Depression estimation using audiovisual features and fisher vector encoding. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 87–91.

Jaiswal, S., Song, S., and Valstar, M. (2019a). Automatic prediction of depression and anxiety from behaviour and personality attributes. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7.

Jaiswal, S., Valstar, M., Kusumam, K., and Greenhalgh, C. (2019b). Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 81–87.

Jaiswal, S., Valstar, M., Kusumam, K., and Greenhalgh, C. (2019c). Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 81–87, New York, NY, USA. Association for Computing Machinery.

Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F., and Turabzadeh, S. (2014). Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM.

Jan, A., Meng, H., Gaus, Y. F. B. A., and Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680.

Jayawardena, S., Epps, J., and Ambikairajah, E. (2020). Ordinal logistic regression with partial proportional odds for depression prediction. *IEEE Transactions on Affective Computing*.

Joshi, J., Dhall, A., Goecke, R., and Cohn, J. F. (2013a). Relative body parts movement for automatic depression analysis. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 492–497. IEEE.

Joshi, J., Goecke, R., Parker, G., and Breakspear, M. (2013b). Can body expressions contribute to automatic depression analysis? In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE.

Kamath, J., Barriera, R. L., Jain, N., Keisari, E., and Wang, B. (2022). Digital phenotyping in depression diagnostics: Integrating psychiatric and engineering perspectives. *World Journal of Psychiatry*, 12(3):393.

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

Karras, T., Laine, S., and Avila, T. (2019). A style-based generator architecture for generative adversarial networks.

Kauer, S. D., Reid, S. C., Crooke, A. H. D., Khor, A., Hearps, S. J. C., Jorm, A. F., Sanci, L., and Patton, G. (2012). Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. *Journal of medical Internet research*, 14(3).

Kendler, K. S. (2016). The phenomenology of major depression and the representativeness and nature of dsm criteria. *American Journal of Psychiatry*, 173(8):771–780.

Kim, B.-K., Lee, H., Roh, J., and Lee, S.-Y. (2015). Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434. ACM.

Kim, H., Kim, S., Kong, S. S., Jeong, Y.-R., Kim, H., Kim, N., et al. (2020). Possible application of ecological momentary assessment to older adults' daily depressive mood: Integrative literature review. *JMIR Mental Health*, 7(6):e13247.

Kossaifi, J., Tran, L., Panagakis, Y., and Pantic, M. (2018). Gagan: Geometry-aware generative adversarial networks.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kusumam, K., Sanchez, E., and Tzimiropoulos, G. (2021). Unsupervised face manipulation via hallucination. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2406–2413. IEEE.

Lakkis, N. A. and Mahmassani, D. M. (2015). Screening instruments for depression in primary care: a concise review for clinicians. *Postgraduate medicine*, 127(1):99–106.

Li, Y., Huang, C., and Loy, C. C. (2019). Dense intrinsic appearance flow for human pose transfer. In *CVPR*.

Lim, J. H. and Ye, J. C. (2017a). Geometric gan. *arXiv preprint arXiv:1705.02894*.

Lim, J. H. and Ye, J. C. (2017b). Geometric gan. *arXiv preprint arXiv:1705.02894*.

Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild.

Lorenz, D., Bereska, L., Milbich, T., and Ommer, B. (2019). Unsupervised part-based disentangling of object shape and appearance. In *CVPR*.

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., Bamis, A., Russell, A., Wang, B., and Bi, J. (2018a). Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–21.

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., Bamis, A., Russell, A., Wang, B., and Bi, J. (2018b). Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1).

Luxton, D. D., McCann, R. A., Bush, N. E., Mishkind, M. C., and Reger, G. M. (2011). mhealth for mental health: Integrating smartphone technology in behavioral healthcare. *Professional Psychology: Research and Practice*, 42(6):505.

Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., and Fritz, M. (2018). Disentangled person image generation. In *CVPR*.

Ma, X., Yang, H., Chen, Q., Huang, D., and Wang, Y. (2016a). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42.

Ma, X., Yang, H., Chen, Q., Huang, D., and Wang, Y. (2016b). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42. ACM.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802.

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A., Haro, J. M., Kerz, M., Lamers, F., et al. (2019). Remote assessment of disease and relapse in major depressive disorder (radar-mdd): a multi-centre prospective cohort study protocol. *BMC psychiatry*, 19:1–11.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K., Annas, P., De Girolamo, G., Difrancesco, S., Haro, J., Horsfall, M., and et al. (2022). Remote assessment of disease and relapse in major depressive disorder (radar-mdd): Recruitment, retention, and data availability in a longitudinal remote measurement study. *European Psychiatry*, 65(S1):S112–S112.

Maurer, D. M., Raymond, T. J., and Davis, B. N. (2018). Depression: screening and diagnosis. *American family physician*, 98(8):508–515.

McManus, S., Bebbington, P. E., Jenkins, R., and Brugha, T. (2016). *Mental health and wellbeing in England: the adult psychiatric morbidity survey 2014*. NHS digital.

Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., and Wang, Y. (2013). Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30.

Mescheder, L., Geiger, A., and Nowozin, S. (2018). Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*.

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

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Moore II, E., Clements, M. A., Peifer, J. W., and Weisser, L. (2007). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107.

Newell, A., Yang, K., and Deng, J. (2016a). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer.

Newell, A., Yang, K., and Deng, J. (2016b). Stacked hourglass networks for human pose estimation.

Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM.

Niu, M., Liu, B., Tao, J., and Li, Q. (2021). A time-frequency channel attention and vectorization network for automatic depression level prediction. *Neurocomputing*, 450:208–218.

Niu, M., Tao, J., Liu, B., and Fan, C. (2019). Automatic depression level detection via lp-norm pooling. *Proc. INTERSPEECH, Graz, Austria*, pages 4559–4563.

Niu, M., Tao, J., Liu, B., Huang, J., and Lian, Z. (2020). Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Transactions on Affective Computing*.

Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org.

Onnela, J.-P. and Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41(7):1691–1696.

Ooi, K. E. B., Low, L.-S. A., Lech, M., and Allen, N. (2011). Prediction of clinical depression in adolescents using facial image analaysis. In *WIAMIS 2011: 12th International Workshop on Image Analysis for Multimedia Interactive Services, Delft, The Netherlands, April 13-15, 2011*. Citeseer.

Organization, W. H. et al. (2017). Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Pampouchidou, A., Simantiraki, O., Vazakopoulou, C.-M., Chatzaki, C., Pediaditis, M., Maridaki, A., Marias, K., Simos, P., Yang, F., Meriaudeau, F., et al. (2017a). Facial geometry and speech analysis for depression detection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1433–1436. IEEE.

Pampouchidou, A., Simos, P., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., and Tsiknakis, M. (2017b). Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*.

Parikh, S. V. and Huniewicz, P. (2015). E-health: an overview of the uses of the internet, social media, apps, and websites for mood disorders. *Current opinion in psychiatry*, 28(1):13–17.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019a). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019b). Semantic image synthesis with spatially-adaptive normalization.

Pérez, H., Escalante, H. J., Villasenor-Pineda, L., Montes-y Gómez, M., Pinto-Avedano, D., and Reyes-Meza, V. (2014). Fusing affective dimensions and audio-visual features from segmented video for depression recognition. In *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC'14)*, pages 49–55. ACM Orlando, Florida, USA.

Pichot, P. (1986). Self-report inventories in the study of depression. In *New results in depression research*, pages 53–58. Springer.

Price, M., Yuen, E. K., Goetter, E. M., Herbert, J. D., Forman, E. M., Acierno, R., and Ruggiero, K. J. (2014). mhealth: A mechanism to deliver more accessible, more effective mental health care. *Clinical Psychology & Psychotherapy*, 21(5):427–436.

Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., and Moreno-Noguer, F. (2018a). Ganimation: Anatomically-aware facial animation from a single image.

Pumarola, A., Agudo, A., Sanfeliu, A., and Moreno-Noguer, F. (2018b). Unsupervised person image synthesis in arbitrary poses. In *CVPR*.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

Rennick-Egglestone, S., Morgan, K., Llewellyn-Beardsley, J., Ramsay, A., McGranahan, R., Gillard, S., Hui, A., Ng, F., Schneider, J., Booth, S., et al. (2019). Mental health recovery narratives and their impact on recipients: systematic review and narrative synthesis. *The Canadian Journal of Psychiatry*, 64(10):669–679.

Riba, E., Mishkin, D., Ponsa, D., Rublee, E., and Bradski, G. (2020). Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., and Pantic, M. (2019). Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 3–12, New York, NY, USA. Association for Computing Machinery.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017a). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017b). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9.

Robinson, J., Khan, N., Fusco, L., Malpass, A., Lewis, G., and Dowrick, C. (2017). Why are there discrepancies between depressed patients' global rating of change and scores on the patient health questionnaire depression module? a qualitative study of primary care in england. *BMJ open*, 7(4):e014519.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., and Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. *JMIR mHealth and uHealth*, 6(8):e9691.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., and Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 17(7).

Sajatovic, M., Chen, P., and Young, R. C. (2015). Chapter nine - rating scales in bipolar disorder. In Tohen, M., Bowden, C. L., Nierenberg, A. A., and Geddes, J. R., editors, *Clinical Trial Design Challenges in Mood Disorders*, pages 105 – 136. Academic Press, San Diego.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Sanchez, E. and Valstar, M. (2018). Triple consistency loss for pairing distributions in gan-based face synthesis. *arXiv preprint arXiv:1811.03492*.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Morency, L.-P., et al. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680.

Shannon, B. J. and Paliwal, K. K. (2003). A comparative study of filter bank spacing for speech recognition. In *Microelectronic engineering research conference*, volume 41, pages 310–12.

Shen, Y., Luo, P., Yan, J., Wang, X., and Tang, X. (2018). Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis.

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32.

Sidorov, M. and Minker, W. (2014). Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 81–86. ACM.

Sloan, D. M., Marx, B. P., and Keane, T. M. (2011). Reducing the burden of mental illness in military veterans: Commentary on kazdin and blase (2011). *Perspectives on Psychological Science*, 6(5):503–506.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. (2016). Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*.

Song, S., Jaiswal, S., Shen, L., and Valstar, M. (2020). Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, pages 1–1.

Song, S., Zhang, W., Liu, J., and Mei, T. (2019). Unsupervised person image generation with semantic parsing transformation. In *CVPR*.

Spek, V., Cuijpers, P., Nyklíček, I., Riper, H., Keyzer, J., and Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological medicine*, 37(3):319–328.

Stassen, H., Kuny, S., and Hell, D. (1998). The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology*, 8(4):303–310.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos.

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., and Sunderland, M. (2011). Psychometric comparison of the phq-9 and bdi-ii for measuring response during treatment of depression. *Cognitive behaviour therapy*, 40(2):126–136.

Tlachac, M., Toto, E., Lovering, J., Kayastha, R., Taurich, N., and Rundensteiner, E. (2021). Emu: Early mental health uncovering framework and dataset. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1311–1318. IEEE.

Todt, J., Hanisch, S., and Strufe, T. (2022). Fant\^ omas: Evaluating reversibility of face anonymizations using a general deep learning attacker. *arXiv preprint arXiv:2210.10651*.

Tom, O. et al. (2014). Depression and multimorbidity in psychiatry and primary care. *The Journal of Clinical Psychiatry*, 75(11):4207.

Torous, J., Andersson, G., Bertagnoli, A., Christensen, H., Cuijpers, P., Firth, J., Haim, A., Hsin, H., Hollis, C., Lewis, S., et al. (2019). Towards a consensus around standards for smartphone apps and digital mental health. *World psychiatry*, 18(1):97.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Tran, D., Ranganath, R., and Blei, D. M. (2017). Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 7.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Uddin, M. A., Joolee, J. B., and Lee, Y.-K. (2020). Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. AVEC '16, page 3–10, New York, NY, USA. Association for Computing Machinery.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

Van Segbroeck, M., Tsiartas, A., and Narayanan, S. S. (2013). A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice. In *INTERSPEECH*, pages 704–708.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, M., Yang, G.-Y., Li, R., Liang, R.-Z., Zhang, S.-H., Hall, P. M., and Hu, S.-M. (2019). Example-guided style consistent image synthesis from semantic labeling.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, page 3–14, New York, NY, USA. Association for Computing Machinery.

Wang, W., Pineda, X. A., Xu, D., Fua, P., Ricci, E., and Sebe, N. (2018). Every smile is unique: Landmark-guided diverse smile generation.

Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Bi, J., Kamath, J., Russell, A., Bamis, A., and Wang, B. (2020). Predicting depressive symptoms using smartphone data. *Smart Health*, 15:100093.

WHO (2020). Institute of health metrics and evaluation. global health data exchange ghdx. http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b. [Accessed 10-Mar-2022].

Wichers, M., Simons, C., Kramer, I., Hartmann, J. A., Lothmann, C., Myin-Germeys, I., Van Bemmel, A., Peeters, F., Delespaul, P., and Van Os, J. (2011). Momentary assessment technology as a tool to help patients with depression help themselves. *Acta psychiatrica scandinavica*, 124(4):262–272.

Williams, L. S., Brizendine, E. J., Plue, L., Bakas, T., Tu, W., Hendrie, H., and Kroenke, K. (2005). Performance of the phq-9 as a screening tool for depression after stroke. *stroke*, 36(3):635–638.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., and Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM.

Wu, X., He, R., Sun, Z., and Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896.

Xu, J., Song, S., Kusumam, K., Gunes, H., and Valstar, M. (2021). Two-stage temporal modelling framework for video-based depression recognition using graph representation. *arXiv preprint arXiv:2111.15266*.

Yang, L., Jiang, D., Han, W., and Sahli, H. (2017a). Dcnn and dnn based multimodal depression recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 484–489. IEEE.

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., and Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 89–96. ACM.

Yang, L., Jiang, D., and Sahli, H. (2018a). Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, 12(1):239–253.

Yang, L., Jiang, D., and Sahli, H. (2020). Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access*, 8:24033–24045.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., and Sahli, H. (2017b). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59. ACM.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., and Sahli, H. (2017c). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59.

Yang, L., Li, Y., Chen, H., Jiang, D., Oveneke, M. C., and Sahli, H. (2018b). Bipolar disorder recognition with histogram features of arousal and body gestures. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 15–21.

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., and Jiang, D. (2017d). Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 45–51.

Yang, Y., Fairbairn, C., and Cohn, J. (2012a). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 99.

Yang, Y., Fairbairn, C., and Cohn, J. F. (2012b). Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yu, Z. and Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM.

Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., Tao, J., and Schuller, B. (2019). Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):423–434.

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2018). Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*.

Zhou, X., Wei, Z., Xu, M., Qu, S., and Guo, G. (2020). Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017b). Unpaired image-to-image translation using cycle-consistent adversarial networks.

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017c). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4):578–584.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., and Bai, X. (2019). Progressive pose attention transfer for person image generation. In *CVPR*.

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., and Dalrymple, K. (2015). How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive psychiatry*, 56:29–34.