ABSTRACT

Title of Dissertation: Towards Multimodal and Context-Aware Emotion Perception

Trisha Mittal
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Dinesh Manocha
Department of Computer Science

Human emotion perception is a part of affective computing, a branch of computing that studies and develops systems and devices that can recognize, interpret, process, and simulate human affects. Research in human emotion perception, however, has been mostly restricted to psychology-based literature which explores the theoretical aspects of emotion perception, but does not touch upon its practical applications. For instance, human emotion perception plays a pivotal role in an extensive array of sophisticated intelligent systems, encompassing domains such as behavior prediction, social robotics, medicine, surveillance, and entertainment. In order to deploy emotion perception in these applications, extensive research in psychology has demonstrated that humans not only perceive emotions and behavior through diverse human modalities but also glean insights from situational and contextual cues.

This dissertation not only enhances the capabilities of existing human emotion perception systems but also forges novel connections between emotion perception and multimedia analysis, social media analysis, and multimedia forensics. Specifically, this work introduces two innovative algorithms that revolutionize the construction of human emotion perception models. These algorithms are then applied to detect falsified multimedia, understand human behavior and psychology on social media networks, and extract the intricate array

of emotions evoked by movies.

In the first part of this dissertation, we delve into two unique approaches to advance emotion perception models. The first approach capitalizes on the power of multiple modalities to perceive human emotion. The second approach leverages the contextual information, such as the background scene, diverse modalities of the human subject, and intricate socio-dynamic inter-agent interactions. These elements converge to predict perceived emotions with better accuracy, culminating in the development of context-aware human emotion perception models.

In the second part of this thesis, we forge connections between emotion perception and three prominent domains of artificial intelligence applications. These domains include video manipulations and deepfake detection, multimedia content analysis, and user behavior analysis on social media platforms. Drawing inspiration from emotion perception, we conceptualize enriched solutions that push the conventional boundaries and redefine the possibilities within these domains.

All experiments in this dissertation have been conducted on all state-of-the-art emotion perception datasets, including IEMOCAP, CMU-MOSEI, EMOTIC, SENDv1, MovieGraphs, LIRIS-ACCEDE, DF-TIMIT, DFDC, Intentonomy, MDID, and MET-Meme. In fact, we propose three additional datasets to this list, namely GroupWalk, VIDEOSHAM, and INTENTGRAM. In addition to providing quantitative results to validate our claims, we conduct user evaluations where applicable, serving as a compelling testament to the remarkable outcomes of our experiments.

Towards Multimodal and Context-Aware Emotion Perception

by

Trisha Mittal

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Dr. Dinesh Manocha, Chair/Advisor
Dr. Min Wu, Dean's Representative
Dr. Ramani Duraiswami
Dr. Aniket Bera
Dr. Viswanathan Swaminathan

Acknowledgements

I would like to thank all the people who have contributed in some way to my Ph.D. journey and to the work described in this thesis. First and foremost, I thank my academic advisor, Professor Dinesh Manocha, for accepting me into his research group, the GAMMA Lab. Dinesh made my Ph.D. a rewarding experience by giving me intellectual freedom in my work, supporting my attendance at various conferences, engaging me in new ideas, and for many useful career-related conversations. Additionally, I would like to thank my committee members Dr. Ramani Duraiswami, Dr. Min Wu, Dr. Viswanathan Swaminathan, and Dr. Aniket Bera for their interest in my work and for agreeing to serve as my committee members. I also want to thank Dr. Vanessa Frias-Martinez for her invaluable suggestions during my preliminary examination.

Every result described in this thesis was accomplished with the help and support of fellow labmates and collaborators. Uttaran Bhattacharya, Rohan Chandra, and I worked together on several different projects in affective computing and driver behavior modeling during the initial years of my Ph.D. I greatly benefited from their keen scientific insights, working styles, and knack for solving research problems. I also want to thank Pooja Guhan, Puneet Mathur, and Sanjoy Chowdhury for their help and contributions to some of my research projects. I had the privilege to intern in fantastic research groups at Adobe Research and Apple AI/ML. My managers, Dr. Viswanathan Swaminathan, and Dr. Barry-John Theobald, and my mentors Dr. Ritwik Sinha, and Dr. Zakaria Aldeneh have all contributed to great research discussions, enriched my learning experience, and

helped me make informed decisions about my career options post-graduation.

I am grateful for the funding sources that allowed me to pursue my research: Intel and ARO grants W911NF1810313, W911NF1910069, and W911NF1910315, Adobe Research Fellowship, and the Outstanding Research Assistant Award for AY 2021-22 by The Graduate School, UMD.

I would like to acknowledge the Department of Computer Science at UMD. I benefited greatly from the graduate courses I took during my initial years, and am grateful to Jason Filipou and Prof. Clyde Kruskal for recruiting me as a teaching assistant for their courses. My graduate journey was a smooth-sailing experience because of our graduate coordinator, Tom Hurst, who quite literally has the solution to all department-level problems. Being an international student, the ISSS staff has been truly helpful in navigating confusing visa-related questions.

Finally, I would like to acknowledge how grateful I am for having a group of friends and family who are always ready to listen to my rants about research and graduate school with a smile on their faces.

Table of contents

List of Figures

List of Tables

Chapter 1

Introduction

Prior work defines emotion perception as "the ability to appreciate the emotion information conveyed by the face, body posture, voice, and contextual information, allowing us to adapt and properly respond to environmental situation" (Nahi et al. 2022). Relatedly, affective computing, a term coined by Dr. Rosalind Picard at MIT, entails developing technologies that can understand, interpret and respond to human emotions (Rosalind W. Picard 1995). Moreover, emotion perception is an important aspect of social cognition and is essential for successful communication and interpersonal relationships. Recognizing and understanding perceived human emotions of others comes naturally to humans, as humans are socially-aware. The overall goal of affective computing is to impart these human emotion recognition capabilities to machines so that human-machine interaction may feel more *natural* in the same way as human-human interactions do. This is becoming increasingly important given the rise in the number of ways in which humans interact with machines via phones, computers, and smart appliances. is a challenging enough task for humans, let alone machines, due to the complexity and variability of emotions and their expressions.

Furthermore, emotion perception has already penetrated into multiple aspects of our daily lives. For instance, in the ongoing quest to make a smart assistant, there is a flood of

emotion-equipped avatars and chatbots in the market. Amazon is making Alexa (Robitzski 2019) better understand the humans with whom it interacts by scanning people's voices for signs of emotions. Similarly, ElliQ robot (Marchese 2022) has been developed which addresses the problem of loneliness and lack of social activity among ageing adults. It uses speech, lighting, sound, images and movement to convey emotion and support.

Emotion perception ideas are also finding applications in digital health. For instance, one specific use case is in designing specialized teaching methodologies for Autistic children (Haber, Voss, and Wall 2020; Hebden 2015). Children with autistic spectrum disorders often struggle to recognize and express emotions, making it difficult to behave appropriately in social situations. They are also being deployed to provide mental health therapy for people suffering from Alzheimer's disease; by offering personalized, AI-curated playlists of songs designed to help alleviate symptoms of depression, anxiety, stress, burnout, and other neurocognitive conditions (Angus 2022; Labbe 2022).

Another interesting use case is the use of Emotion AI for driving applications; to monitor the driver's life signs and mental stress in real time in an effort to ensure driving safety (McManus 2017; Times 2022). Emotion recognition can also be used to predict customer behavior, improve customer experience, and increase sales (C. T. Allen, Machleit, and Kleine 1992; Denham et al. 2000). In social and cultural domains, emotion recognition can be used to synthesize and animate digital characters and avatars, enabling more realistic and engaging communication (Roth et al. 2016; Heidicker, Langbehn, and Steinicke 2017). Emotion recognition is also relevant in the fields of social robotics, where robots can recognize and respond to human emotions in real-time, creating more personalized and engaging interactions (Bauer et al. 2009; Narayanan et al. 2020). Emotion recognition has critical implications in the field of surveillance, where it can be used to detect and prevent criminal activity, identify potential threats, and improve security (Clavel et al. 2008a; Clavel et al. 2008b). In the domain of digital teaching and counseling, emotion recognition can be used to personalize and enhance the learning and counseling experience, improving the effectiveness

of therapy and education (Baur et al. 2013; Jamy Li et al. 2016).

Emotion perception is an interdisciplinary research area and involves contributions from all fields such as psychology, computer science and social sciences. In psychology, emotion perception involves the study of how people recognize and interpret emotional experiences. The earliest study in emotion perception dates back to the 1920s, when for the first time researchers showed images of faces of people to gather the perceived emotion labels. This was shortly followed by Basic Emotions Theory proposed by Paul Ekman. The theory suggested that there are six universal emotions that can be recognized across cultures: happiness, sadness, anger, fear, disgust, and surprise (P. Ekman 1971). This is till date a very common emotion label taxonomy used. Subsequently, there have been multiple other theories that have been proposed to explain how humans process emotions (Zajonc 1980; LeDoux 1998). In the 2000s another theory, the theory of constructed emotion was proposed which suggested that emotions are not 6 discrete categories (Barrett, Mesquita, and E. R. Smith 2010).

In computer science the focus is on developing algorithms to recognize emotions from various sources, including facial expressions, speech, and physiological data. For instance, in the early days researchers relied on manual coding of facial expressions to identify emotional states. One of the first systems for automatic emotion recognition was the Facial Action Coding System (FACS), which allowed for automatic facial expression (P. Ekman 1978). By 1980s and 1990s, various other modalities speech, text and physiological signals such as heart rate and skin conductance had been modeled and used for emotion perception (Rosalind W. Picard 1995). In the 1990s, Rosalind Picard introduced the concept of affective computing, which involves the development of technologies that can recognize, interpret, and respond to human emotions. And, as of today, we now have much more advanced deep learning architectures for emotion perception.

Similarly, in social sciences various questions around emotion perception are being explored since a long time. Some of these include understanding how emotions shape

social interactions and relationships (Hochschild 1979); and the role of emotions in economic decision-making (Davidson, K. Scherer, and Goldsmith 2003). More recently research is on understanding how emotions can influence voting, public opinion, and social movements (Jasper 1998) and how culture can influence emotional experiences (Markus and Kitayama 1991).

Similarly, both 'multimodal learning' and 'context-aware' learning are both being used to enrich machine learning algorithms today. Multimodal learning is an approach to machine learning that integrates information from multiple sensory modalities, such as vision, speech, and text. A lot of prior work in focuses on unimodal emotion perception, which considers only a single input modality and these algorithms do perform well. Several studies in psychology (Aviezer, Trope, and Todorov 2012; Soleymani, Pantic, and Pun 2011; Pantic et al. 2005), however, suggest that an ideal system for automatic human emotion recognition should be multimodal, simultaneously taking into account multiple cues from various modalities. Multimodal emotion perception models have been shown to outperform unimodal models (S. Yoon et al. 2019; Gunes and Piccardi 2007; Majumder et al. 2018; A. B. Zadeh et al. 2018; Choi, Song, and C. W. Lee 2018; Sahay et al. 2018) as they make use of the complementary nature of information from different cues, resulting in better inference models that are also more robust to sensor noise. However, one of the longstanding challenges in developing multimodal emotion perception models is finding the most efficient mechanism for combining or "fusing" multiple modalities (Baltrusaitis, Ahuja, and L. Morency 2017).

Context-aware learning take into account the context in which the data was generated or the task is being performed. And, hence context can mean different things in different situations. Furthermore, intersecting studies in affective computing and psychology (Barrett, Mesquita, and E. R. Smith 2010; Ledgerwood 2014; McNulty and Fincham 2012) also suggest that emotional processes cannot be interpreted without context, and that context not only produces emotion but also shapes how emotion is perceived. However, *context* itself is quite subjective; with not much consensus in the literature. For instance, one

organization of contextual features is in three levels, ranging from micro-level (person) to macro-level (cultural) (Greenaway, Kalokerinos, and Williams 2018). Similarly, more literature in emotion recognition suggests several broad categories of contextual features, including person, and situation (Aldao 2013; Barrett, Mesquita, and Gendron 2011; Mesquita and Boiger 2014).

As AI continues to become more ubiquitous in our lives, it is increasingly important to develop AI systems that can understand and respond to human emotions beyond the applications that have already been explored. In this dissertation, we focus and explore our attention on three such AI applications; i) video manipulations and deepfake detection; ii) multimedia content analysis, and iii) user behavior analysis on social media platforms.

The availability of video editing software and artificial intelligence (AI) available to common people has proliferated manipulated video content (Khelifi and Bouridane 2017; Y. He et al. 2021) edited with malicious intent. These edits can be more traditional; copy-move and splicing manipulations, aesthetic edits, and temporal edits; there is also a surge of manipulated videos known as "deepfakes". These are purely Ai generated videos with manipulations focused on a single person. This calls for the need for detection methods for flagging such videos. Prior work in video forensics is limited to very specific attack methods with little benchmarking due to a lack of publicly available video manipulation datasets. On the other hand, because deepfakes are synthesized using deep-learning-based generative methods; the datasets are huge in magnitude. As a result, there are a lot of deepfake detection methods that have been proposed in the last few years, however, these are limited to relying on visual artifacts on faces that appear as a result of the synthesis methods.

Perceiving emotions in images and videos is an integral aspect of affective computing and has applications in various fields such as digital content management (D. Joshi et al. 2014; Y. Wang and B. Li 2015), marketing (McDuff, El Kaliouby, Cohn, et al. 2014; Hussain, Mingda Zhang, X. Zhang, et al. 2017; Ye and Kovashka 2018), education (Downs and Strand 2008; Alqahtani

and Ramzan 2019), and healthcare (Cohn et al. 2009). There is growing interest in dynamically modeling emotions over time; also known as "time series emotion recognition." Several time-series emotion datasets have been proposed to aid in solving this problem. While most datasets focus on single-person emotional narratives in controlled settings, movie databases are also being explored for time-series emotion perception tasks. The theory of "emotional causality" (Coëgnarts and Kravanja 2016; Athanasiadou and Tabakowska 2010; Niemeier and Dirven 1997; Kövecses 2003) has been developed to understand how humans reason and interpret emotions, but it has been relatively unexplored in the context of time-series emotion perception. To reason about emotions invoked in various clips of a movie, it is important to develop a causal understanding of the story.

Finally, the popularity of social media platforms and their usage has increased dramatically in recent years. Consequently, social media has become a rich source for researchers to address a variety of societal problems. While a lot of these problem statements focus on the *content* aspect of social media, given the increased popularity it is also important to understand *the impact social media can have on the users*. There are two characteristics of social media usage that make understanding their impact on users of grave concern. Firstly, users often have little control over the content they consume on social media feeds. And secondly, almost all conversations on social media are $1:n$. To this end, it becomes important to understand the intent behind why users are sharing content on social media; and to make users emotionally aware of the content they consume online. A consequence of a vast number of $1:n$ conversations is immense unidentified *emotion contagion*. While detecting the occurrence of this contagion is a difficult problem to solve, researchers have shown that contagion exists on specific platforms like Twitter, Facebook, and, Weibo, and have also presented hypotheses on factors that tend to cause stronger or weaker contagion on these platforms.

Main Contributions: Below we summarize our contributions presented in this dissertation.

1. **Improved Emotion Perception Models:** We develop two emotion perception models; one uses multiple modalities to infer emotions and the other models' various definitions of context to infer perceived human emotion.

   (a) We present M3ER, a multimodal emotion perception model that combines input modalities in a multiplicative manner and is robust to sensor noise.

   (b) We also present EmotiCon, a context-aware emotion perception model that uses three definitions of context.

   (c) Additionally, we also present GroupWalk dataset; captured in uncontrolled settings with both faces and gaits that have emotion-label annotations. GroupWalk dataset is a collection of 45 videos captured in multiple real-world settings of people walking in dense crowd settings. The videos have about 3544 agents annotated with their emotion labels.

2. **Video Manipulation and Deepfake Detection Methods:** We focus our attention on both; more recent kind of AI-synthesized manipulated videos "deepfakes" and the more traditional video manipulations.

   (a) We present a novel approach that simultaneously exploits the audio (speech) and video (face) modalities and the perceived emotion features extracted from both modalities to detect any falsification or alteration in the input video.

   (b) We also present a new manipulated high-resolution video dataset called VIDEOSHAM. VIDEOSHAM consists of 823 videos that are manipulated using six spatial and temporal attacks manipulating videos at the scene level targeting, not just faces, but also the background context, text, and audio, aesthetic edits, adding/removing entities, and temporal edits. We also present a user evaluation to understand the kind of manipulated videos from VIDEOSHAM that deceive humans and state-of-the-art methods. We also present some ideas

from multimodal and affective models that can be used to flag such manipulations.

3. **Affective Analysis of Multimedia Content:** To better understand the emotions and affect presented in multimedia content, we present Affect2MM, a learning-based method for capturing the dynamics of emotion over time. Affect2MM aligns with the psychological theory of "emotional causality" to better model the emotions evoked by each clip of a movie. To better model this temporal causality in movies for long-range multimedia content like movies, we use attention methods and Granger causality to explicitly model the temporal causality (between clips in movies).

4. **User Behavior Analysis on Social Media Platforms:** We focus on understanding how content shared on social media platforms impacts the users given the increased amount of usage.

   (a) We present INTENT-O-METER, a learning-based model that can predict the intent of users when they make a social media post. In addition to visual (image) and textual (caption) features, the model leverages Theory of Reasoned Action (TRA) factoring in (i) the creator's attitude towards sharing a post, and (ii) the social norm or perception towards the post in determining the creator's intention. We also present INTENTGRAM, an intent prediction dataset curated from public Instagram profiles using Apify[1]. We also present a user evaluation by integrating the intent prediction model with a web application interface (similar to Instagram) to understand users' feedback on the use of such intent labels along with social media posts.

   (b) We also present detailed analysis and insights into what are some factors that could lead to stronger or weaker contagion on the popular social media application, Instagram.

[1] https://apify.com

Overview of the Thesis: We organize the material in the dissertation as follows; In Chapter 2, we discuss our contributions to emotion perception models using multiple modalities and contextual cues. We present our work on the use of affective computing and multimodal learning in detecting video manipulations and deepfake detection in Chapter 3. This is followed by a discussion on temporal affective analysis of multimedia content in Chapter 4. In Chapter 5 we discuss the impact of social media platforms and their usage on users. We identify two key characteristics and present a learning model, a dataset, a user evaluation, and an analysis of this impact on users. We end this with a discussion of limitations in our existing work and some directions for future work in Chapter 6.

Chapter 2

Using Multiple Modalities and Context for Emotion Perception

One of the primary tasks in developing efficient multimodal systems for perceiving emotions is to combine and collate information from the various modalities by which humans express emotion. These modalities include but are not limited to, facial expressions, speech and voice modulations, written text, body postures, gestures, and walking styles. Many researchers have advocated combining more than one modality to infer perceived emotion for various reasons, including:

(a) *Richer information:* Cues from different modalities can augment or complement each other, and hence lead to more sophisticated inference algorithms.

(b) *Robustness to Sensor Noise:* Information on different modalities captured through sensors can often be corrupted due to signal noise, or be missing altogether when the particular modality is not expressed, or cannot be captured due to occlusion, sensor artifacts, etc. We call such modalities *ineffectual*. Ineffectual modalities are

especially prevalent in in-the-wild datasets.

However, combining multiple modalities for emotion recognition presents its own set of challenges. One critical decision is determining which modalities should be integrated and how. Some modalities tend to occur together more frequently than others and are therefore more straightforward to collect and use in conjunction. For example, some of the most widely used benchmark datasets for multimodal emotion recognition, including IEMOCAP (Busso et al. 2008) and CMU-MOSEI (A. B. Zadeh et al. 2018), a feature commonly co-occurring modalities, such as facial expressions, associated speech, and transcribed text. As the volume of data available on social media sites and the internet continues to grow (e.g., YouTube), it has become easier to obtain data for these three modalities, often with the help of automatic caption generation. Many of the other existing multimodal datasets (Ringeval et al. 2013; Dhall, Goecke, Lucey, et al. 2012) are also subsets of these three modalities. Additionally, using information from body posture and gaits has also been shown to be a promising cue to infer emotions (Kleinsmith and Bianchi-Berthouze 2012; Meeren, Heijnsbergen, and Gelder 2005; Bhattacharya et al. 2020). In our work, we try and use all of these modalities.

One additional challenge is the lack of consensus on the optimal method for combining, or "fusing," multiple modalities (Baltrusaitis, Ahuja, and L. Morency 2017). Currently, the two most commonly employed techniques are early fusion (also referred to as "feature-level" fusion) and late fusion (also called "decision-level" fusion). Early fusion combines the input modalities into a single feature vector, which is then used to make a prediction. In contrast, late fusion methods use each of the input modalities to make an individual prediction, which is subsequently combined for the final classification. Previous research on emotion recognition has primarily investigated early fusion (Sikka et al. 2013) and late fusion (Gunes and Piccardi 2007) techniques using additive combinations. Additive combinations assume that every modality is potentially useful and thus should be incorporated into the joint representation. However, this assumption is not ideal for in-the-wild datasets, which are

more susceptible to sensor noise.

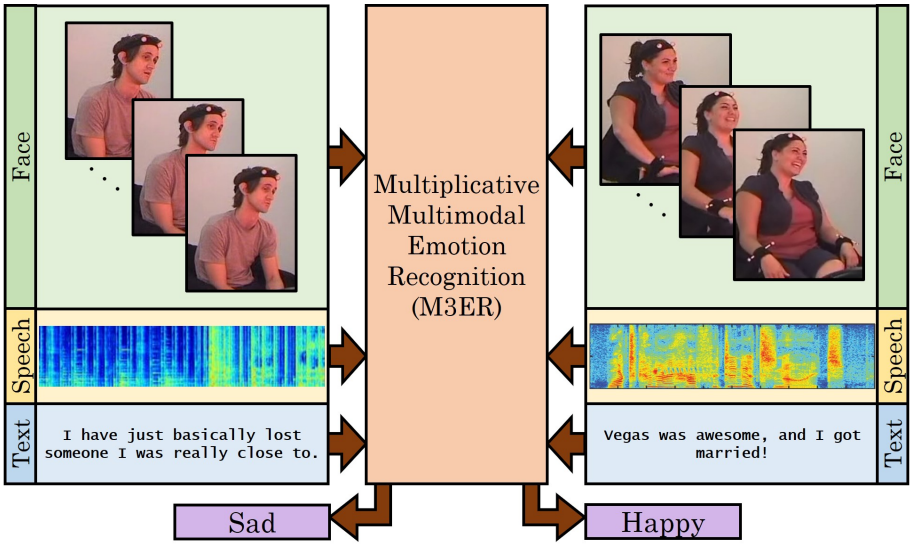
While modalities and cues extracted from a person can provide valuable information regarding perceived emotions, understanding the context in which the emotions are displayed is equally important. Frege's context principle (Resnik 1967) argues against interpreting the meaning of a word in isolation, but rather within the context of a sentence. Similarly, in the field of psychology, context plays a crucial role in emotion recognition. There are different interpretations of the term 'context' among researchers, including:

(a) *Context 1 (Multiple Modalities):* Incorporating cues from different modalities was one of the initial definitions of context (a.k.a multimodal emotion perception). As already discussed, combining modalities provides complementary information, which leads to better inference and also performs better on in-the-wild datasets. An additional goal of this work is to make Emotion Recognition systems work for real-life scenarios. This implies using modalities that do not require sophisticated equipment to be captured and are readily available. Psychology researchers (Aviezer, Trope, and Todorov 2012) have conducted experiments by mixing faces and body features corresponding to different emotions and found that participants guessed the emotions that matched the body features. This is also because of the ease of "mocking" one's facial expressions. Subsequently, researchers (Kleinsmith and Bianchi-Berthouze 2012; Meeren, Heijnsbergen, and Gelder 2005) found the combination of faces and body features to be a reliable measure of inferring human emotion. As a result, it would be useful to combine such face and body features for context-aware emotion recognition.

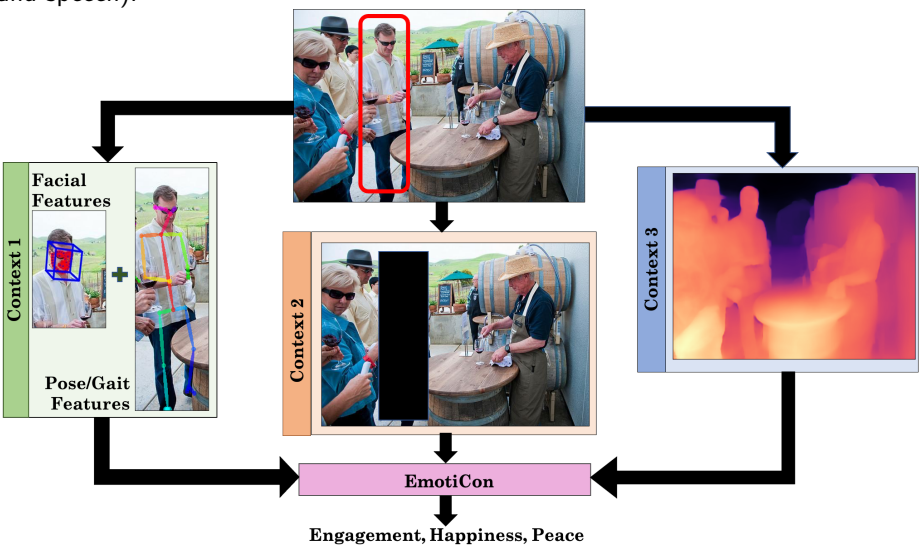
(b) *Context 2 (Background Context):* Semantic understanding of the scene from visual cues in the image helps in getting insights about the agent's surroundings and activity, both of which can affect the perceived emotional state of the agent.

(c) *Context 3 (Socio-Dynamic Inter-Agent Interactions):* Researchers in psychology suggest that the presence or absence of other agents affects the perceived emotional

state of an agent. When other agents share an identity or are known to the agent, they often coordinate their behaviors. This varies when other agents are strangers. Such interactions and proximity to other agents have been less explored for perceived emotion recognition.



(a) M3ER: The proposed multimodal emotion recognition model uses a data-driven multiplicative fusion technique to combine the three input modalities (face, text, and speech).



(b) EmotiCon: The proposed context-aware emotion recognition model incorporates three interpretations of context to perform emotion recognition from videos and images.

Figure 2.1: **Multimodal and Context-Aware Emotion Perception Models:** We present a multimodal emotion perception model, M3ER (Figure 2.1a), and a context-aware emotion perception model, EmotiCon (Figure 2.1b) to improve the performance of emotion perception models.

Towards this end, we propose a multimodal emotion perception model, M3ER (Figure 2.1a, Section 2.2), and a context-aware emotion perception model, EmotiCon (Figure 2.1b, Section 2.3) to improve performance of emotion recognition models. To sum-

marize, we make the following main contributions.

1. We present M3ER (Figure 2.1a, Section 2.2), a multimodal emotion recognition algorithm which uses a data-driven multiplicative fusion technique with deep neural networks. Our input consists of the feature vectors for three modalities — face, speech, and text. To make M3ER robust to noise, we propose a novel preprocessing step where we use Canonical Correlational Analysis (CCA) (Hotelling 1936) to differentiate between an ineffectual and effectual input modality signal. We also present a feature transformation method to generate proxy feature vectors for ineffectual modalities given the true feature vectors for the effective modalities. This enables our network to work even when some modalities are corrupted or missing.
2. We also present EmotiCon (Figure 2.1b, Section 2.3), a context-aware multimodal emotion recognition algorithm. Consistent with Ferge's Context principle, in this work, we try to incorporate three interpretations of context to perform emotion recognition from videos and images. Though extendable to any number of modalities, we release a new dataset GroupWalk for emotion recognition. To the best of our knowledge, there exist very few datasets captured in uncontrolled settings with both faces and gaits that have emotion-label annotations. To enable research in this domain, we make GroupWalk publicly available with emotion annotations. GroupWalk is a collection of 45 videos captured in multiple real-world settings of people walking in dense crowd settings. The videos have about 3544 agents annotated with their emotion labels.

For M3ER, we compare our work with prior methods by testing our performance on two benchmark datasets IEMOCAP (Busso et al. 2008) and CMU-MOSEI (A. B. Zadeh et al. 2018) We report an accuracy of 82.7**%** on the IEMOCAP dataset and 89.0**%** on the CMU-MOSEI dataset, which is a collective 5% accuracy improvement on the absolute over prior methods. We show ablation experiment results on both datasets, where almost 75% of the data has at least one modality corrupted or missing, to demonstrate the

importance of our contributions. As per the annotations in the datasets, we classify IEMOCAP into 4 discrete emotions (*angry, happy, neutral, sad*) and CMU-MOSEI into 6 discrete emotions (*anger, disgust, fear, happy, sad, surprise*).

Similarly for EmotiCon, we compare our work with prior methods by testing our performance on EMOTIC (Kosti, Jose M Alvarez, et al. 2017b), a benchmark dataset for context-aware emotion recognition. We report an improved AP score of 35.48 on EMOTIC, which is an improvement of $7 - 8$ over prior methods (Kosti, Jose M Alvarez, et al. 2019; Jiyoung Lee et al. 2019; Minghui Zhang, Liang, and Ma 2019). We also report AP scores of our approach and prior methods on the new dataset, GroupWalk. We perform ablation experiments on both datasets, to justify the need for the three components of EmotiCon. As per the annotations provided in EMOTIC, we perform a multi-label classification over 26 discrete emotion labels. On GroupWalk too, we perform a multi-label classification over 4 discrete emotions (*anger, happy, neutral, sad*).

The rest of the chapter is structured as follows: We discuss prior work in the domain of multimodal and context-aware emotion perception in Section 2.1. Then in Section 2.2 and Section 2.3, we discuss M3ER, the multimodal emotion perception model, and EmotiCon, the context-aware emotion perception model respectively. We also elaborate on the empirical analysis performed to evaluate the methods. We conclude with a discussion regarding limitations and some future directions in building more robust emotion perception models in Section 2.4.

2.1 Prior Work in Emotion Perception Models

In this section, we give a brief overview of previous research in emotion recognition in psychology (Section 2.1.1), and prior work in unimodal and multimodal emotion recognition (Section 2.1.2-2.1.3). We follow up with a discussion on context-aware emotion recognition (Section 2.1.4) and existing context-aware datasets (Section 2.1.5). We also discuss some of the existing fusion methods used in the emotion recognition lit-

erature (Section 2.1.6) and conclude with a brief discussion on Canonical Correlational Analysis (Section 2.1.7).

2.1.1 Emotion Recognition in Psychology Research

Understanding and interpreting human emotion is a fundamental area of interest in psychology. Early attempts at predicting emotions solely based on facial expressions were questioned for their limited reflection of the complex human sensory system (Russell, Bachorowski, and Fernández-Dols 2003). Furthermore, relying solely on facial expressions can be unreliable due to the ease of producing mocking expressions, especially in the presence of an audience (P. Ekman 1993; Fernández-Dols and Ruiz-Belda 1995). Consequently, it is essential to consider multiple cues to accurately predict human emotions. Research indicates that a multimodal approach that considers various cues, such as body language, vocal tone, and context, is more consistent with the human sensory system (Soleymani, Pantic, and Pun 2011; Aviezer, Trope, and Todorov 2012; Pantic et al. 2005; Meeren, Heijnsbergen, and Gelder 2005). The significance of contextual features in interpreting emotional processes cannot be overstated. Researchers in psychology agree that emotions cannot be interpreted without context and that context plays a crucial role in producing and shaping emotions (Barrett, Mesquita, and E. R. Smith 2010; Ledgerwood 2014; McNulty and Fincham 2012; Aldao 2013; Barrett, Mesquita, and Gendron 2011; Mesquita and Boiger 2014). Therefore, it is essential to consider various contextual features, such as the person, situation, and broader cultural context (Greenaway, Kalokerinos, and Williams 2018; Martinez 2019). Experiments conducted by Martinez et al.(Martinez 2019) demonstrated the necessity of context, as participants could infer affect even when the faces and bodies were masked in silent videos. Greenway et al.(Greenaway, Kalokerinos, and Williams 2018) organize contextual features into three levels, ranging from micro-level features like the individual to macro-level features like culture. Situational features, such as the presence and proximity of others, also play an essential role in eliciting emotions, as shown by research indicating that people express more emotion in the

presence of others and especially in situations where they know each other (Yamamoto and Suzuki 2006; Jakobs, Manstead, and Fischer 2001).

2.1.2 Unimodal Emotion Recognition

Most of the initial efforts in recognizing human emotions have been focused on a single modality. Facial expressions have been the most commonly explored modality in this domain, owing to the availability of facial expression datasets and advancements in computer vision methods (Saragih, Lucey, and Cohn 2009; Akputu, Seng, and Y. L. Lee 2013). However, researchers have also explored other modalities such as speech or voice expressions (Klaus R Scherer, Johnstone, and Klasmeyer 2003), body gestures (Navarretta 2012), and physiological signals like respiratory and heart signals (Knapp, Jonghwa Kim, and André 2010).

2.1.3 Multimodal Emotion Recognition

Initially, multimodal emotion recognition was explored using classifiers such as Support Vector Machines, linear regression, and logistic regression, particularly when the size of the datasets was less than 500 (Sikka et al. 2013; Gunes and Piccardi 2007; Castellano, Kessous, and Caridakis 2008). However, with the development of bigger datasets, researchers started exploring deep learning architectures for multimodal emotion recognition (S. Yoon et al. 2019; Gunes and Piccardi 2007; Majumder et al. 2018; A. B. Zadeh et al. 2018; Choi, Song, and C. W. Lee 2018; Sahay et al. 2018). In all the multimodal approaches, feature extraction is performed on each of the input modalities using either hand-crafted formulations or deep learning architectures. Some of the architectures that have been explored include Bi-Directional Long Short Term Memory (BLSTM) networks (S. Yoon et al. 2019), Deep Belief Networks (DBNs) (Gunes and Piccardi 2007), Convolutional Neural Networks (Choi, Song, and C. W. Lee 2018), hierarchical networks (Majumder et al. 2018), and Relational Tensor Networks (Sahay et al. 2018).

Table 2.1: **Summary of Context-Aware Emotion Recognition Datasets:** We compare GroupWalk dayaset with existing context-rich emotion recognition datasets such as EMOTIC (Kosti, Jose M Alvarez, et al. 2017a), AffectNet (Mollahosseini, Hasani, and Mahoor 2019), CAER and CAER-S (Jiyoung Lee et al. 2019), and AFEW (Dhall, Goecke, Lucey, et al. 2012).

Data type	Dataset	Dataset Size	Agents Annotated	Setting	Emotion Labels	Context
Images	EMOTIC (Kosti, Jose M Alvarez, et al. 2017a)	18,316 images	34,320	Web	26 Categories	Yes
	AffectNet (Mollahosseini, Hasani, and Mahoor 2019)	450,000 images	450,000	Web	8 Categories	No
	CAER-S (Jiyoung Lee et al. 2019)	70,000 images	70,000	TV Shows	7 Categories	Yes
Videos	AFEW (Dhall, Goecke, Lucey, et al. 2012)	1,809 clips	1,809	Movie	7 Categories	No
	CAER (Jiyoung Lee et al. 2019)	13,201 clips	13,201	TV Show	7 Categories	Yes
	IEMOCAP (Busso et al. 2008)	12 hrs	-	TV Show	4 Categories	Yes
	GroupWalk	45 clips(10 mins each)	3544	Real Settings	4 Categories	Yes

2.1.4 Context-Aware Emotion Recognition

Recent advancements in context-aware emotion recognition have been achieved through deep-learning network architectures. Kosti et al. (Kosti, Jose M Alvarez, et al. 2019) and Lee et al. (Jiyoung Lee et al. 2019) propose similar architectures that consist of two-stream networks followed by a fusion network. One stream focuses on a specific modality such as the face (Jiyoung Lee et al. 2019) or body posture (Kosti, Jose M Alvarez, et al. 2019), while the other captures contextual information. Lee et al. (Jiyoung Lee et al. 2019) consider everything except the face as context and mask the face in the image, while (Kosti, Jose M Alvarez, et al. 2019) uses a Region Proposal Network (RPN) to extract contextual elements from the image. These elements form nodes in an affective graph, which is then encoded using a Graph Convolution Network (GCN) to capture the context. Additionally, group emotion recognition has also been explored (Garg 2019; K. Wang et al. 2018), where the goal is to label the emotion of the entire group of individuals in a frame under the assumption that they share some social identity.

2.1.5 Context-Aware Emotion Recognition Datasets

In the past, emotion recognition datasets have often focused on only one modality, such as faces or body features, or have been collected in controlled environments. For instance, the

GENKI database (Whitehill et al. 2009) and the UCDSEE dataset (Tracy, Robins, and Schriber 2009) concentrate primarily on facial expressions collected in lab settings. The Emotion Recognition in the Wild (EmotiW) challenges (Dhall, Goecke, J. Joshi, et al. 2016) contains three databases: the AffectNet database (Mollahosseini, Hasani, and Mahoor 2019) (which comprises facial expressions from the wild), SFEW (a subset of AFEW with only face frames annotated), and HAPPEI database, which aims to estimate group-level emotions. Some recent works have acknowledged the potential of using context for emotion recognition and pointed out the scarcity of such datasets. For instance, the Context-Aware Emotion Recognition (CAER) dataset (Minghui Zhang, Liang, and Ma 2019) is a collection of TV show clips with 7 discrete emotion annotations. The EMOTIC dataset (Kosti, Jose M Alvarez, et al. 2019) is a collection of images from datasets such as MSCOCO (T.-Y. Lin et al. 2014) and ADE20K (B. Zhou et al. 2019), along with images downloaded from web searches. The dataset consists of $23,571$ images, with approximately $34,320$ people annotated for 26 discrete emotion classes. We have summarized and compared these datasets in Table 2.1.

2.1.6 Combination of Multiple Modalities

Previous studies in emotion recognition (Sikka et al. 2013; Gunes and Piccardi 2007; Castellano, Kessous, and Caridakis 2008; S. Yoon et al. 2019; Gunes and Piccardi 2007) have relied on additive combinations using either early or late fusion techniques. However, in the real world, every modality is not equally reliable for every data point due to factors like sensor noise, occlusions, etc. To address this, recent works have explored more sophisticated data-driven (Choi, Song, and C. W. Lee 2018), hierarchical (Majumder et al. 2018), and attention-based (S. Yoon et al. 2019; Choi, Song, and C. W. Lee 2018) fusion techniques. Multiplicative combination methods (Kuan Liu et al. 2018) explicitly model the relative reliability of each modality, where more reliable modalities are assigned a greater weight in the joint prediction. Additionally, the reliability of modalities may vary from sample to sample, making it crucial to learn which modalities are more reliable for each sample. This approach has

been successful in tasks like user profiling and physical process recognition (Kuan Liu et al. 2018).

2.1.7 Canonical Correlational Analysis (CCA)

The main goal of Canonical Correlation Analysis (CCA) (Hotelling 1936) is to project input vectors onto a common space to maximize their component-wise correlation. To improve the performance of CCA, several extensions have been proposed, such as Deep CCA (Andrew et al. 2013), Generalized CCA (Kettenring 1971), and Kernel CCA (Welling 2005). These extensions learn non-linear transformations of the input vectors to maximize their correlation. CCA-based approaches have also been employed in the field of multimodal emotion recognition (Shan, Gong, and McOwan 2007), to obtain highly correlated feature vectors from each input modality prior to their fusion. In our study, we utilize CCA to assess the correlation between input modalities and identify effective and ineffective modalities.

2.2 M3ER: Multimodal Emotion Perception Model

In this section, we go over the proposed multimodal emotion perception model, M3ER. We formulate the problem in Section 2.2.1. We then explain the three components in detail in M3ER in Section 2.2.2. In Section 2.2.3 we discuss the implementation details and end with the experiments conducted to evaluate and analyze the performance of M3ER in Section 2.2.4.

2.2.1 Problem Formulation

We denote the set of speech, text, and facial modality tuples as $\mathcal{M} = \{(s, t, f)\}$, where each tuple corresponds to a data sample, such as a video or an image. For a particular data sample with the modality tuple $m = (s, t, f) \in \mathcal{M}$, the feature vector for each modality is denoted as $f_i, i \in m$. We denote the set of predicted emotions as $\mathcal{E} = \{\text{happy}, \text{sad}, \text{angry}, \text{neutral}\}$. The proxy feature vectors generated for speech, text, and

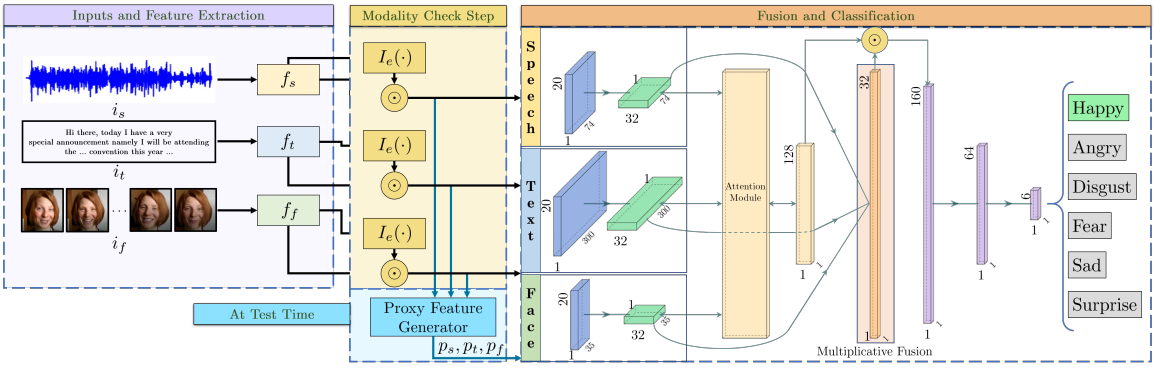


Figure 2.2: **M3ER Model Architecture:** We use three modalities, speech, text, and facial features. We first extract features to obtain f_s, f_t, f_f from the raw inputs, i_s, i_t and i_f (purple box). The feature vectors then are checked if they are effective. We use an indicator function I_e (Equation 2.2) to process the feature vectors (yellow box). These vectors are then passed into the classification and fusion network of M3ER to get a prediction of the emotion (orange box). At the inference time, if we encounter a noisy modality, we regenerate a proxy feature vector (p_s, p_t or p_f) for that particular modality (blue box).

face vectors are represented by p_s, p_t, p_f, respectively. Finally, we define an indicator function, $I_e(f)$ that outputs either a vector of zeros or ones of the same dimension as f, depending on the conditions of the function definition.

Problem 2.2.1. *Given a tuple $m \in \mathcal{M}$ represented as (s, t, f), our goal is to compute*

$$e_m = \arg\max_{e \in \mathcal{E}} \mathbb{P}(emotion = e|m) \tag{2.1}$$

where $\mathbb{P}(a|b)$ denotes the conditional probability of a given b.

We present an overview of our multimodal perceived emotion recognition model in Figure 2.2. During training, we first extract feature vectors (f_s, f_t, f_f) from raw inputs (i_s, i_t, i_f) (purple box in the Figure 2.2). These are then passed through the modality check step (yellow box in the Figure 2.2) to distinguish between effective and ineffectual signals, and discarding the latter if any (See Section 2.2.2.1). The feature vectors as returned by the modality check step go through three deep-layered feed-forward neural network channels (orange box in Figure 2.2). Finally, we add our multiplicative fusion layer to combine the three modalities (Section 2.2.2.3). At test time, the data point

once again goes through the modality check step. If a modality is deemed ineffectual, we regenerate a proxy feature vector (blue box in Figure 2.2) which is passed to the network for the emotion classification (Section 2.2.2.2).

2.2.2 Approach

There are three main components to M3ER: Modality Check Step, Regenerating Proxy Feature Vectors, and Multiplicative Fusion.

2.2.2.1 Modality Check Step

To enable perceived emotion recognition in real-world scenarios, where sensor noise is inevitable, we introduce the Modality Check step which filters ineffectual data. It has been observed in emotion prediction studies (Shan, Gong, and McOwan 2007), that for participants whose emotions were predicted correctly, each of their corresponding modality signals correlated with at least one other modality signal. We directly exploit this notion of correlation to distinguish between features that could be effective for emotion classification (effective features) and features that are noisy (ineffectual features).

More concretely, we use Canonical Correlation Analysis (CCA) to compute the correlation score, ρ, of every pair of input modalities. We compare the correlation against a heuristically chosen threshold, τ, and introduce the following indicator function,

$$I_e(f_i) = \begin{cases} 0 & \rho(f_i, f_j) < \tau, (i,j) \in \mathcal{M}, i \neq j, \\ \mathbb{1} & else. \end{cases} \tag{2.2}$$

For all features, we apply the following operation, $I_e(f) \odot f$, which discards ineffectual features and retains the effective ones. Here, $\odot$ denotes element-wise multiplication.

We show how to compute the correlation score between two modality feature vectors. Given a pair of feature vectors, f_i, f_j, with $i, j \in \mathcal{M}$, we first compute the the projective

transformations, $H^i_{i,j}$ and $H^j_{i,j}$, for both feature vectors, respectively. Also note that these feature vectors f_i, f_j are reduced to the same lower dimensions (100, here). We obtain the projected vector by applying the projective transformation. Thus, in our example above,

$$f_i^{'} = H^i_{i,j} f_i,$$

and,

$$f_j^{'} = H^j_{i,j} f_j,$$

Finally, we can compute the correlation score for the pair $\{f_i, f_j\}$ using the formula:

$$\rho(f_i^{'}, f_j^{'}) = \frac{\mathrm{cov}(f_i^{'}, f_j^{'})}{\sigma_{f_i^{'}} \sigma_{f_j^{'}}}$$

and check them against an empirically chosen threshold (τ). $\forall i \in m$, we check

$$\rho(f_i^{'}, f_j^{'}) < \tau,$$

where $\forall\, (i,j) \in \mathcal{M}, i \neq j$.

For implementation purposes, we keep the $H^j_{i,j}$ for all pairs of modalities precomputed based on the training set. At inference time, we simply compute the projected vectors $f_i^{'}, f_j^{'}$ and $\rho(f_i^{'}, f_j^{'})$.

2.2.2.2 Regenerating Proxy Feature Vectors

When one or more modalities have been deemed ineffectual at test time in the modality check step, we generate proxy feature vectors for the ineffectual modalities using the following equation, $p_i = \mathcal{T} f_i$, where $i \in \mathcal{M}$ and $\mathcal{T}$ is any linear transformation. We illustrate the details below.

Generating exact feature vectors for missing modalities is challenging due to the non-linear relationship between the modalities. However, we empirically show that by relaxing

the non-linear constraint, there exists a linear algorithm that approximates the feature vectors for the missing modalities with high classification accuracy. We call these resulting vectors: proxy feature vectors.

Suppose that during test time, the feature vector for the speech modality is corrupt and identified as ineffectual, while f_f is identified as effective during the Modality Check Step. Our aim is then to regenerate a proxy feature vector, p_s, for the speech modality. More formally, we are given, say, a new, unseen face modality feature vector, f_f, the set of observed face modality vectors, $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, and the set of corresponding observed speech modality vectors, $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. Our goal is to generate a proxy speech vector, p_s, corresponding to f_f.

We begin by preprocessing the inputs to construct bases, $\mathcal{F}_b = \{v_1, v_2, \ldots, v_p\}$ and $\mathcal{S}_b = \{w_1, w_2, \ldots, w_q\}$ from the column spaces of $\mathcal{F}$ and $\mathcal{S}$. Under the relaxed constraint, we assume there exists a linear transformation, $\mathcal{T} : \mathcal{F}_b \to \mathcal{S}_b$. Our algorithm proceeds without assuming knowledge of $\mathcal{T}$:

1. The first step is to find $v_j = \arg\min_j d(v_j, f_f)$, where d is any distance metric. We chose the L_2 norm in our experiment s. We can solve this optimization problem using any distance metric minimization algorithm such as the K-nearest neighbors algorithm.

2. Compute constants $a_i \in \mathbb{R}$ by solving the following linear system, $f_f = \sum_{i=1}^{p} a_i v_i$. Then,

$$p_s = \mathcal{T} f_f = \sum_{i=1}^{p} a_i \mathcal{T} v_i = \sum_{i=1}^{p} a_i w_i.$$

Our algorithm can be extended to generate proxy vectors from effective feature vectors corresponding to multiple modalities. In this case, we would apply the steps above to each of the effective feature vectors and take the mean of both the resulting proxy vectors.

2.2.2.3 Multiplicative Modality Fusion

The key idea in the original work (Kuan Liu et al. 2018) for multiplicative combination is to explicitly suppress the weaker (not so expressive) modalities, which indirectly boost the stronger (expressive) modalities. They define the loss for the i^{th} modality as follows.

$$c^{(y)} = -\sum_{i=1}^{M} \prod_{j \neq i} \left(1 - p_j^{(y)}\right)^{\beta/(M-1)} \log p_i^{(y)} \tag{2.3}$$

where y is the true class label, M is the number of modalities, β is the hyperparameter that down-weights the unreliable modalities and $p_i^{(y)}$ is the prediction for class y given by the network for the i^{th} modality. This indirectly boosts the stronger modalities. In our approach, we reverse this concept and propose a modified loss. We explicitly boost the stronger modalities in the combination network. The difference is subtle but has key significance on the results. In the original formulation, the modified loss was given by Equation 2.3. We empirically show that the modified loss gives better classification accuracies than the originally proposed loss function in Section 2.2.4. The original loss function tries to ignore or tolerate the mistakes of the modalities making wrong predictions by explicitly suppressing them, whereas, in our modified version, we ignore the wrong predictions by simply not addressing them and rather focusing on modalities giving the right prediction. In the original loss, calculating the loss for each modality depends on the probability given by all the other modalities. This has a higher computation cost due to the product term. Furthermore, if either of the input modalities produces an outlier prediction due to noise in the signal, it affects the prediction of all other modalities. Our proposed modified loss is as follows:

$$c^{(y)} = -\sum_{i=1}^{M} \left(p_i^{(y)}\right)^{\beta/(M-1)} \log p_i^{(y)} \tag{2.4}$$

This fusion layer is applied to combine the three input modalities.

M3ER is a modular algorithm that can work on top of existing networks for multimodal classification. Given a network for multiple modalities, we can replace the fusion step and incorporate the modality check and proxy vector regeneration of the M3ER and improve classification accuracies. In the next section, we demonstrate this point by incorporating M3ER in state-of-the-art networks for two datasets, IEMOCAP and CMU-MOSEI.

2.2.3 Implementation Details

In this section, we discuss the datasets (Section 2.2.3.1) that we use to evaluate our method, M3ER, followed with feature extraction (Section 2.2.3.2) and network architecture (Section 2.2.3.3) details. We end with mentioning training and hyperparameters used to obtain the results on the CMU-MOSEI dataset and the IEMOCAP dataset (Section 2.2.3.4).

2.2.3.1 Datasets

The *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) dataset (Busso et al. 2008) consists of text, speech, and face modalities of 10 actors recorded in the form of conversations using a Motion Capture camera. The conversations include both scripted and spontaneous sessions. The labeled annotations consist of four emotions — angry, happy, neutral, and sad. The *CMU Multimodal Opinion Sentiment and Emotion Intensity* (CMU-MOSEI) (A. B. Zadeh et al. 2018) contains $23,453$ annotated video segments from $1,000$ distinct speakers and 250 topics acquired from social media channels. The labels in this dataset comprise six emotions —*angry, disgust, fear, happy, sad,* and *surprise*.

2.2.3.2 Feature Extraction

CMU-MOSEI Dataset: To extract f_t from the CMU-MOSEI dataset, we use the 300-dimensional pre-trained GloVe word embeddings (Pennington, Socher, and Manning 2014). To compute f_s from the CMU-MOSEI dataset, we follow the approach of Zadeh et

al. (A. B. Zadeh et al. 2018) and obtain the 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features, glottal source parameters among others. Lastly, to obtain f_f, we use the combination of face embeddings obtained from state-of-the-art facial recognition models, facial action units, and facial landmarks for CMU-MOSEI.

IEMOCAP Dataset: We use the 300 dimensional pre-trained GloVe word embeddings (Pennington, Socher, and Manning 2014) to extract f_t for the IEMOCAP dataset. To compute f_s for IEMOCAP, we follow Chernykh et al. (Chernykh and Prikhodko 2017a) and use 34 acoustic features. Lastly, to obtain f_f, IEMOCAP captures facial data comprising 189 facial expressions using a Motion Capture camera.

2.2.3.3 Classification Network Architecture Details

CMU-MOSEI Dataset: For training on the CU-MOSEI dataset, we integrate our multiplicative fusion layer into Zadeh et al. (A. Zadeh et al. 2018b) memory fusion network (MFN). Each of the input modalities is first passed through single-hidden-layer LSTMs, each of output dimension 32. The outputs of the LSTMs, along with a 128-dimensional memory variable initialized to all zeros (yellow box in the network Figure 2.2), are then passed into an *attention module* as described by the authors of MFN. The operations inside the attention module are repeated for a fixed number of iterations t, determined by the maximum sequence length among the input modalities ($t = 20$ in our case). The outputs at the end of every iteration in the attention module are used to update the memory variable as well as the inputs to the LSTMs. After the end of t iterations, the outputs of the 3 LSTMs are combined using multiplicative fusion to a 32 dimensional feature vector. This feature vector is concatenated with the final value of the memory variable, and the resultant 160 dimensional feature vector is passed through a 64 dimensional fully connected layer followed by a 6 dimensional fully connected to generate the network outputs.

IEMOCAP Dataset: For M3ER-IEMOCAP, we use Tripathi et al.'s (Samarth Tripathi, Sarthak Tripathi, and Beigi 2018) multiple fully connected layers of dimensions 128, 64 and 4

before a softmax for the speech input as can be seen in Figure 2.3. However, as opposed to multiple LSTM layers, we use a single LSTM layer with 64 hidden units before two dense layers of dimensions 64 and 4 followed by a softmax layer for the text modality. For the facial input, we use three convolutional layers with filter sizes 32, 64, and 128 all with a stride length of 2. These convolutional layers are followed by two fully connected layers (dimensions 64 and 4) succeeded by a softmax layer. All three softmax layers then go into the multiplicative layer (shown in orange in Figure 2.3).

2.2.3.4 Training Details

CMU-MOSEI Dataset: For training with M3ER on the CMU-MOSEI dataset, we split the CMU-MOSEI dataset into training (70%), validation (10%), and testing (20%) sets. We use a batch size of 256 and train it for 500 epochs. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.01. All our results were generated on an NVIDIA GeForce GTX 1080 Ti GPU.

IEMOCAP Dataset: For training M3ER-IEMOCAP, we split the IEMOCAP dataset into training (85%) and testing (15%) sets. We use a batch size of 128 and train it for 100 epochs. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.001. All our results were generated on an NVIDIA GeForce GTX 1080 Ti GPU.

2.2.4 Experiments and Results

In this section, we list the state-of-the-art algorithms with which we compare M3ER using standard classification evaluation metrics in Section 2.2.4.1. We also discuss the performance of M3ER on these state-of-the-art datasets (Section 2.2.4.2). We perform exhaustive ablation experiments to motivate the benefits of our contributions and discuss these ablation experiments in Section 2.2.4.3.

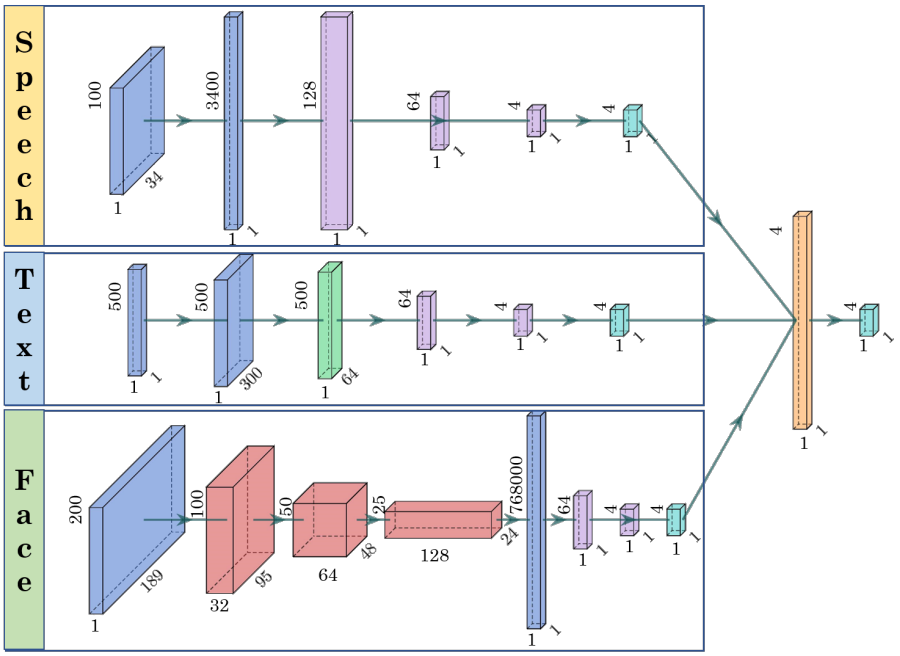


Figure 2.3: **M3ER Model Architecture (IEMOCAP Dataset):** We use three modalities, speech, text, and facial features for evaluating M3ER on IEMOCAP dataset too. We use a variety of layers, fully-connected (purple), softmax (cyan), LSTM (green), and convolutional layers (red). The multiplicative layer is shown in orange.

Table 2.2: **Ablation Experiments on M3ER:** We remove one component of M3ER at a time, and report the F1 and MA scores on the IEMOCAP and the CMU-MOSEI datasets, to showcase the effect of each of these components. Modifying the loss function leads to an increase of 6-7% in both F1 and MA. Adding the modality check step on datasets with ineffectual modalities leads to an increase of 2-5% in F1 and 4-5% in MA, and adding the proxy feature regeneration step on the same datasets leads to a further increase of 2-7% in F1 and 5-7% in MA.

(a) Ablation Experiments performed on IEMOCAP Dataset.

Ineffectual modalities?	Experiments	Angry		Happy		Neutral		Sad		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Kuan Liu et al. 2018)	0.794	80.6%	0.750	76.9%	0.695	68.0%	0.762	80.8%	0.751	76.6%
	M3ER	**0.862**	**86.8%**	**0.862**	**81.6%**	**0.745**	**74.4%**	**0.828**	**88.1%**	**0.824**	**82.7%**
Yes	M3ER– Modality Check Step – Proxy Feature Vector	0.704	71.6%	0.712	70.4%	0.673	64.7%	0.736	79.8%	0.706	71.6%
	M3ER– Proxy Feature Vector	0.742	75.7%	0.745	73.7%	0.697	66.9%	0.778	84.0%	0.741	75.1%
	M3ER	**0.799**	**82.2%**	**0.743**	**76.7%**	**0.727**	**67.5%**	**0.775**	**86.3%**	**0.761**	**78.2%**

(b) Ablation Experiments performed on CMU-MOSEI Dataset.

Ineffectual modalities?	Experiments	Angry		Disgust		Fear		Happy		Sad		Surprise		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Kuan Liu et al. 2018)	0.889	79.9%	0.945	89.6%	0.963	93.1%	0.587	55.8%	0.926	85.3%	0.949	90.0%	0.878	82.3%
	M3ER	**0.919**	**86.3%**	**0.927**	**92.1%**	**0.904**	**88.9%**	**0.836**	**82.1%**	**0.899**	**89.8%**	**0.952**	**95.0%**	**0.902**	**89.0%**
Yes	M3ER– Modality Check Step – Proxy Feature Vector	0.788	73.3%	0.794	80.0%	0.843	85.0%	0.546	55.7%	0.832	79.5%	0.795	80.1%	0.764	75.6%
	M3ER– Proxy Feature Vector	0.785	77.8%	0.799	83.2%	0.734	77.5%	0.740	77.1%	0.840	86.0%	0.781	83.5%	0.783	80.9%
	M3ER	**0.816**	**81.3%**	**0.844**	**86.8%**	**0.918**	**89.4%**	**0.780**	**75.7%**	**0.873**	**86.1%**	**0.932**	**91.3%**	**0.856**	**85.0%**

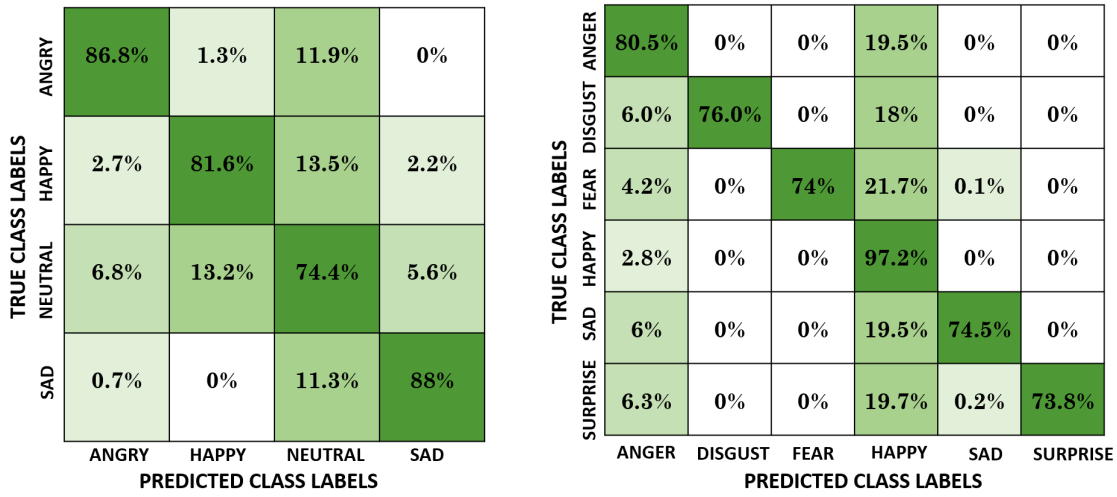


Figure 2.4: **Analysis of M3ER's performance using a Confusion Matrix:** For each emotion class, we show the percentage of inputs belonging to that class that was correctly classified by M3ER (dark green cells) and the percentage of inputs that were misclassified into other classes (pale green and white cells) for both the datasets. *Left:* Confusion matrix for classification on IEMOCAP dataset. *Right:* Confusion matrix for classification on CMU-MOSEI dataset.

2.2.4.1 Evaluation Metrics and Methods

We use two standard metrics, F1 scores and mean classification accuracies (MAs), to evaluate all the methods. However, some prior methods have not reported MA, while others have not reported F1 scores. We, therefore, leave out the corresponding numbers in our evaluation as well and compare the methods with only the available numbers. For the IEMOCAP dataset, we compare our accuracies with the following state-of-the-art methods.

(a) Yoon et al. (S. Yoon et al. 2019) use only two modalities of the IEMOCAP dataset, text and speech, using an attention mechanism that learns to align the relevant text with the audio signal instead of explicitly combining outputs from the two modalities separately. The framework uses two Bi-linear LSTM networks.

(b) Kim et al. (Y. Kim, H. Lee, and Provost 2013) focus on feature selection parts and hence use DBNs which they claim are better equipped at learning high-order non-linear relationships. They empirically show that non-linear relationships help in emotion

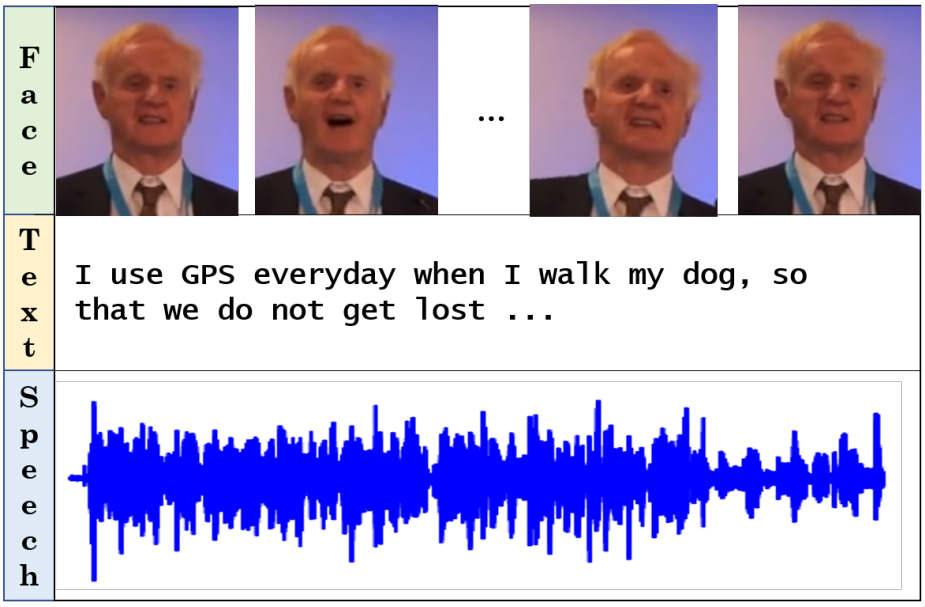


Figure 2.5: **Example Misclassification by M3ER:** This is the text and face input of a 'happy' data point from CMU-MOSEI dataset that our model, M3ER misclassifies as 'angry'. IN this example, the man is giving a funny speech with animated and exaggerated facial looks which appear informative but lead us to a wrong class label.

recognition.

(c) Majumdar et al. (Majumder et al. 2018) recognize the need for a more explainable and intuitive method for fusing different modalities. They propose a hierarchical fusion that learns bimodal and trimodal correlations for data fusion using deep neural networks.

For the CMU-MOSEI dataset, we compare our F1 scores with the following state-of-the-art methods.

(a) Zadeh et al. (A. B. Zadeh et al. 2018) propose a Dynamic Fusion Graph (DFG) for fusing the modalities. The DFG can model n-modal interactions with an efficient number of parameters. It can also dynamically alter its structure and choose a fusion graph based on the importance of each n-modal dynamics. They claim that this is a more interpretable fusion as opposed to naive late fusion techniques.

(b) Choi et al. (Choi, Song, and C. W. Lee 2018) use the text and speech modality of the CMU-MOSEI dataset. They extract feature vectors for text and speech spectro-

Table 2.3: **Comparing M3ER's Performance with SOTA methods:** We compare the F1 scores and the mean classification accuracies (MA) of M3ER on the two datasets, IEMOCAP and CMU-MOSEI, with three prior state-of-the-art methods. Numbers not reported by prior methods are marked with '-'. We observe around 5-10% increase in MA and 1-23% increase in F1 score.

Dataset	Method	F1	MA
IEMOCAP	Kim et al. (Y. Kim, H. Lee, and Provost 2013)	-	72.8%
	Majumdar et al. (Majumder et al. 2018)	-	76.5%
	Yoon et al. (S. Yoon et al. 2019)	-	77.6%
	M3ER	**0.824**	**82.7%**
CMU-MOSEI	Sahay et al. (Sahay et al. 2018)	0.668	-
	Zadeh et al. (A. B. Zadeh et al. 2018)	0.763	-
	Choi et al. (Choi, Song, and C. W. Lee 2018)	0.895	-
	M3ER	**0.902**	**89.0%**

grams using Convolutional Neural Networks (CNNs) architectures. They then use a trainable attention mechanism to leaner the non-linear dependence between the two modalities.

(c) Sahay et al. (Sahay et al. 2018) propose a tensor fusion network that explicitly models n-modal inter-modal interactions using an n-fold Cartesian product from modality embeddings.

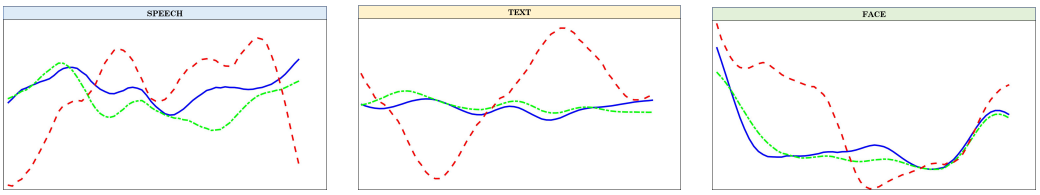


Figure 2.6: **M3ER's Step 2 Visualization (Regenerating Ineffectual Feature Vector):** We show the quality of the regenerated proxy feature vectors for each of the three modalities. For the three graphs, we demonstrate the original feature vector (blue), the ineffectual version of the modality because of added white Gaussian noise (red), and the regenerated feature vector (green). The mean L_2 norm distance between the original and the regenerated vector for the speech, text, and face modality are all around 0.01% of the L_2 norm of the respective data.

2.2.4.2 Experimental Analysis

We discuss and analyze the results of various experiments below.

Figure 2.7: **M3ER's Qualitative Results on CMU-MOSEI Dataset:** We qualitatively show data points correctly classified by M3ER from all the 6 class labels of CMU-MOSEI. The labels as classified by M3ER in row order from top left, are *Anger, Disgust, Fear, Happy, Sad, Surprise*.

1. **Comparison with state-of-the-art:** Evaluation of F1 scores and MAs of all the methods is summarized in Table 2.3. We observe an improvement of 1-23% in F1 scores and 5-10% in MAs when using our method.
2. **Confusion Matrix:** We also show the confusion matrix (Figure 2.4) to analyze the per-class performance of M3ER on IEMOCAP and CMU-MOSEI. We observe that more than 73% of the samples per class were correctly classified by M3ER. We see no confusion (0%) between some emotion labels in the two confusion matrices, for instance, 'sad' and 'happy' in IEMOCAP and 'fear' and 'surprise' in CMU-MOSEI. Interestingly, we see a small set of data points getting confused between 'happy' and 'angry' labels for both datasets. We reason that this is because, in both situations, people often tend to exaggerate their cues.
3. **Qualitative Results:** Additionally, we show one sample per class from the CMU-MOSEI dataset that was correctly classified by M3ER in Figure 2.7.
4. **Failure Case:** We also qualitatively show a data point in Figure 2.5 where M3ER fails to classify correctly. We observe that exaggerations of facial expressions and speech have led to a 'happy' sample being classified by our model as 'angry', a pattern also observed from the confusion matrices.

2.2.4.3 Ablation Experiments

We explain and list down the ablation experiments conducted to validate M3ER below-

1. **Original vs M3ER Multiplicative Fusion Loss:** We first compare the original multiplicative fusion loss (Kuan Liu et al. 2018) (Equation 2.3) with our modified loss (Equation 2.4 on both IEMOCAP and CMU-MOSEI. As shown in Table 2.2, using our modified loss results in an improvement of 6-7% in both F1 score and MA. Next, to motivate the necessity of checking the quality of signals from all the modalities and implementing corrective measures in the case of ineffectual features, we corrupt the datasets by adding white Gaussian noise with a signal-to-noise ratio of 0.01 to at least one modality in approximately 75% of the samples in the datasets. We then compare the performance of the various ablated versions of M3ER as summarized in Table 2.2 and detailed below.
2. **M3ER – Modality Check Step – Proxy Feature Vector:** This version simply applies the multiplicative fusion with the modified loss on the datasets. We show that this results in a drop of 4-12% in the overall F1 score and 9-12% in the overall MA from the non-ablated version of M3ER.
3. **M3ER – Proxy Feature Vector:** In this version, we perform the modality check step to filter out the ineffectual modality signals. This results in an improvement of 2-5% in the overall F1 score and 4-5% in the overall MA from the previous version. However, we do not replace the filtered-out modalities with generated proxy features, thus having fewer modalities to work with. This results in a drop of 2-7% in the overall F1 score and 5-7% in the overall MA from the non-ablated version of M3ER. Finally, with all the components of M3ER in place, we achieve an overall F1 score of 0.761 on IEMOCAP and 0.856 on CMU-MOSEI, and an overall MA of 78.2% on IEMOCAP and 85.0% on CMU-MOSEI. Additionally, we also show in Figure 2.6 that the mean L_2 norm distance between the proxy feature vectors

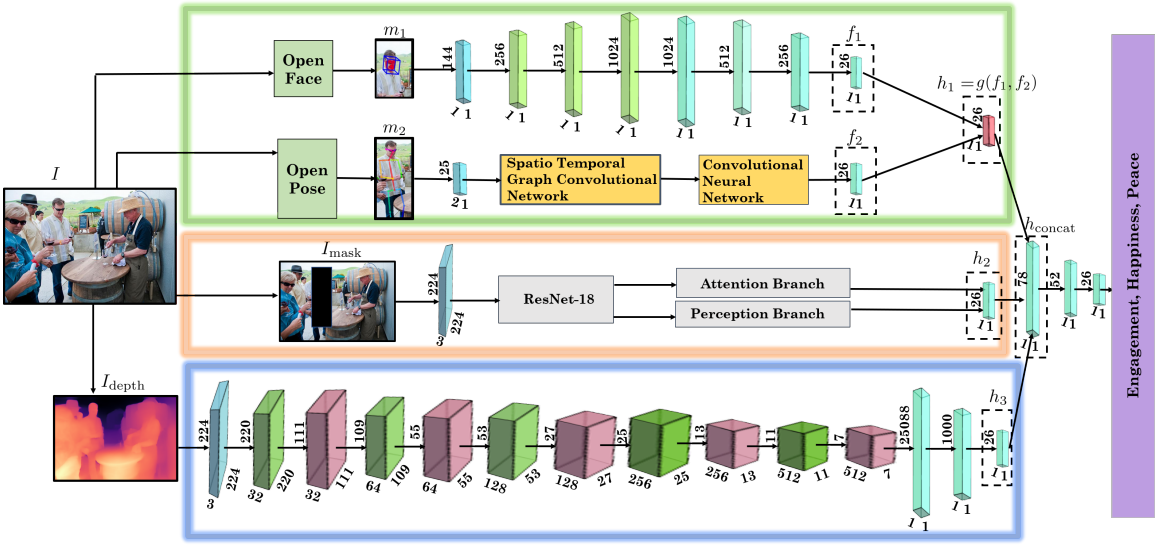


Figure 2.8: **EmotiCon Model Architecture:** We use three interpretations of context. We first extract features for the two modalities to obtain f_1 and f_2 and inputs I_{mask} and I_{depth} from the raw input image, I. These are then passed through the respective neural networks to obtain h_1, h_2 and h_3. To obtain h_1, we use a multiplicative fusion layer (red color) to fuse inputs from both modalities, faces, and gaits. h_1, h_2 and h_3 are then concatenated to obtain h_{concat}.

regenerated by M3ER in and the ground truth data is around 0.01% of the L_2 norm of the respective data.

2.3 EmotiCon: Context-Aware Emotion Perception

In this section, we go over the proposed context-aware emotion perception model, EmotiCon. We formulate the problem in Section 2.3.1. We then explain the three interpretations of context used in detail in EmotiCon in Section 2.3.2. In Section 2.3.3 we discuss the implementation details and end with the experiments conducted to evaluate and analyze the performance of EmotiCon in Section 2.3.4.

2.3.1 Problem Formulation

Our input consists of an RGB image, I. We process I to generate the input data for each network corresponding to the three contexts. The network for Context 1 consists of n streams corresponding to n distinct modalities denoted as $m_1, m_2, \ldots, m_n$. Each distinct

layer outputs a feature vector, f_i. The n feature vectors $f_1, f_2, \ldots, f_n$ are combined via multiplicative fusion (Mittal, Bhattacharya, et al. 2020b) to obtain a feature encoding, $h_1 = g(f_1, f_2, \ldots, f_n)$, where $g(\cdot)$ corresponds to the multiplicative fusion function. Similarly, h_2, and h_3 are computed through the networks corresponding to the second and third Contexts. h_1, h_2, and h_3 are concatenated to perform multi-label emotion classification.

We are presented with the following problem formulation:

Problem 2.3.1. *Given a data sample I as input, e.g. a video frame or an image, along with a list of m context features, $h_1, h_2, \ldots, h_m$, our goal is to predict the most likely emotions $e_1, e_2, ..., e_k$ present in I.*

We present an overview of our context-aware multimodal emotion recognition model, EmotiCon, in Figure 2.8.

2.3.2 Approach

EmotiCon models three definitions of context. We go over them in detail in this section.

2.3.2.1 Context 1 (Multiple Modalities)

In real life, people appear in a multi-sensory context that includes a voice, a body, and a face; these aspects are also perceived as a whole. Combining more than one modality to infer emotion is beneficial because cues from different modalities can complement each other. They also seem to perform better on in-the-wild datasets (Mittal, Bhattacharya, et al. 2020b) than other unimodal approaches. Our approach is extendable to any number of modalities available. To validate this claim, other than EMOTIC and GroupWalk, which have two modalities, faces, and gaits, we also show results on the IEMOCAP dataset which is face, text, and speech as three modalities. From the input image I, we obtain $m_1, m_2, \ldots, m_n$ using processing steps as explained in Section 2.3.3. These inputs are then passed through their respective neural network architectures to obtain $f_1, f_2, \ldots, f_n$.

To make our algorithm robust to sensor noise and averse to noisy signals, we combine these features multiplicatively to obtain h_1. As shown in previous research (Kuan Liu et al. 2018; Mittal, Bhattacharya, et al. 2020b), multiplicative fusion learns to emphasize reliable modalities and to rely less on other modalities. To train this, we use the modified loss function proposed previously (Mittal, Bhattacharya, et al. 2020b) defined as:

$$L_{\text{multiplicative}} = -\sum_{i=1}^{n} (p_i^e)^{\frac{\beta}{n-1}} \log p_i^e \tag{2.5}$$

where n is the total number of modalities being considered, and p_i^e is the prediction for emotion class, e, given by the network for the i^{th} modality.

2.3.2.2 Context 2 (Situational/Background Context)

Our goal is to identify semantic context from images and videos to perform perceived emotion recognition. Semantic context includes the understanding of objects –excluding the primary agent– present in the scene, their spatial extents, keywords, and the activity being performed. For instance, in Figure 2.1b, the input image consists of a group of people gathered around with drinks on a bright sunny day. The "bright sunny day", "drink glasses", "hats" and "green meadows" constitute semantic components and may affect the judgment of one's perceived emotion.

Motivated by multiple approaches in the computer vision literature (Zheng et al. 2019; Fukui et al. 2019) surrounding semantic scene understanding, we use an attention mechanism to train a model to focus on different aspects of an image while *masking* the primary agent, to extract the semantic components of the scene. The mask, $I_{\text{mask}} \in \mathbb{R}^{224\times224}$, for an input image I is given as

$$I_{\text{mask}} = \begin{cases} I(i,j) & \text{if } I(i,j) \notin \text{bbox}_{\text{agent}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

where $\mathrm{bbox}_{\mathrm{agent}}$ denotes the bounding box of the agent in the scene.

2.3.2.3 Context 3 (Inter-Agent Interactions/Socio-Dynamic Context)

When an agent is surrounded by other agents, their perceived emotions change. When other agents share an identity or are known to the agent, they often coordinate their behaviors. This varies when other agents are strangers. Such interactions and proximity can help us infer the emotion of agents better.

Prior experimental research has used walking speed, distance, and proximity features to model socio-dynamic interactions between agents to interpret their personality traits.

Some of these algorithms, like the social force model (Helbing and Molnar 1995), are based on the assumption that pedestrians are subject to attractive or repulsive forces that drive their dynamics. Non-linear models like RVO (Yeh et al. 2008) aim to model collision avoidance among individuals while walking to their individual goals. But, both of these methods do not capture cohesiveness in a group.

We propose an approach to model these socio-dynamic interactions by computing proximity features using depth maps. The depth map, $I_{\mathrm{depth}} \in \mathbb{R}^{224\times224}$, corresponding to input image, I, is represented through a 2D matrix where,

$$I_{\mathrm{depth}}(i,j) = d(I(i,j),c) \tag{2.7}$$

$d(I(i,j),c)$ represents the distance of the pixel at the i^{th} row and j^{th} column from the camera center, c. We pass I_{depth} as input depth maps through a CNN and obtain h_3.

In addition to depth map-based representation, we also use Graph Convolutional Networks (GCNs) to model the proximity-based socio-dynamic interactions between agents. GCNs have been used to model similar interactions in traffic networks (Yan, Y. Xiong, and D. Lin 2018) and activity recognition (S. Guo et al. 2019). The input to a GCN network consists of the spatial coordinates of all agents, denoted by $X \in \mathbb{R}^{n\times2}$, where n represents the number of agents in the image, as well as the unweighted adjacency matrix, $A \in \mathbb{R}^{n\times n}$,

of the agents, which is defined as follows,

$$A(i,j) = \begin{cases} e^{-d(v_i,v_j)} & \text{if } d(v_i,v_j) < \mu, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

The function $f = e^{-d(v_i,v_j)}$ (Belkin and Niyogi 2003) denotes the interactions between any two agents.

2.3.3 Implementation Details

In this section, we discuss the datasets (Section 2.3.3.1) that we use to evaluate our method, EmotiCon, followed with feature extraction and data processing (Section 2.3.3.2) and network architecture (Section 2.3.3.3) details. We end with mentioning training and hyperparameters used to obtain the results on the EMOTIC dataset and the GroupWalk dataset (Section 2.3.3.4).

2.3.3.1 Datasets

We use two datasets for experimentation.

1. **EMOTIC Dataset:** The EMOTIC dataset contains 23,571 images of 34,320 annotated people in unconstrained environments. The annotations consist of the apparent emotional states of the people in the images. Each person is annotated for 26 discrete categories, with multiple labels assigned to each image.
2. **GroupWalk Dataset:** GroupWalk consists of 45 videos that were captured using stationary cameras in 8 real-world setting including a hospital entrance, an institutional building, a bus stop, a train station, a marketplace, a tourist attraction, a shopping place and more. The annotators annotated agents with clearly visible faces and gait across all videos. 10 annotators annotated a total of 3544 agents. The annotations consist of the following emotion labels– Angry, Happy, Neutral,

and Sad. Efforts to build on this dataset are still ongoing. To prepare train and test splits for the dataset, we randomly selected 36 videos for the training and 9 videos for testing.

While perceived emotions are essential, other affects such as dominance and friendliness are important for carrying out joint and/or group tasks. Thus, we additionally label each agent for dominance and friendliness. More details about the annotation process, labelers, and label processing are presented below.

(i) Annotation Procedure: Annotators were allowed to view the videos as many times as they wanted and had to categorize the emotion they perceived looking at the agent into 7 categories - "Somewhat Happy", "Extremely Happy", "Somewhat Sad", Extremely Sad", "Somewhat Angry", "Extremely Angry", "Neutral". In addition to perceived emotions, the annotators were also asked to annotate the agents in terms of dominance (5 categories- "Somewhat Submissive", "Extremely Submissive", "Somewhat Dominant", "Extremely Dominant", "Neutral") and friendliness (5 categories- "Somewhat Friendly", "Extremely Friendly", "Somewhat Unfriendly", "Extremely Unfriendly", "Neutral"). Attempts to build the dataset are still ongoing. For the sake of completeness, we show the friendliness label distribution and dominance label distribution for every annotator in Figure 2.11 and Figure 2.12 respectively.

(ii) Labels Processing: 4 major labels that have been considered are Angry, Happy, Neutral, and Sad. One can observe that the annotations are either "Extreme" or "Somewhat" variants of these major labels (except Neutral). Target labels were now generated for each agent. Each of them is of the size 1 x 4 with the 4 columns representing the 4 emotions being considered and are initially all 0. For a particular agent id, if the annotation by an annotator was an "Extreme" variant of Happy, Sad, or Angry, 2 was added to the number in the column representing the corresponding major label. Otherwise, for all the other cases,

1 was added to the number in the column representing the corresponding major label. Once we have gone through the entire dataset, we normalize the target label vector so that vector is a combination of only 1s and 0s.

(iii) Analysis: We show the emotion label distribution for every annotator in Figure 2.9. To understand the trend of annotator agreement and disagreement across the 10 annotators, we gather agents labeled similarly in majority (more than 50% of annotators annotated the agent with the same labels) and then study the classes they were confused most with. We show this pictorially for two classes Happy and Sad in Figure 2.10. For instance, we see that Happy and Sad labels are often confused with the label 'Neutral'. In addition, we also show the label distributions for every annotator for Friendliness as well as Dominance in Figure 2.11 and Figure 2.12 respectively.

2.3.3.2 Data Processing

We discuss the data processing for all three context definitions.

(a) **Context1:** We use OpenFace (Baltrušaitis, Robinson, and L.-P. Morency 2016) to extract a 144-dimensional face modality vector, $m_1 \in \mathbb{R}^{144}$ obtained through multiple facial landmarks. We compute the 2D gait modality vectors, $m_2 \in \mathbb{R}^{25\times 2}$ using OpenPose (Z. Cao et al. 2017) to extract 25-coordinates from the input image I. For each coordinate, we record the x and y pixel values.

(b) **Context2:** We use RobustTP (Chandra et al. 2019), which is a pedestrian tracking method to compute the bounding boxes for all agents in the scene. These bounding boxes are used to compute I_{mask} according to Equation 2.6.

(c) **Context3:** We use Megadepth (Z. Li and Snavely 2018) to extract the depth maps from the input image I. The depth map, I_{depth}, is computed using Equation 2.7.

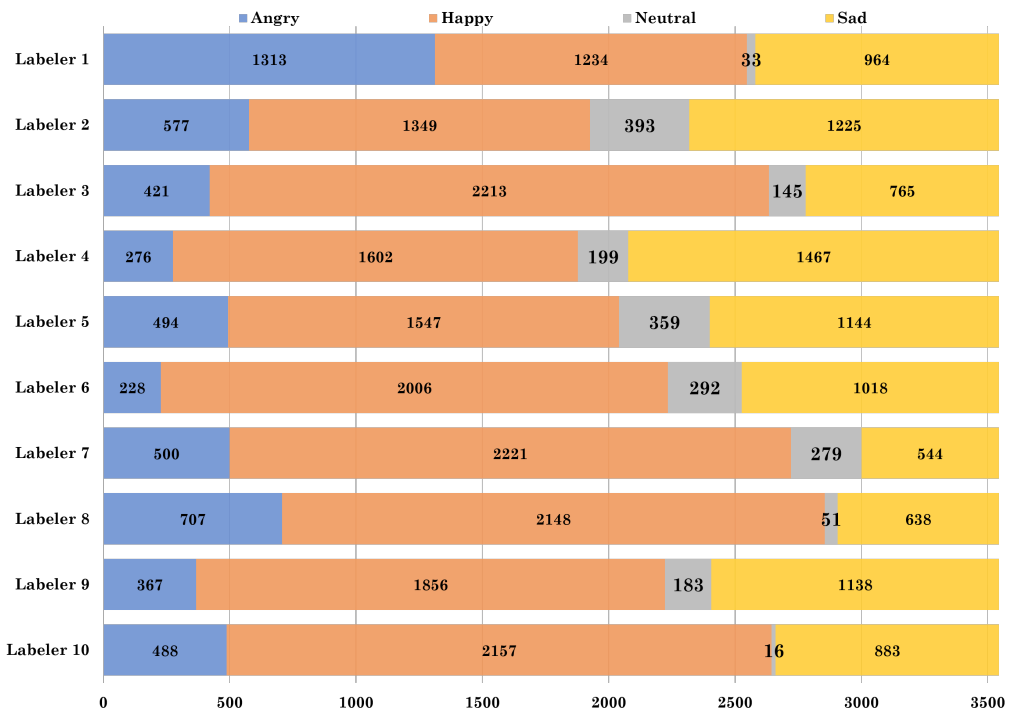


Figure 2.9: **Annotator Annotations of GroupWalk Dataset:** We depict the emotion class labels for GroupWalk by 10 annotators. A total of 3544 agents were annotated from 45 videos.

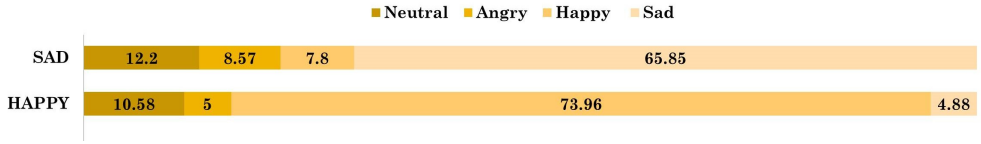


Figure 2.10: **Annotator Agreement/Disagreement in Labeling of GroupWalk Dataset:** For two emotion classes (Happy and Sad), we depict the trend of annotator disagreement.

2.3.3.3 Network Architecture Details

We discuss the network architecture for all three context definitions.

(a) **Context1:** Given a face vector, m_1, we use three 1D convolutions (depicted in light green color in Figure 2.8) with batch normalization and ReLU non-linearity. This is followed by a max pool operation and three fully-connected layers (cyan color in Figure 2.8) with batch normalization and ReLU. For m_2, we use the ST-GCN architecture proposed by (Bhattacharya et al. 2020), which is currently the state-of-the-art network for emotion classification using gaits. Their method was originally designed to deal with 2D pose information for 16 body joints. We modify their

setup for 2D pose inputs for 25 joints. We show the different layers and hyper-parameters used in Figure 2.8. The two networks give us f_1 and f_2, which are then multiplicatively fused (depicted in red color in Figure 2.8) to generate h_1.

(b) **Context2:** For learning the semantic context of the input image I, we use the Attention Branch Network (ABN) (Fukui et al. 2019) on the masked image I_{mask}. ABN contains an attention branch that focuses on attention maps to recognize and localize important regions in an image. It outputs these potentially important locations in the form of h_2.

(c) **Context3:** We perform two experiments using both a depth map and a GCN. For a depth-based network, we compute the depth map, I_{depth}, and pass it through a CNN. The CNN is composed of 5 alternating 2D convolutional layers (depicted in dark green color in Figure 2.8) and max pooling layers (magenta color in Figure 2.8). This is followed by two fully connected layers of dimensions 1000 and 26 (cyan color in Figure 2.8).

For the graph-based network, we use two graph convolutional layers followed by two linear layers of dimension 100 and 26.

(d) **Fusing Context Interpretations:** To fuse the feature vectors from the three context interpretations, we use an early fusion technique. We concatenate the feature vectors before making any individual emotion inferences.

$$h_{\text{concat}} = [h_1, h_2, h_3]$$

We use two fully connected layers of dimensions 52 and 26, followed by a softmax layer. This output is used for computing the loss and the error, and then back-propagating the error back to the network.

(e) **Loss Function:** Our classification problem is a multi-label classification problem where we assign one or more than one emotion label to an input image or video.

To train this network, we use the multi-label soft margin loss function and denote it by $L_{\text{classification}}$. The loss function optimizes a multi-label one-versus-all loss based on max-entropy between the input x and output y.

So, we combine the two loss functions, $L_{\text{multiplicative}}$ (from Eq. 2.5) and $L_{\text{classification}}$ to train EmotiCon.

$$L_{\text{total}} = \lambda_1 L_{\text{multiplicative}} + \lambda_2 L_{\text{classification}} \tag{2.9}$$

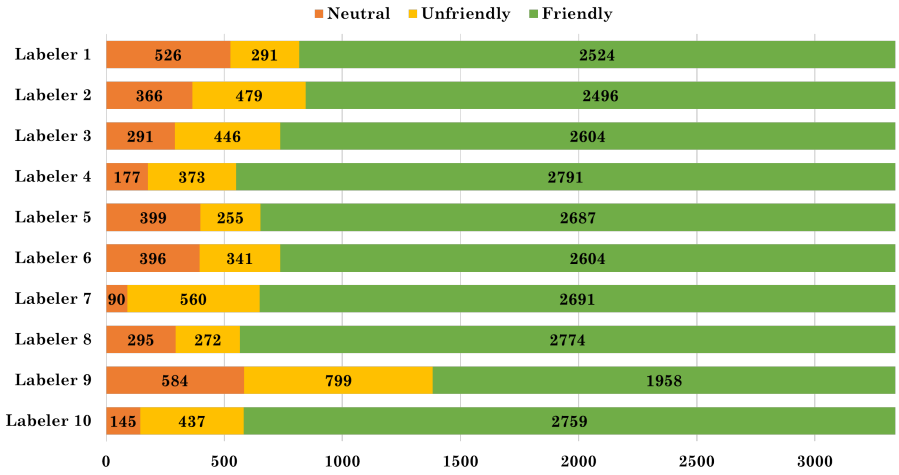


Figure 2.11: **Friendliness Labeler Annotations of GroupWalk Dataset:** We depict the friendliness labels for GroupWalk by 10 labelers. A total of 3341 agents were annotated from 45 videos.

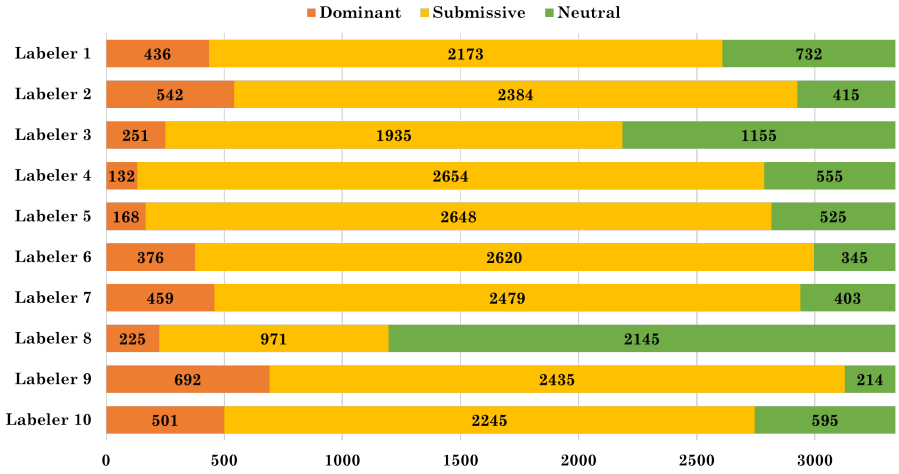


Figure 2.12: **Dominance Labeler Annotationsof GroupWalk Dataset:** We depict the dominance labels for GroupWalk by 10 labelers. A total of 3341 agents were annotated from 45 videos.

2.3.3.4 Training Details

For training EmotiCon on the EMOTIC dataset, we use the standard train, validation, and test split ratios provided in the dataset. For GroupWalk, we split the dataset into training (85%) and testing (15%) sets. In GroupWalk each sample point is an agent ID; hence the input is all the frames for the agent in the video. To extend EmotiCon on videos, we perform a forward pass for all the frames and take an average of the prediction vector across all the frames and then compute the AP scores and use this for loss calculation and backpropagating the loss. We use a batch size of 32 for EMOTIC and a batch size of 1 for GroupWalk. We train EmotiCon for 75 epochs. We use the Adam optimizer with a learning rate of 0.0001. All our results were generated on NVIDIA GeForce GTX 1080 Ti GPU. All the code was implemented using PyTorch (Paszke et al. 2017).

2.3.4 Experiments and Results

In this section, we list the state-of-the-art algorithms against which we compare EmotiCon's performance (Section 2.3.4.1). We discuss and analyze some qualitative and quantitative experiments and results in Section 2.3.4.2. In the end, we perform exhaustive ablation experiments to motivate the benefits of our contributions and discuss these ablation experiments in Section 2.3.4.3.

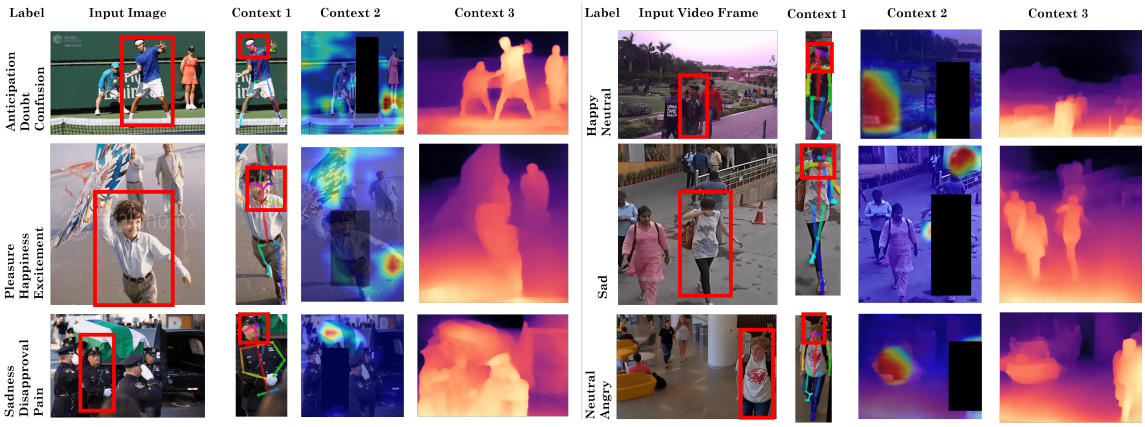


Figure 2.13: **EmotiCon's Qualitative Results on the EMOTIC and GroupWalk Dataset:** We show the classification results on three examples, each from the EMOTIC dataset (left) and GroupWalk Dataset (right), respectively. In the top row example (left) and middle row example (right), the depth map clearly marks the tennis player about to swing to convey anticipation, and the woman coming from the hospital to convey sadness, respectively. In the bottom row (left) and bottom row (middle) examples, the semantic context of the coffin and the child's kite is clearly identified to convey sadness and pleasure, respectively.

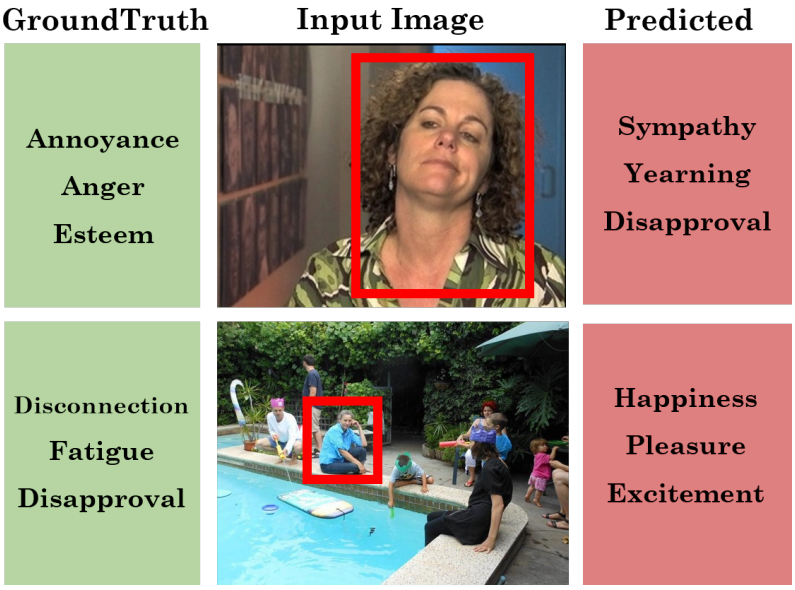


Figure 2.14: **Example Misclassification by EmotiCon:** We show two examples where M3ER incorrectly classifies the labels. In the first example, EmotiCon is confused about the prediction due to lack of any context. In the second example, there are a lot of contexts available, which also becomes confusing.

Table 2.4: **Comparing EmotiCon's Performance with SOTA methods:** We report the AP scores on the EMOTIC and the GroupWalk datasets. EmotiCon outperforms all the three methods for most of the classes and also overall.

(a) AP Scores for EMOTIC Dataset.

Labels	Kosti et al. (Kosti, Jose M Alvarez, et al. 2017a)	Zhang et al. (Minghui Zhang, Liang, and Ma 2019)	Lee et al. (Jiyoung Lee et al. 2019)	**EmotiCon** GCN-Based	**EmotiCon** Depth-Based
Affection	27.85	46.89	19.9	36.78	**45.23**
Anger	09.49	10.87	11.5	14.92	**15.46**
Annoyance	14.06	11.23	16.4	18.45	**21.92**
Anticipation	58.64	62.64	53.05	68.12	**72.12**
Aversion	07.48	5.93	16.2	16.48	**17.81**
Confidence	78.35	72.49	32.34	59.23	**68.65**
Disapproval	14.97	11.28	16.04	**21.21**	19.82
Disconnection	21.32	26.91	22.80	25.17	**43.12**
Disquietment	16.89	16.94	17.19	16.41	**18.73**
Doubt/Confusion	29.63	18.68	28.98	33.15	**35.12**
Embarrassment	03.18	1.94	15.68	11.25	**14.37**
Engagement	87.53	88.56	46.58	90.45	**91.12**
Esteem	17.73	13.33	19.26	22.23	**23.62**
Excitement	77.16	71.89	35.26	82.21	**83.26**
Fatigue	09.70	13.26	13.04	**19.15**	16.23
Fear	14.14	4.21	10.41	11.32	**23.65**
Happiness	58.26	73.26	49.36	68.21	**74.71**
Pain	08.94	6.52	10.36	12.54	**13.21**
Peace	21.56	32.85	16.72	**35.14**	34.27
Pleasure	45.46	57.46	19.47	61.34	**65.53**
Sadness	19.66	25.42	11.45	**26.15**	23.41
Sensitivity	09.28	5.99	**10.34**	9.21	8.32
Suffering	18.84	23.39	11.68	22.81	**26.39**
Surprise	**18.81**	9.02	10.92	14.21	17.37
Sympathy	14.71	17.53	17.125	24.63	**34.28**
Yearning	08.34	10.55	9.79	12.23	**14.29**
mAP	27.38	28.42	20.84	32.03	**35.48**

(b) AP Scores for GroupWalk Dataset.

Labels	Kosti et al (Kosti, Jose M Alvarez, et al. 2017a)	Zhang et al .(Minghui Zhang, Liang, and Ma 2019)	Lee et al. (Jiyoung Lee et al. 2019)	**EmotiCon** GCN-Based	**EmotiCon** Depth-Based
Anger	58.46	-	42.31	65.13	**69.42**
Happy	69.12	-	56.79	72.46	**73.18**
Neutral	42.27	-	39.24	44.51	**48.51**
Sad	63.83	-	54.33	68.25	**72.24**
mAP	58.42	-	48.21	62.58	**65.83**

2.3.4.1 Evaluation Metrics and Methods

We use the standard metric Average Precision (AP) to evaluate all our methods. For both EMOTIC and GroupWalk datasets, we compare our methods with the following state-of-the-art methods.

(a) Kosti et al. (Kosti, Jose M Alvarez, et al. 2019) propose a two-stream network followed by a fusion network. The first stream encodes context and then feeds the entire image as input to the CNN. The second stream is a CNN for extracting body features.

Table 2.5: **Ablation Experiments on EmotiCon:** Keeping the Context interpretation 1 throughout, we remove the other two Context interpretations one by one and compare the AP scores for emotion classification on both the datasets.

(a) Ablation Experiments performed on the EMOTIC Dataset.

Labels	Context Interpretations			
	Only 1	Only 1 and 2	Only 1 and 3	**1, 2 and 3**
Affection	29.87	41.83	30.15	45.23
Anger	08.52	11.41	8.36	15.46
Annoyance	09.65	17.37	12.91	21.92
Anticipation	46.23	67.59	60.53	72.12
Aversion	06.27	11.71	09.46	17.81
Confidence	51.92	65.27	59.63	68.85
Disapproval	11.81	17.35	15.41	19.82
Disconnection	31.74	41.46	32.56	43.12
Disquietment	07.57	12.69	12.24	18.73
Doubt/Confusion	21.62	31.28	29.51	35.12
Embarrassment	08.43	10.51	12.25	14.37
Engagement	78.68	84.62	81.51	91.12
Esteem	18.32	18.79	09.42	23.62
Excitement	73.19	80.54	76.14	83.26
Fatigue	06.34	11.95	14.15	16.23
Fear	14.29	21.36	22.29	23.65
Happiness	52.52	69.51	71.51	74.71
Pain	05.75	09.56	11.10	13.21
Peace	13.53	30.72	30.15	34.27
Pleasure	58.26	61.89	59.81	65.53
Sadness	19.94	19.74	22.27	23.41
Sensitivity	03.16	04.11	8.15	8.32
Suffering	15.38	20.92	12.83	26.39
Surprise	05.29	16.45	16.26	17.37
Sympathy	22.38	30.68	22.17	34.28
Yearning	04.94	10.53	9.82	14.29
mAP	24.06	31.53	29.63	35.48

(b) Ablation Experiments performed on the GroupWalk Dataset.

Labels	Context Interpretations			
	Only 1	Only 1 and 2	Only 1 and 3	**1, 2 and 3**
Anger	58.51	63.83	66.15	69.42
Happy	61.24	64.16	68.87	73.18
Neutral	40.36	41.57	44.15	48.51
Sad	62.17	67.22	70.35	72.24
mAP	55.57	59.20	62.38	65.83

The fusion network combines features of the two CNNs and estimates the discrete emotion categories.

(b) Zhang et al. (Minghui Zhang, Liang, and Ma 2019) build an affective graph with nodes as the context elements extracted from the image. To detect the context elements, they use a Region Proposal Network (RPN). This graph is fed into a Graph Convolutional Network (GCN). Another parallel branch in the network encodes the body features using a CNN. The outputs from both branches are concatenated to infer an emotion label.

(c) Lee et al. (Jiyoung Lee et al. 2019) present network architecture, CAER-Net consisting of two subnetworks, a two-stream encoding network, and an adaptive fusion network. The two-stream encoding network consists of a face stream and a context stream where facial expression and context (background) are encoded. An adaptive fusion network is used to fuse the two streams.

We use the publicly available implementation for Kosti et al. (Kosti, Jose M Alvarez, et al. 2019) and train the entire model on GroupWalk. Both Zhang et al. (Minghui Zhang, Liang, and Ma 2019) and Lee et al. (Jiyoung Lee et al. 2019) do not have publicly available implementations. We reproduce the method by Lee et al. (Jiyoung Lee et al. 2019) to the best of our understanding. For Zhang et al. (Minghui Zhang, Liang, and Ma 2019), while we report their performance on the EMOTIC dataset, with limited implementation details, it was difficult to build their model to test their performance on GroupWalk.

2.3.4.2 Analysis and Discussion

We summarize our experiments and results below-

1. **Comparison with state-of-the-art:** We summarize the evaluation of the APs for all the methods on the EMOTIC and GroupWalk datasets in Table 2.4. For EmotiCon, we report the AP scores for both GCN-based and Depth Map-based

implementations of Context 3. On both the EMOTIC and GroupWalk datasets, EmotiCon outperforms the state-of-the-art.

2. **Generalize to more Modalities:** A major factor for the success of EmotiCon is its ability to combine different modalities effectively via multiplicative fusion. Our approach learns to assign higher weights to more expressive modalities while suppressing weaker ones. For example, in instances where the face may not be visible, EmotiCon infers the emotion from context (See Figure 2.13, middle row(right)). This is in contrast to Lee et al. (Jiyoung Lee et al. 2019), which relies on the availability of face data. Consequently, they perform poorly on both the EMOTIC and GroupWalk datasets, as both datasets contain many examples where the face is not visible clearly.
3. **GCN versus Depth Maps:** GCN-based methods do not perform as well as depth-based but are a close second. This may be due to the fact that on average most images of the EMOTIC dataset contain 5 agents. GCN-based methods in the literature have been trained on datasets with a lot more agents in each image or video. Moreover, with a depth-based approach, EmotiCon learns a 3D aspect of the scene in general and is not limited to inter-agent interactions.
4. **Failure Cases:** We show two examples from EMOTIC dataset in Figure 2.14 where EmotiCon fails to classify correctly. We also show the ground truth and predicted emotion labels. In the first image, EmotiCon is unable to gather any context information. On the other hand, in the second image, there is a lot of context information like the many visual elements in the image and multiple agents. This leads to an incorrect inference of the perceived emotion.
5. **Qualitative Results:** We show qualitative results for three examples, each from both the datasets, respectively, in Figure 2.13. The first column is the input image marking the primary agents, the second column shows the corresponding extracted face and gait, the third column shows the attention maps learned by the model,

and lastly, in the fourth column, we show the depth map extracted from the input image.

The heatmaps in the attention maps indicate what the network has learned. In the bottom row (left) and bottom row (middle) examples, the semantic context of the coffin and the child's kite is clearly identified to convey sadness and pleasure, respectively. The depth maps corresponding to the input images capture the idea of proximity and inter-agent interactions. In the top row example (left) and middle row example (right), the depth map clearly marks the tennis player about to swing to convey anticipation, and the woman coming from the hospital to convey sadness, respectively.

2.3.4.3 Ablation Results

To motivate the importance of Context 2 and Context 3, we run EmotiCon on both EMOTIC and GroupWalk dataset removing the networks corresponding to both contexts, followed by removing either of them one by one. The results of the ablation experiments have been summarized in Table 2.5. We choose to retain Context 1 in all these runs because it is only Context 1 that is capturing information from the agent itself.

We observe from the qualitative results in Figure 2.13 that Context 2 seems more expressive in the images of EMOTIC dataset, while Context 3 is more representative in GroupWalk. This is supported by the results reported in Table 2.5, columns 2 and 3. To understand why this happens, we analyze the two datasets closely. EMOTIC dataset was collected for the task of emotion recognition with context. it is a dataset of pictures collected from multiple datasets and scraped from the Internet. As a result, most of these images have a rich background context. Moreover, we also found that more than half the images of EMOTIC contain at most 3 people. These are the reasons we believe that interpretation 2 helps more in EMOTIC than interpretation 3. In the GroupWalk Dataset, the opposite is true. The number of people per frame is much higher. This density gets

captured best in interpretation 3 helping the network to make a better inference.

2.4 Conclusion, Limitations and Future Work

We presented M3ER, a multimodal emotion recognition model that uses a multiplicative fusion layer and EmotiCon, a context-aware emotion recognition model. M3ER is robust to the sensor because of a modality check step that distinguishes between good and bad signals to regenerate a proxy feature vector for bad signals. We use multiplicative fusion to decide on a per-sample basis which modality should be relied on more for making a prediction. Currently, we have applied our results to databases with three input modalities, namely face, speech, and text. EmotiCon borrows and incorporates three context interpretations from psychology. We use multiple modalities (faces and gaits), situational context, and also socio-dynamic inter-agent context information. We make an effort to use easily available modalities that can be easily captured or extracted using commodity hardware (e.g., cameras). To foster more research on emotion recognition with naturalistic modalities, we also release a new dataset called GroupWalk dataset. The dataset consists of 45 videos that were captured using stationary cameras in 8 real-world setting including a hospital entrance, an institutional building, a bus stop, a train station, a marketplace, a tourist attraction, a shopping place and more.

Our models have limitations and often confuse certain class labels. Further, we currently perform binary classification per class; however, human perception is rather subjective in nature and would resemble a probability distribution over these discrete emotions. Thus, it would be useful to consider multi-class classification in the future. Further, we currently perform multi-class classification over discrete emotion labels for both the models; it would be useful to move towards the continuous model of emotions (Valence, Arousal, and Dominance).

Using multiplicative fusion was one approach towards optimally fusing various modalities; however using more more elaborate fusion techniques can help improve the accuracies

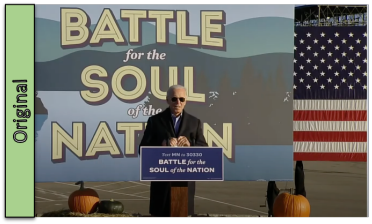
further. It would be further useful to extend M3ER for more than three modalities and use more than three definitions of context for EmotiCon.

Chapter 3

Video Manipulation and Deepfake Detection Using Affective Cues

The proliferation of accessible video editing software and artificial intelligence (AI) tools has led to an increase in manipulated video content (Khelifi and Bouridane 2017; Y. He et al. 2021). While digital manipulation is commonplace in the creative process, in some cases video manipulation has malicious intent. Social media often amplifies such false information through the circulation of manipulated videos (Anderson 2018; Figueira and L. Oliveira 2017). A recent survey by Pew Research Center showed that exposure to such false information is of widespread concern (Silver 2020). Therefore, there has been a significant increase in cases of misinformation, fraud, and cybercrimes in the last decade. Such video manipulations pose a great threat to politics and can manipulate elections (Watts, Rothschild, and Mobius 2021; J. Allen et al. 2020), alter political narratives, weaken the public's trust in a country's leadership, and an increasing hatred among various social groups. Another common occurrence is corporate fraud and scams where people use altered audio

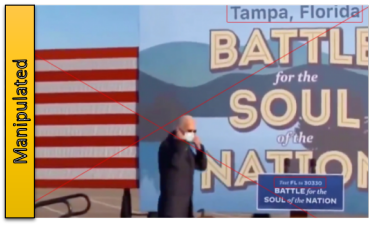
(a1) The original photo, from Getty Images shows an armed man parked in front of a car.



(b1) This is an original clip of a presidential candidate addressing public in the US state, Minnesota.

(c1) An original image shows three missiles being launched by Iran's government.

(a2) The photo above was altered by digitally placing the armed man in front of a peaceful protest, insinuating violence.



(b2) The clip above is altered by changing the location and the signs on the podium to a different US state, Florida.

(c2) In an altered image released on Iran's Revolutionary Guards website, claimed that 4 missiles were launched simultaneously.

Figure 3.1: **Instances of Real-world Video Manipulations: (a)** (Brunner 2020), **(b)** (News 2020), and **(c)** (NPR 2008) are all examples of videos on social media spatially manipulated with the intent to mislead the public. Similar such instances also occur for temporally manipulated videos; but it was not feasible to depict them here.

to impersonate other people to extort cash and other resources. Lastly, many video manipulations often result in numerous cybercrimes (Harris 2018; Spivak 2018; Botha and Pieterse 2020). To further illustrate our motivations in this work, we depict such instances of video manipulations in Figure 3.1.

Recent advances in computer vision and deep learning techniques have enabled the creation of sophisticated and compelling forged versions of social media images and videos (also known as *"deepfakes"*). Some of the common deepfake generation tools and libraries include, FaceSwap [1], FakeApp [2], DeepFaceLab [3], DFaker [4], and, FaceSwap-GAN [5]. Due to the surge in AI-synthesized deepfake content, multiple attempts have been

[1] FaceSwap: https://github.com/deepfakes/faceswap
[2] FakeApp: https://www.malavida.com/en/soft/fakeapp/gref
[3] DeepFaceLab: https://github.com/iperov/DeepFaceLab
[4] DFaker: https://github.com/dfaker/df
[5] FaceSwap-GAN: https://github.com/shaoanlu/faceswap-GAN.

Table 3.1: **Summary of Characteristics of Video Manipulation Datasets:** We compare VIDEOSHAM with state-of-the-art video manipulation datasets.

Faces	Datasets	Release Date	# Videos		Source		Attacks	Human Density	Context	Modality	
			Real	Fake	Original	Manipulated				Visual	Audio
Only	UADFV (X. Yang, Y. Li, and Lyu 2019)	Nov-18	49	49	YouTube	Deep Learning	3	1	✗	✓	✗
	DF-TIMIT (Korshunov and Marcel 2018a)	Dec-18	640	320	VidTIMIT (Sanderson 2002)	Deep Learning	3, 4	1	✗	✓	✓
	FaceForensics++ (Rossler et al. 2019)	Jan-19	1000	4000	YouTube	Deep Learning	3, 4	1	✗	✓	✗
	DFD†	Sep-19	0	3000	YouTube	Deep Learning	3	1	✗	✓	✗
	CelebDF (Y. Li, X. Yang, et al. 2019)	Nov-19	5907	5639	YouTube	Deep Learning	3	1	✗	✓	✗
	DFDC (Dolhansky et al. 2019)	Oct-21	23654	104,500	Actors	Unknown	3	1	✗	✓	✗
	DeeperForensics 1.0 (L. Jiang et al. 2020)	Jan-21	50,000	10,000	Actors	Deep Learning	3	1	✗	✓	✗
	WildDeepFake (Zi et al. 2020)	Jan-21	3,805	3,509	Internet	Internet	3, 4, 5	1	✗	✓	✗
	KoDF (Kwon et al. 2021)	Aug-21	62,166	175,776	Actors	Deep Learning	3, 4, 5	1	✗	✓	✓
	FakeAVCeleb (Khalid, Tariq, and Woo 2021)	Sep-21	490+	20,000+	VoxCeleb2 (Chung, Nagrani, and Zisserman 2018)	Deep Learning	3, 4	1	✗	✓	✓
	ForgeryNet (Y. He et al. 2021)	July-21	91,630	121,617	Multiple	Deep Learning	3, 4	1	✗	✓	✓
	SR-DF (J. Wang et al. 2021)	Apr-21	1,000	4,000	YouTube	Deep Learning	3, 4	1	✗	✓	✓
	Khelifi et al. (Khelifi and Bouridane 2017)	Jan-19	200	200	Multiple	User Generated	6, 7	1	✗	✓	✗
Beyond	MTVFD (Al-Sanjary, Ahmed, and Sulong 2016)	2016	30	30	YouTube	User Generated	1, 2	≤ 1	✓	✗	✗
	Liao et al (Liao and T.-Q. Huang 2013)	2013	10	8	Multiple	User Generated	1	≤ 1	✓	✓	✗
	Su et al (Su, T. Huang, and J. Yang 2015)	2015	7	7	SONY DSCP10	User Generated	1	≤ 1	✓	✓	✗
	Ours	Nov-21	413	413	Online Videos	User Generated		upto 40	✓	✓	✓

† Google AI blog.

made to release benchmark datasets (Korshunov and Marcel 2018a; Rossler et al. 2019; AI n.d.; Dolhansky et al. 2019) and algorithms (P. Zhou, X. Han, et al. 2017; Afchar et al. 2018; X. Yang, Y. Li, and Lyu 2019; Y. Li and Lyu 2018; Matern, Riess, and Stamminger 2019; Rossler et al. 2019; H. H. Nguyen, Fang, et al. 2019; H. H. Nguyen, Yamagishi, and Echizen 2019; Sabir et al. 2019; Güera and Delp 2018; Verdoliva and Bestagini 2019) for deepfake detection. We summarize some of these deepfake datasets in Table 3.1. DeepFake detection methods classify an input video or image as *"real"* or *"fake"*.

Prior methods exploit only a single modality, i.e., only the facial cues from these videos either by employing temporal features or by exploring the visual artifacts within frames. Other than these modalities, multimodal approaches have also exploited the contextual information in video data to detect fakes (Papadopoulou et al. 2017). There are many other applications of video processing that use and combine multiple modalities for audio-visual speech recognition (Gurban et al. 2008), emotion recognition (Mittal, Bhattacharya, et al. 2020b; A. B. Zadeh et al. 2018), and language and vision tasks (Hodosh, Young, and Hockenmaier 2013;

Bigham et al. 2010). These applications show that combining multiple modalities can provide complementary information and lead to stronger inferences. Even for detecting deepfake content, we can extract many such modalities like facial cues, speech cues, background context, hand gestures, and body posture and orientation from a video. When combined, multiple cues or modalities can be used to detect whether a given video is real or fake.

But facial manipulations represent *only a fraction* of all manipulated content circulated on social media. For example, modifications also include changing the background context (Figure 3.1(b)), text and audio (Figure 3.1(c)) in media, aesthetic edits, adding/removing entities (Figure 3.1(a)), and temporal edits (Figure 3.1(d)). These manipulations can be performed in a matter of clicks due to the availability of state-of-the-art video editing tools like Adobe AfterEffects™, Adobe PremierePro™, Filmora, GIMP, and many others. To our knowledge, no benchmark video dataset exists that extends beyond deepfake-only facial manipulations to include the vast range of manipulations described above.

Main Contributions: Towards this problem of manipulated videos, our contributions are two-fold. In our first work, we present an audio-visual deepfake detection method and in the second work, we expand the scope and look at more generic video manipulations. More formally, the following are our main contributions in this domain-

1. We present a novel approach that simultaneously exploits the audio (speech) and video (face) modalities and the perceived emotion features extracted from both modalities to detect any falsification or alteration in the input video. To model these multimodal features and the perceived emotions, our learning method uses a Siamese network-based architecture. At training time, we pass a real video along with its deepfake through our network and obtain modality and perceived emotion embedding vectors for the face and speech of the subject. We use these embedding vectors to compute the triplet loss function to minimize the similarity between the modalities from the fake video and maximize the similarity between modalities for

the real video.

2. We release a new manipulated high-resolution video dataset called VIDEOSHAM (Figure 3.2). VIDEOSHAM offers several benefits over existing manipulated video datasets. Firstly, the videos in VIDEOSHAM are manipulated using six spatial and temporal attacks (See Table 3.2) manipulating videos at the scene level targeting, not just faces, but also the background context, text, and audio, aesthetic edits,

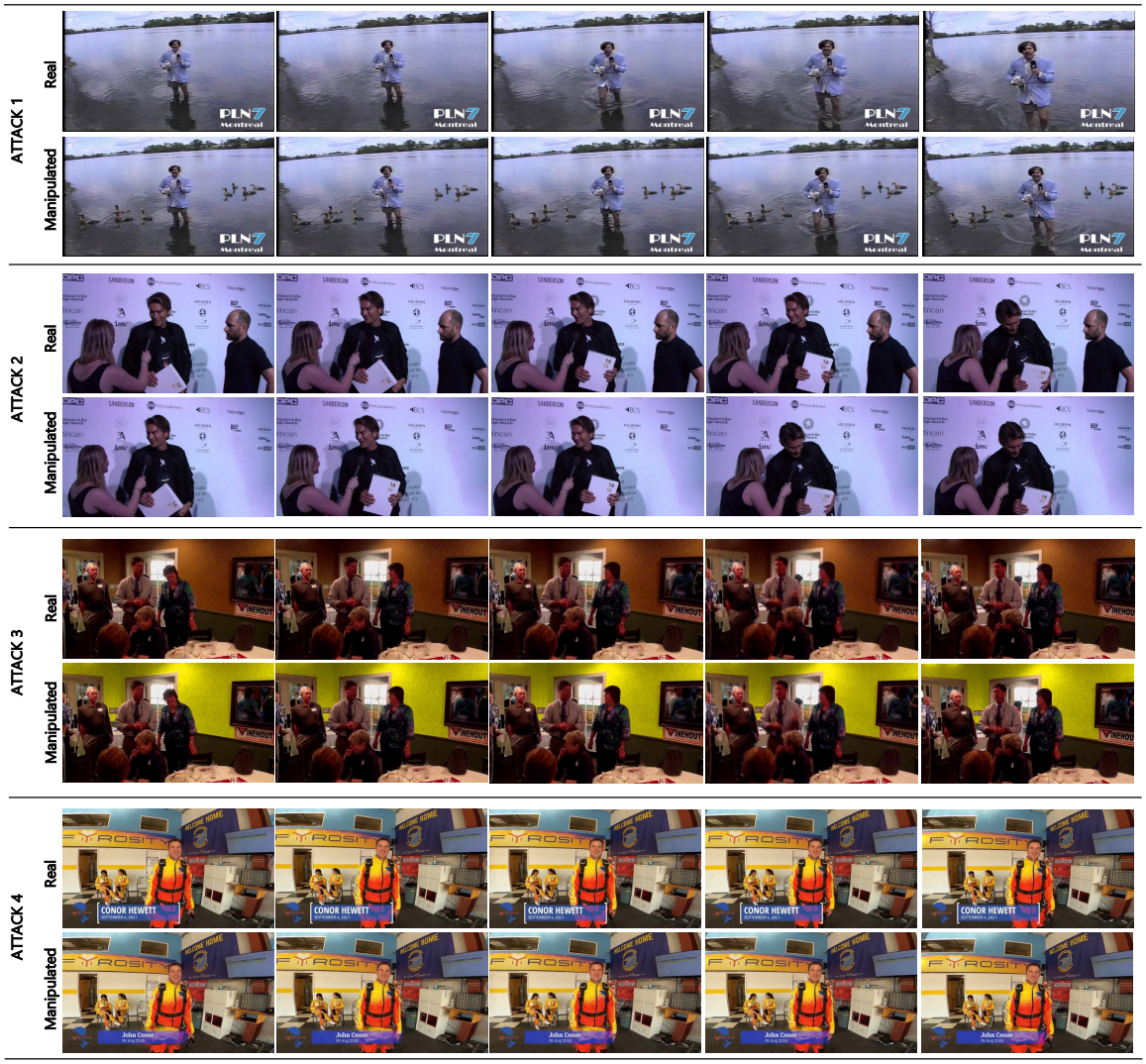


Figure 3.2: **VideoSham Dataset's Qualitative Examples:** We present a series of frames, both for real and manipulated videos for the 4 *spatial* attacks. In ATTACK 1, we add ducks in the water behind the person talking in the microphone. In ATTACK 2, we remove the man in the black shirt to the right corner. In ATTACK 3, we change the color of the walls to yellow. And, finally, in ATTACK 4, we alter the name of the person talking. We were not able to add examples of the 2 *temporal* attacks (ATTACK 5, ATTACK 6) here.

adding/removing entities, and temporal edits. Secondly, although there exist image manipulation datasets that go beyond faces, they cannot be used to detect video manipulations, which require dedicated video datasets. The latter, however, is hard to create due to the manual labor involved. In this work, we go beyond images to release the first video manipulation dataset containing beyond-face manipulations.

We validate our audio-visual deepfake detection model on two benchmark deepfake detection datasets, DeepFakeTIMIT Dataset (Korshunov and Marcel 2018a), and DFDC (Dolhansky et al. 2019). We report the Area Under Curve (AUC) metric on the two datasets for our approach and compare with several prior works. We report the per-video AUC score of 84.4%, which is an improvement of about 9% over SOTA on DFDC, and our network performs at par with prior methods on the DF-TIMIT dataset.

VideoSham consists of 413 real-world videos and their corresponding manipulated versions (total of 826 videos). The videos have diverse scene backgrounds, are context-rich, and contain up to 9 subjects on average. VideoSham is the largest dataset containing manipulated videos generated by professional video editors with varied attacks. A user study was conducted on Amazon Mechanical Turk (AMT) to understand the kind of attack methods that mislead humans the most. In addition, we analyze the performance of existing state-of-the-art deepfake detection algorithms and video forensics algorithms on VideoSham. We find that these techniques are less than 50% effective in distinguishing between a real and a manipulated video. We also discuss some promising ideas from multimodal learning and affective computing that could be helpful in detecting some of these attacks.

The rest of the chapter is structured as follows: We discuss prior work in the domain of video manipulation techniques and detection methods in Section 3.1. Then in Section 3.2 we go over the proposed audio-visual deepfake detection method that uses the correlation between the modality embeddings and the emotion embeddings to detect a deepfake video. And, finally in Section 3.3, we discuss our contribution, the VideoSham dataset.

We elaborate on the characteristics of the dataset and the various experiments conducted to anlayze the dataset further. We conclude with a discussion regarding limitations and some future directions in building more robust video manipulation detection methods in Section 3.4.

3.1 Prior Work in Video Manipulation Detection

In this section, we discuss previous works in detection of manipulated and deceptive media content. To begin with, we first discuss the video manipulation techniques used to create such fake videos in Section 3.1.1. Then in Section 3.1.2, we summarize various datasets and benchmarks for video manipulations. We also survey different techniques used for detecting deepfake videos in Section 3.1.3 and generic video forensic methods in Section 3.1.4.

Table 3.2: **Summary of Attacks Used to Manipulate Videos:** We summarize the various attacks that have been explored in prior literature for manipulating images and videos.

	S.No.	Attack	Method/Software	Description
Spatial	1	Copy-Move and Splicing	Adobe Photoshop™, AfterEffects™	Select and copy-paste
	2	Retouching/Lighting	Adobe Lightroom ™	Change Brightness and Contrast Median Filter
	3	Face Swapping (FS)	FakeApp, FaceSwap (Korshunova et al. 2017) FaceShifter (L. Li et al. 2019) FSGAN (Nirkin, Keller, and Hassner 2019) DeepFaceLab (Perov et al. 2020)	Face transfer
	4	Face Re-enactment (FR)	Neural Textures (Thies, Zollhöfer, and Nießner 2019) First-Order-Motion (Siarohin et al. 2019) Face2Face (Thies, Zollhofer, et al. 2016) IcFace (Tripathy, Kannala, and Rahtu 2020) FSGAN (Nirkin, Keller, and Hassner 2019)	Guided face deformations
	5	Audio-driven FR (AFR) Audio-driven FR (AFR) ATFHP (R. Yi et al. 2020)	Wav2Lip (Prajwal et al. 2020) APB2FACE (Jiangning Zhang et al. 2020)	Audio-guided face reenactment
Temporal	6	Temporal	Adobe Lightroom ™	Frame Dropping, Frame Insertion Shifting in time, Frame Swapping
Geometric	7	Geometric	Adobe Lightroom ™	Cropping, Resizing Rotation, Shifting

3.1.1 Video Manipulation Techniques/Attacks

Manipulation techniques, or attacks, are broadly categorized as spatial (Amerini et al. 2011), temporal (Khelifi and Bouridane 2017), and geometric (Khelifi and Bouridane 2017) in the litera-

ture (see Table 3.2). Basic examples of spatial attacks include copy-move and image/video splicing which correspond to spatially or temporally shifting an object to a different location in the same video or a different video, respectively. Retouching, another common attack, involves aesthetic edits like adjusting brightness, contrast, and other parameters of digital content. More recently, people have used AI to alter facial features to create deepfake videos. AI-based techniques are comprised of two major attack approaches, Face Swapping (Nirkin, Keller, and Hassner 2019; Perov et al. 2020) and Face Re-enactment (Thies, Zollhofer, et al. 2016; Thies, Zollhöfer, and Nießner 2019; Siarohin et al. 2019). Face Swapping switches the subject's face with the face of another person and Face Re-enactment alters the subject's facial expressions. Temporal attacks involve swapping, duplicating, inserting, and deleting frames of video, giving the impression that the video has been sped up or slowed down. Finally, geometric attacks include operations like cropping and rotations.

3.1.2 Video Manipulation Datasets

Creating benchmarks of video manipulations is a challenging task as this may require per-frame manipulations. Some of the datasets (like Khelifi et al. (Khelifi and Bouridane 2017), MTVFD (Al-Sanjary, Ahmed, and Sulong 2016), Liao et al. (Liao and T.-Q. Huang 2013), Su et al. (Su, T. Huang, and J. Yang 2015), Media Forensics Challenge (Guan et al. 2019)) are very small in volume containing $7-200$ videos each, these datasets are also **not publicly available**. Most of these videos have 0 or 1 subjects present in the frame with very little background context. More recently, AI-synthesized attacks like *face swapping*, *face re-enactment*, and *audio-driven face re-enactment* have led to the creation of datasets like UADFV (X. Yang, Y. Li, and Lyu 2019), FaceForensics++ (Rossler et al. 2019), DeeperForensics1.0 (L. Jiang et al. 2020), WildDeepFake (Zi et al. 2020). Because these datasets are generated using learning methods; some of these datasets have upto 100k videos. However all of these datasets have strictly 1 subject per video with the face being predominant part of the frame with no background context at all. Many datasets are missing audio except DFDC (Dolhansky et al.

2019), DF-TIMIT (Korshunov and Marcel 2018a), KoDF (Kwon et al. 2021), FakeAVCeleb (Khalid, Tariq, and Woo 2021), ForgeryNet (Y. He et al. 2021) and SR-DF (J. Wang et al. 2021).

3.1.3 Deepfake Detection Methods

The goal of the deepfake detection approaches is to algorithmically distinguish fake videos from real videos.

Unimodal Methods: Most prior work in deepfake detection decompose videos into frames and explore visual artifacts across frames. For instance, Li et al. (Y. Li and Lyu 2018) propose a Deep Neural Network (DNN) to detect fake videos based on artifacts observed during the face warping step of the generation algorithms. Similarly, Yang et al. (X. Yang, Y. Li, and Lyu 2019) look at inconsistencies in the head poses in the synthesized videos and Matern et al. (Matern, Riess, and Stamminger 2019) capture artifacts in the eyes, teeth and facial contours of the generated faces. Prior works have also experimented with a variety of network architectures. For instance, Nguyen et al. (H. H. Nguyen, Yamagishi, and Echizen 2019) explore capsule structures, Rossler et al. (Rossler et al. 2019) use the XceptionNet, and Zhou et al. (P. Zhou, X. Han, et al. 2017) use a two-stream Convolutional Neural Network (CNN) to achieve SOTA in general-purpose image forgery detection. Previous researchers have also observed and exploited the fact that temporal coherence is not enforced effectively in the synthesis process of deepfakes. For instance, Sabir et al. (Sabir et al. 2019) leveraged the use of spatio-temporal features of video streams to detect deepfakes. Likewise, Guera and Delp et al. (Güera and Delp 2018) highlight that deepfake videos contain intra-frame consistencies and hence use a CNN with a Long Short Term Memory (LSTM) to detect deepfake videos.

Multimodal Methods: While unimodal DeepFake Detection methods have focused only on the facial features of the subject, there has not been much focus on using the multiple modalities that are part of the same video. Jeo and Bang et al. (Jeon, Y. Bang, and Woo 2019) propose FakeTalkerDetect, which is a Siamese-based network to detect the fake videos

generated from the neural talking head models. They perform a classification based on distance. However, the two inputs to their Siamese network are a real and fake video. Korshunov et al. (Korshunov and Marcel 2018b) analyze the lip-syncing inconsistencies using two channels, the audio and visual of moving lips. Krishnamurthy et al. (Krishnamurthy et al. 2018) investigated the problem of detecting deception in real-life videos, which is very different from deepfake detection. They use an MLP-based classifier combining video, audio, and text with Micro-Expression features. Our approach to exploiting the mismatch between two modalities is quite different and complimentary to these methods. However, it is clear that due to the nature of the datasets (single-person, face-centered videos), these approaches focus only on facial cues and audio cues.

3.1.4 Video Forensic Methods

Developments in video forensics literature focus on two specific attacks; Copy-Move and Splicing (Row 1 in Table 3.2) and Temporal attacks (Row 6 in Table 3.2). Most conventional copy-move forgery detection methods mainly consist of three components (Cozzolino, Poggi, and Verdoliva 2015): (1) feature extraction, (2) matching, and (3) post-processing. A variety of features have been explored, e.g., DCT (Discrete Cosine Transform) (Mahmood et al. 2016), DWT (Discrete Wavelet Transform) and KPCA (Kernel Principal Component Analysis) (Bashar et al. 2010), Zernike moments (Ryu et al. 2013). Consequently, some end-to-end deep learning based copy-move forgery detection methods were proposed (Yue Wu, Abd-Almageed, and Natarajan 2018a; Yue Wu, Abd-Almageed, and Natarajan 2018b; Y. Li and Lyu 2019). However these efforts are limited to images. Another interesting development, still in naive stages is deep learning methods to detect inpainting in videos (P. Zhou, Yu, et al. 2021). Some of the methods in detecting temporal attacks (also called intra-frame manipulations) use the consistency of velocity field (Yuxing Wu et al. 2014) and optical flow (Q. Wang et al. 2014). These methods can recognize frame insertion and frame deletion attacks. Similarly, Zhao et al. (D.-N. Zhao, R.-K. Wang, and Lu 2018) use inter-frame similarity analysis

Table 3.3: **Notations Used for our Audio-Visual Deepfake Detection Method:** We highlight the notation and symbols used in the paper.

Symbol	Description
x_y	$x \in \{f, s\}$ denote face and speech features extracted from OpenFace and pyAudioAnalysis. $y \in \{\text{real}, \text{fake}\}$ indicate whether the feature x is real or fake. E.g. f_{real} denotes the *face features* extracted from a *real* video using OpenFace.
a_c^b	$a \in \{e, m\}$ denote emotion embedding and modality embedding. $b \in \{f, s\}$ denote face and speech cues. $c \in \{\text{real}, \text{fake}\}$ indicate whether the embedding a is real or fake. E.g. m_{real}^f denotes the *face modality* embedding generated from a *real* video.
ρ_1	Modality Embedding Similarity Loss (Used in Training)
ρ_2	Emotion Embedding Similarity Loss (Used in Training)
d_m	Face/Speech Modality Embedding Distance (Used in Testing)
d_e	Face/Speech Emotion Embedding Distance (Used in Testing)

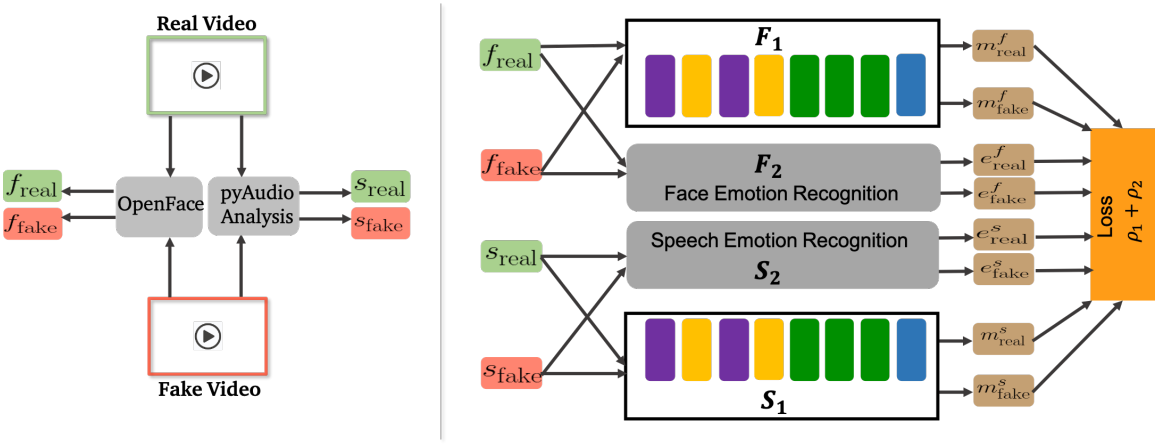
to detect frame duplications in the videos. Finally, Long et al. (Long et al. 2019) propose a coarse-to-fine framework based on deep Convolutional Neural Networks (CNN) to detect potential frame duplications.

3.2 Audio-Visual Deepfake Detection Method

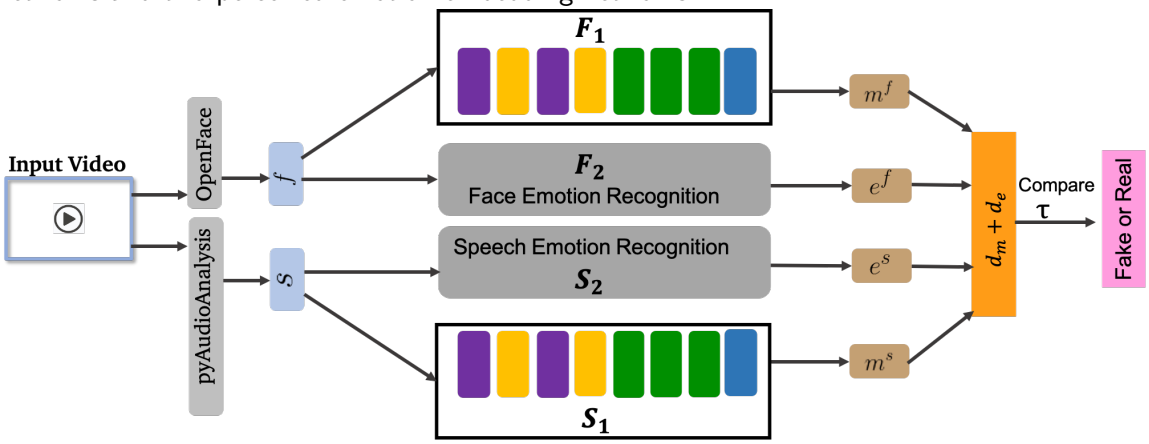
In this section, we present our multimodal approach to detecting deepfake videos. We briefly describe the problem statement and give an overview of our approach in Section 3.2.1. We also elaborate on how our approach is similar to a Siamese Network architecture. We then explain our approach in Section 3.2.2. We elaborate on the modality embeddings and the perceived emotion embedding, the two main components and also explain the similarity score and modified triplet losses used for training the network. We then explain the implementation details in Section 3.2.3 and the experimental results in Section 3.2.4.

3.2.1 Problem Statement and Overview

Problem 3.2.1. *Given an input video with audio modality s and visual modality f, our goal is to detect if it is a deepfake video.*



(a) **Training Routine:** *(left)* We extract facial and speech features from the raw videos (each subject has a real and fake video pair) using OpenFace and pyAudioAnalysis, respectively. *(right)* The extracted features are passed to the training network that consists of two modality embedding networks and two perceived emotion embedding networks.



(b) **Testing Routine:** At runtime, given an input video, our network predicts the label (real or fake).

Figure 3.3: **Audio-Visual Deepfake Detection Model Architecture:** We present an overview diagram for both the training and testing routines of our model. The networks consist of 2D convolutional layers (purple), max-pooling layers (yellow), fully-connected layers (green), and normalization layers (blue). F_1 and S_1 are modality embedding networks and F_2 and S_2 are perceived emotion embedding networks for face and speech, respectively.

Overviews of our training and testing routines are given in Figure 3.3a and Figure 3.3b, respectively. During training, we select one "real" and one "fake" video containing the same subject. We extract the face as well as the speech features, f_{real} and s_{real}, respectively, from the real input video. In a similar fashion, we extract the face and speech features (using OpenFace (Baltrušaitis, Robinson, and L.-P. Morency 2016) and pyAudioAnaly-

sis (Giannakopoulos 2015)), f_{fake} and s_{fake}, respectively, from the fake video. More details about the feature extraction from the raw videos have been presented in Section 3.2.3.2. The extracted features, $f_{\text{real}}, s_{\text{real}}, f_{\text{fake}}, s_{\text{fake}}$, form the inputs to the networks (F_1, F_2, S_1, and S_2), respectively. We train these networks using a combination of two triplet loss functions designed using the similarity scores, denoted by ρ_1 and ρ_2. ρ_1 represents similarity among the facial and speech modalities, and ρ_2 is the similarity between the affect cues (specifically, perceived emotion) from the modalities of both real and fake videos.

Our training method is similar to a Siamese network because we also use the same weights of the network (F_1, F_2, S_1, S_2) to operate on two different inputs, one real video and the other a fake video of the same subject. Unlike regular classification-based neural networks, which perform classification and propagate that loss back, we instead use similarity-based metrics for distinguishing real and fake videos. We model this similarity between these modalities using Triplet loss (explained elaborately in Section 3.2.2.4).

During testing, we are given a single input video, from which we extract the face and speech feature vectors, f and s, respectively. We pass f into F_1 and F_2, and pass s into S_1 and S_2, where F_1, F_2, S_1, and S_2 are used to compute distance metrics, $dist_1$ and $dist_2$. We use a threshold τ, learned during training, to classify the video as real or fake.

We list all notations used throughout in Table 3.3.

3.2.2 Approach

There are two main embedding correlations we compute to infer a video is fake or not. We explain the details below.

3.2.2.1 F_1 and S_1: Video/Audio Modality Embeddings

F_1 and S_1 are neural networks that we use to learn the unit-normalized embeddings for the face and speech modalities, respectively. In Figure 3.3, we depict F_1 and S_1 in both

training and testing routines. They are composed of 2D convolutional layers (purple), max-pooling layers (yellow), and fully connected layers (green). ReLU non-linearity is used between all layers. The last layer is a unit-normalization layer (blue). For both face and speech modalities, F_1 and S_1 return 250-dimensional unit-normalized embeddings.

The training is performed using the following equations:

$$\begin{aligned} m^f_{\text{real}} &= F_1(f_{\text{real}}),\ m^f_{\text{fake}} = F_1(f_{\text{fake}}), \\ m^s_{\text{real}} &= S_1(s_{\text{real}}),\ m^s_{\text{fake}} = S_1(s_{\text{fake}}) \end{aligned} \tag{3.1}$$

And the testing is done using the equations:

$$m^f = F_1(f),\ m^s = S_1(s) \tag{3.2}$$

3.2.2.2 F_2 and S_2: Video/Audio Perceived Emotion Embedding

F_2 and S_2 are neural networks that we use to learn the unit-normalized affect embeddings for the face and speech modalities, respectively. F_2 and S_2 are based on the Memory Fusion Network (MFN) (A. Zadeh et al. 2018a), which is reported to have SOTA performance on both emotion recognition from multiple views or modalities like face and speech. MFN is based on a recurrent neural network architecture with three main components: a system of LSTMs, a Memory Attention Network, and a Gated Memory component. The system of LSTMs takes in different views of the input data. In our case, we adopt the trained single-view version of the MFN, where the face and speech are treated as separate views, i.e. F_2 takes in the video (view only) and S_2 takes in the audio (view only). We pre-trained the F_2 MFN with video and the S_2 MFN with audio from CMU-MOSEI dataset (A. B. Zadeh et al. 2018). The CMU-MOSEI dataset describes the perceived emotion space with 6 discrete emotions following the Ekman model (P. Ekman, Freisen, and Ancoli 1980): happy, sad, angry, fearful, surprise, and disgust, and a "neutral" emotion to denote the absence of any of these emotions. For face and speech modalities in our network, we use 250-

dimensional unit-normalized features constructed from the cross-view patterns learned by F_2 and S_2 respectively.

The training is performed using the following equations:

$$\begin{aligned} e^f_{\text{real}} = F_2(f_{\text{real}}), e^f_{\text{fake}} = F_2(f_{\text{fake}}), \\ e^s_{\text{real}} = S_2(s_{\text{real}}), e^s_{\text{fake}} = S_2(s_{\text{fake}}). \end{aligned} \tag{3.3}$$

And the testing is done using the equations:

$$e^f = F_2(f),\ e^s = S_2(s). \tag{3.4}$$

3.2.2.3 Training Routine

At training time, we use a fake and a real video with the same subject as the input. After, passing extracted features from raw videos ($f_{\text{real}}, f_{\text{fake}}, s_{\text{real}}, s_{\text{fake}}$) through F_1, F_2, S_1, and S_2, we obtain the unit-normalized modality and perceived emotion embeddings as described in Eqs. 3.1-3.4.

Considering an input real and fake video, we first compare f_{real} with f_{fake}, and s_{real} with s_{fake} to understand what modality was manipulated *more* in the fake video. Considering, we identify the face modality to be manipulated more in the fake video, based on these embeddings we compute the first similarity between the real and fake speech and face embeddings as follows:

$$\textbf{Similarity Score 1}:\ L_1 = d(m^s_{\text{real}}, m^f_{\text{real}}) - d(m^s_{\text{real}}, m^f_{\text{fake}}), \tag{3.5}$$

where d denotes the Euclidean distance.

In simpler terms, L_1 is computing the distance between two pairs, $d(m^s_{\text{real}}, m^f_{\text{real}})$ and $d(m^s_{\text{real}}, m^f_{\text{fake}})$. We expect $m^s_{\text{real}}, m^f_{\text{real}}$ to be closer to each other than $m^s_{\text{real}}, m^f_{\text{fake}}$ as it contains a fake face modality. Hence, we expect to maximize this difference. To use this

correlation metric as a loss function to train our model, we formulate it using the notation of Triplet Loss

$$\textbf{Similarity Loss 1}:\ \rho_1 = \max\big(L_1 + m_1, 0\big), \tag{3.6}$$

where m_1 is the margin used for convergence of training.

If we had observed that speech is the more manipulated modality in the fake video, we would formulate L_1 as follows:

$$L_1 = d(m^f_{\text{real}}, m^s_{\text{real}}) - d(m^f_{\text{real}}, m^s_{\text{fake}}).$$

Similarly, we compute the second similarity as the difference in affective cues extracted from the modalities from both real and fake videos. We denote this as follows:

$$\textbf{Similarity Score 2}:\ L_2 = d(e^s_{\text{real}}, e^s_{\text{fake}}) - d(f^s_{\text{real}}, e^f_{\text{fake}}). \tag{3.7}$$

As per prior psychology studies, we expect that similar un-manipulated modalities point towards similar affective cues. Hence, because the input here has a manipulated face modality, we expect $e^s_{\text{real}}, e^s_{\text{fake}}$ to be closer to each other than to $e^f_{\text{real}}, e^f_{\text{fake}}$. To use this as a loss function, we again formulate this using a Triplet loss.

$$\textbf{Similarity Loss 2}:\ \rho_2 = \max(L_2 + m_2, 0), \tag{3.8}$$

where m_2 is the margin.

Again, if the speech was the highly manipulated modality in the fake video, we would formulate L_2 as follows:

$$L_2 = d(e^f_{\text{real}}, e^f_{\text{fake}}) - d(e^f_{\text{real}}, e^s_{\text{fake}}).$$

We use both the similarity scores as the cumulative loss and propagate this back into the

network.

$$Loss = \rho_1 + \rho_2 \tag{3.9}$$

3.2.2.4 Testing Routine

At test time, we only have a single input video that is to be labeled real or fake. After extracting the features, f and s from the raw videos, we perform a forward pass through F_1, F_2, S_1 and S_2, as depicted in Figure 3.3b to obtain modality and perceived emotion embeddings.

To make an inference about real and fake, we compute the following two distance values:

$$\begin{aligned} &\textbf{Distance 1}\text{: } d_m = d(m^f, m^s), \\ &\textbf{Distance 2}\text{: } d_e = d(e^f, e^s). \end{aligned} \tag{3.10}$$

To distinguish between real and fake, we compare d_m and d_e with a threshold, that is, τ empirically learned during training as follows:

$$\text{If } d_m + d_e > \tau,$$

we label the video as a fake video.

Computation of τ: To compute τ, we use the best-trained model and run it on the training set. We compute d_m and d_e for both real and fake videos of the train set. We average these values and find an equidistant number, which serves as a good threshold value. Based on our experiments, the computed value of τ was almost consistent and didn't vary much between datasets.

3.2.3 Implementation Details

In this section, we discuss the datasets (Section 3.2.3.1) that we use to evaluate our proposed audio-visual deepfake detection method, followed with feature extraction details (Section 3.2.3.2). We end with mentioning training and hyperparameters used to obtain the results on the DFDC dataset and the DF-TIMIT dataset (Section 3.2.3.3).

3.2.3.1 Datasets

We perform experiments on the DF-TIMIT (Korshunov and Marcel 2018a) and DFDC (Dolhansky et al. 2019) datasets. We used the entire DF-TIMIT dataset and were able to use randomly sampled $18,000$ videos from DFDC dataset due to computational overhead. Both the datasets are split into training (85%), and testing (15%) sets.

3.2.3.2 Feature Extraction

In our approach (See Figure 3.3), we first extract the face and speech features from the real and fake input videos. We use existing SOTA methods for this purpose. In particular, we use OpenFace (Baltrušaitis, Robinson, and L.-P. Morency 2016) to extract 430-dimensional facial features, including the 2D landmarks positions, head pose orientation, and gaze features. To extract speech features, we use pyAudioAnalysis (Giannakopoulos 2015) to extract 13 Mel Frequency Cepstral Coefficients (MFCC) speech features. Prior works (Chernykh and Prikhodko 2017b; Bougiatiotis and Giannakopoulos 2018; Cai et al. 2019) using audio or speech signals for various tasks like perceived emotion recognition, and speaker recognition also use MFCC features to analyze audio signals.

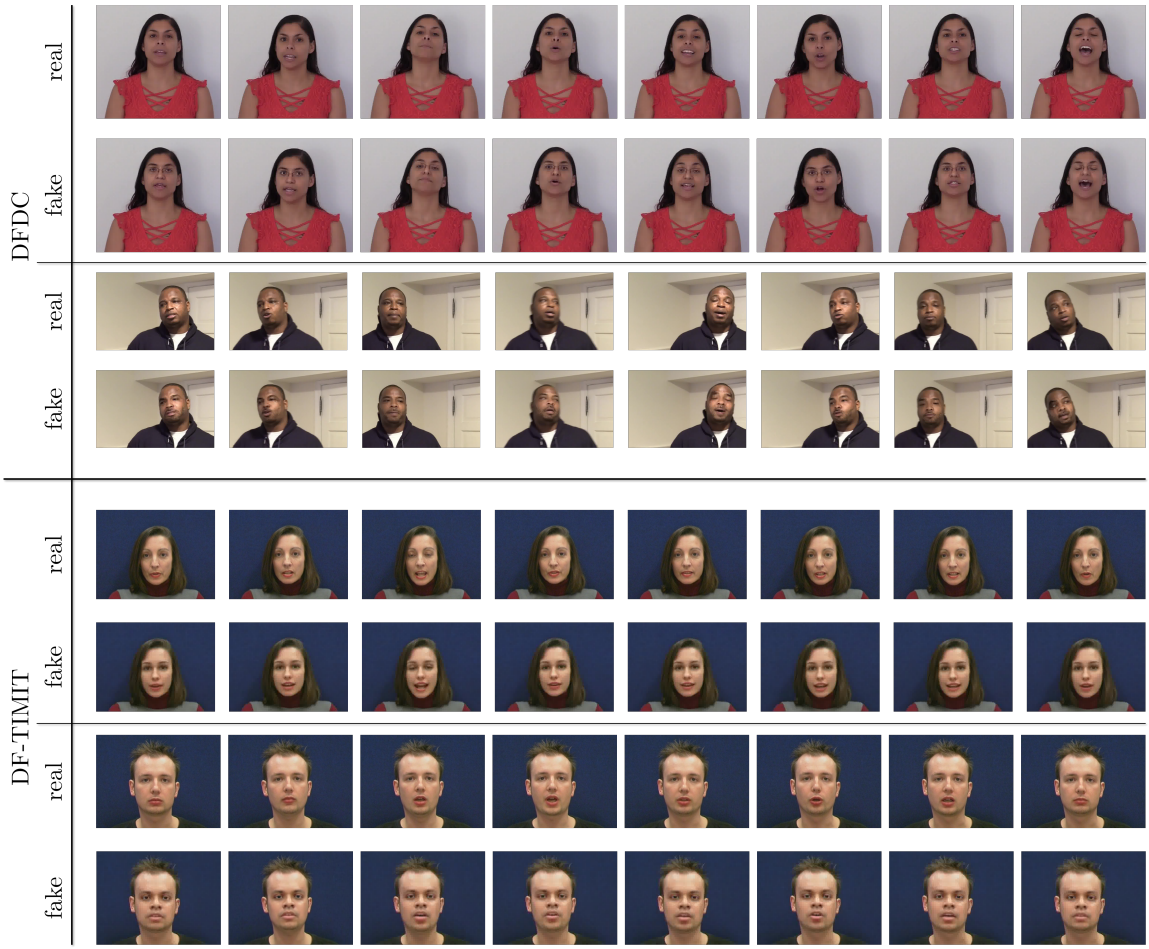


Figure 3.4: **Audio-Visual Deepfake Detection Model's Qualitative Results:** We show results of our model on the DFDC and DF-TIMIT datasets. Our model uses the subjects' audio-visual modalities as well as their perceived emotions to distinguish between real and deepfake videos. The perceived emotions from the speech and facial cues in fake videos are different; however in the case of real videos, the perceived emotions from both modalities are the same.

3.2.3.3 Training Parameters

On the DFDC Dataset, we trained our models with a batch size of 128 for 500 epochs. Due to the significantly smaller size of the DF-TIMIT dataset, we used a batch size of 32 and trained it for 100 epochs. We used Adam optimizer with a learning rate of 0.01. All our results were generated on an NVIDIA GeForce GTX1080 Ti GPU.

3.2.4 Experiments and Results

In this section, we list the state-of-the-art algorithms with which we compare our audio-visual deepfake detection against in Section 3.2.4.1. We also discuss some ablation experiments we performed in Section 3.2.4.3. To further analyze our model, we perform some correlation analysis and present our findings in Section 3.2.4.2. We end with a discussion on some qualitative results in Section 3.2.4.4.

3.2.4.1 Comparison with state-of-the-art Methods

We report and compare per-video AUC Scores of our method against 9 prior deepfake video detection methods on DF-TIMIT and DFDC. To ensure a fair evaluation, while the subset of DFDC the 9 methods were trained and tested are unknown, we select $18,000$ samples randomly and report our numbers. Moreover, as per the nature of the approaches the prior 9 methods report per-frame AUC scores. We have summarized these results in Table 3.4. The following are the prior methods used to compare the performance of our approach on the same datasets.

1. Two-stream (P. Zhou, X. Han, et al. 2017): uses a two-stream CNN to achieve SOTA performance in image-forgery detection. They use standard CNN network architectures to train the model.
2. MesoNet (Afchar et al. 2018) is a CNN-based detection method that targets the microscopic properties of images. AUC scores are reported on two variants.
3. HeadPose (X. Yang, Y. Li, and Lyu 2019) captures inconsistencies in headpose orientation across frames to detect deepfakes.
4. FWA (Y. Li and Lyu 2018) uses a CNN to expose the face warping artifacts introduced by the resizing and interpolation operations.
5. VA (Matern, Riess, and Stamminger 2019) focuses on capturing visual artifacts in the eyes, teeth and facial contours of synthesized faces. Results have been reported on

Table 3.4: **Comparing Our Audio-Visual Deepfake Detection Method with SOTA Methods:** We compare the AUC scores for our method against the state-of-the-art methods on DFDC dataset and the DF-TIMIT dataset. Blue denotes best and green denotes second-best. Our model improves the state-of-the-art method by approximately 9% on the DFDC dataset and achieves accuracy similar to the state-of-the-art on the DF-TIMIT dataset.

S.No.	Methods	Datasets		
		DF-TIMIT (Korshunov and Marcel 2018a)		DFDC (Dolhansky et al. 2019)
		LQ	HQ	
1	Capsule (H. H. Nguyen, Yamagishi, and Echizen 2019)	78.4	74.4	53.3
2	Multi-task (H. H. Nguyen, Fang, et al. 2019)	62.2	55.3	53.6
3	HeadPose (X. Yang, Y. Li, and Lyu 2019)	55.1	53.2	55.9
4	Two-stream (P. Zhou, X. Han, et al. 2017)	83.5	73.5	61.4
5	VA-MLP (Matern, Riess, and Stamminger 2019)	61.4	62.1	61.9
	VA-LogReg	77.0	77.3	66.2
6	MesoInception4	80.4	62.7	73.2
	Meso4 (Afchar et al. 2018)	87.8	68.4	75.3
7	Xception-raw (Rossler et al. 2019)	56.7	54.0	49.9
	Xception-c40	75.8	70.5	69.7
	Xception-c23	95.9	94.4	72.2
8	FWA (Y. Li and Lyu 2018)	**99.9**	93.2	72.7
	DSP-FWA	**99.9**	**99.7**	75.5
	Our Method	96.3	94.9	**84.4**

two standard variants of this method.

6. Xception (Rossler et al. 2019) is a baseline model trained on the FaceForensics++ dataset based on the XceptionNet model. AUC scores have been reported on three variants of the network.
7. Multi-task (H. H. Nguyen, Fang, et al. 2019) uses a CNN to simultaneously detect manipulated images and segment manipulated areas as a multi-task learning problem.
8. Capsule (H. H. Nguyen, Yamagishi, and Echizen 2019) uses capsule structures based on a standard DNN.
9. DSP-FWA is an improved version of FWA (Y. Li and Lyu 2018) with a spatial pyramid pooling module to better handle the variations in resolutions of the original target faces.

While we outperform the DFDC dataset, we have comparable values for the DF-TIMIT dataset. We believe this is because all 640 videos in the DF-TIMIT dataset are face-centered with no body pose. In DFDC, the videos are collected with full-body poses, with the face taking less than 50% of the pixels in each frame. The FWA and DSP-FWA methods identify deepfakes by detecting artifacts caused by affine warping of manipulated faces to match the configuration of the source's face. This is especially useful for the face-centered DF-TIMIT dataset than the DFDC dataset.

Table 3.5: **Ablation Experiments on our Audio-Visual Deepfake Detection Model:** To motivate our model, and the two components (modality and emotion embeddings correlation) we perform ablation studies where we remove one correlation at a time for training and report the AUC scores on both the DF-TIMIT dataset and the DFDC dataset.

Methods	Datasets		
	DF-TIMIT (Korshunov and Marcel 2018a)		DFDC (Dolhansky et al. 2019)
	LQ	HQ	
Our Method w/o Modality Similarity (ρ_1)	92.5	91.7	78.3
Our Method w/o Emotion Similarity (ρ_2)	94.8	93.6	82.8
Our Method	**96.3**	**94.9**	**84.4**

3.2.4.2 Interpreting the Correlations

To better understand the learned embeddings, we plot the distance between the unit-normalized face and speech embeddings learned from F_1 and S_1 on $1,000$ randomly chosen points from the DFDC train set in Figure 3.7(a). We plot $d(m^s_{\text{real}}, m^f_{\text{real}})$ in blue and $d(m^s_{\text{fake}}, m^f_{\text{fake}})$ in orange. It is interesting to see that the peak or the majority of the subjects from real videos have a smaller separation, 0.2 between their embeddings as opposed to the fake videos (0.5). We also plot the number of videos, both fake and real, with a mismatch of perceived emotion labels extracted using F_2 and S_2 in Figure 3.7(b). Of a total of $15,438$ fake videos, $11,301$ showed a mismatch in the labels extracted from face and speech modalities. Similarly, out of $3,180$ real videos, 815 also showed a label mismatch.

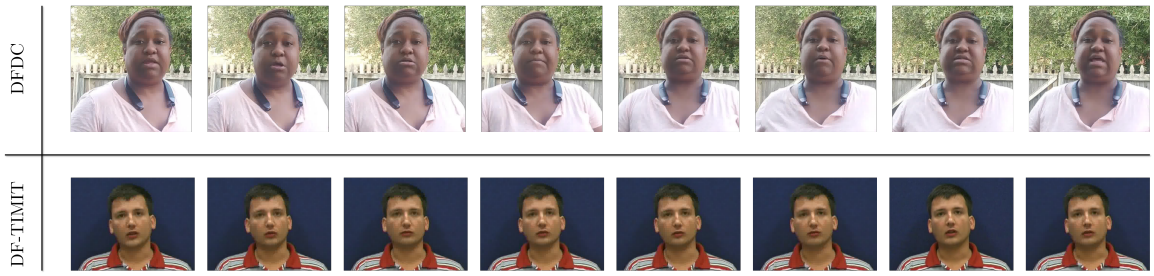


Figure 3.5: **Example Misclassification by the Proposed Audio-Visual Deepfake Detection Method:** We show one sample each from DFDC and DF-TIMIT where our model predicted the two fake videos as real due to incorrect perceived emotion embeddings.

3.2.4.3 Ablation Experiments

As explained in Section 3.2.2.4, we use two distances, based on the modality embedding similarities and perceived emotion embedding similarities, to detects fake videos. To understand and motivate the contribution of each similarity, we perform an ablation study where we run the model using only one correlation for training. We have summarized the results of the ablation experiments in Table 3.5. The modality embedding similarity helps to achieve better AUC scores than the perceived emotion embedding similarity.

3.2.4.4 Qualitative Results

We show some selected frames of videos from both the datasets in Figure 3.4 along with the labels (real/fake). For the qualitative results shown for DFDC, the real video predicted a "neutral" perceived emotion label for both speech and face modality, whereas in the fake video, the face predicted "surprise" and speech predicted "neutral". This result is indeed interpretable because the fake video was generated by manipulating only the face modality and not the speech modality. We see a similar perceived emotion label mismatch for the DF-TIMIT sample as well.

Figure 3.6: **Example Results on In-The-Wild Videos of the proposed Audio-Visual Deepfake Detection Method:** We observe that our model succeeds in the wild. We collect several popular deepfake videos from online social media and our model achieves reasonably good results.

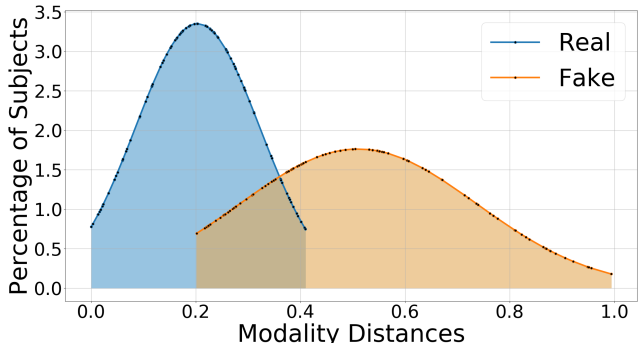
Failure Cases: Our approach models the correlation between two modalities and the associated affective cues to distinguish between real and fake modalities. However, there are multiple instances where the deepfake videos do not contain such a mismatch in terms of perceived emotional classification based on different modalities. This is also because every human being expressed his/her emotions differently. As a result, our model fails to classify such videos as fake. Similarly, both face and speech are modalities that are easy to fake. As a result, it is possible that our method also classifies a real video as a fake video due to this mismatch. In Figure 3.5, we show one such video from both datasets, where our model failed.

Results on Videos in the Wild: We tested the performance of our model on two such deepfake videos obtained from an online social platform (*YouTube Video 1* n.d.; *YouTube Video 2* n.d.). Some frames from this video have been shown in Figure 3.6. While the model successfully classified the first video as a deepfake, it could not be for the second deepfake video.

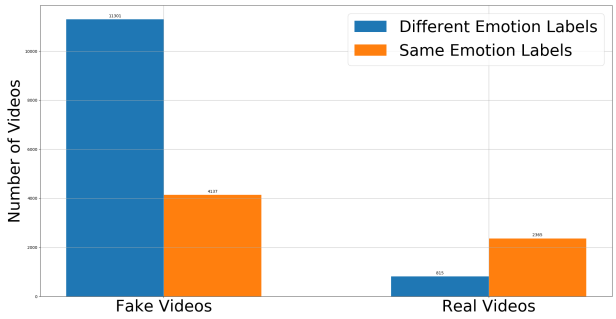
3.3 Video Manipulation Beyond Faces

We now focus on video manipulations which are not restricted to just manipulations on the faces. Towards that end, we present a video manipulation dataset, VIDEOSHAM dataset. We discuss the dataset creation process and compare it with other state-of-the-art datasets

in Section 3.3.1. We further analyze this dataset in Section 3.3.2 with three experiments.



(a) **Modality Embedding Distances:** We plot the percentage of subject videos versus the distance between the face and speech modality embeddings. The figure shows that the distribution of real videos (blue curve) is centered around a lower modality embedding distance (0.2). In contrast, the fake videos (orange curve) are distributed around a higher distance center (0.5). **Conclusion:** We show that audio-visual modalities are more similar in real videos as compared to fake videos.



(b) **Perceived Emotion Embedding in Real and Fake Videos:** The blue and orange bars represent the total number of videos where the perceived emotion labels, obtained from the face and speech modalities, do *not* match, and match, respectively. Of the total $15,438$ fake videos, 73.2% videos were found to contain a mismatch between perceived emotion labels, and for real videos this was only 24%. **Conclusion:** We show that the perceived emotions of subjects, from multiple modalities, are strongly similar in real videos, and often mismatched in fake videos.

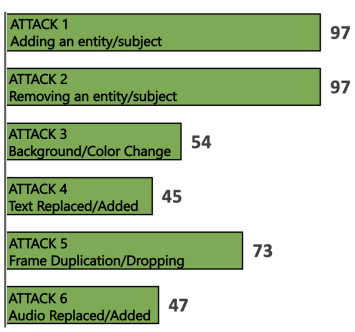
Figure 3.7: **Analysis of the Audio-Visual Deepfake Detection Model (Interpreting Modality and Emotion Embeddings):** We provide an intuitive interpretation of the learned embeddings from F_1, S_1, F_2, S_2 with visualizations. These results back our hypothesis of perceived emotions being highly correlated in real videos as compared to fake videos.

3.3.1 VideoSham Dataset

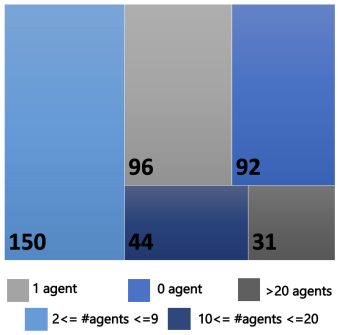
In this section, we present details on the dataset creation process and discuss some of the salient features and characteristics of VIDEOSHAM.

3.3.1.1 Source Videos

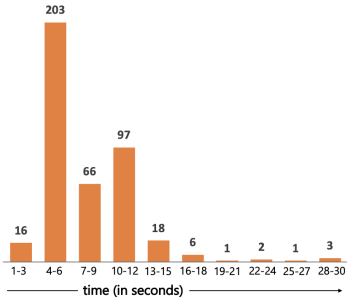
We have a total of 836 videos comprising of 413 original videos and 413 manipulated versions, each corresponding to one of the original videos. We obtain our source videos from an online video website (vimeo [6]) and only include videos attributed with a CC-BY (Creative Commons) license. In addition, we avoid videos with brands, children, objectionable content, TV show/movie clips and videos with copyrighted music. We trim these original videos to a specific length (upto $5-30$ seconds) before we perform any manipulation attack.



(a) **Attack distribution:** Distribution of videos that are attacked with different manipulation techniques. Attacks $1-4$ are spatial attacks, and Attacks $5-6$ are temporal attacks.



(b) **Density distribution:** Distribution of videos according to the number of persons present in each video. This is considerably high w.r.t. the existing datasets.



(c) **Duration distribution:** Distribution of videos according to duration or length of each video (in seconds). The average length of our videos is 8 seconds.

Figure 3.8: **Characteristics and Summary of VideoSham Dataset Statistics:** We visually present various statistics for VIDEOSHAM for better insights.

[6] www.vimeo.com.

3.3.1.2 Manipulation Attacks

We employ a total of 6 manipulation attacks for creating our dataset. As per prior literature, we also categorize these attacks into spatial and temporal attacks[7]. We visually show the distribution of the attacks in Figure 3.8a. We describe each of the attacks below.

- ATTACK 1 (Adding an entity/subject): In this attack, we select an entity or a subject from some other sources and place them in the current video. This attack is somewhat similar to a copy-move attack.

- ATTACK 2 (Removing an entity/subject): In this attack, we basically select an entity or a subject in the video and remove it from all the frames and fill in the gap with background settings. To do this, we used content-aware fill in Adobe AfterEffects™ and some deep learning methods for generating masks (K. He, Gkioxari, et al. 2017) and performing video inpainting (D. Kim et al. 2019; D. Kim et al. 2020).

- ATTACK 3 (Background/Color Change): We focus on a particular aspect of the video, and change the background of the video, or the color of a small entity in the video.

- ATTACK 4 (Text Replaced/Added): We perform edits like adding some text in the video or removing or replacing already existing text in the video.

- ATTACK 5 (Frames Duplication / Removal/ Dropping): This attack is specifically to render the video temporally inconsistent. We choose to perform one of these manipulations, randomly duplicating frames, and removing or dropping frames in the video. This also includes slowing down a video.

- ATTACK 6 (Audio Replaced): Audio modality is a very important aspect of videos. To manipulate this, we replace the existing audio with some other audio.

[7]We do not use geometric attacks, as they have been shown to be easily detected.

We visually depict the 4 spatial attacks (ATTACK 1, ATTACK 2, ATTACK 3, and ATTACK 4) in Figure 3.2.

3.3.1.3 Manipulated Videos

We worked with 3 professional video editors hired on Upwork [8]. The editors were shortlisted based on their experience and were well-versed with Adobe AfterEffects™, the software used for creating these edits. Each editor was assigned tasks, i.e. source videos, start and end timestamps to be edited, and a one-line description of the manipulation to be performed. We provide all videos and the attacks performed for every video.

3.3.1.4 Dataset Analysis

In Figure 3.8a, we present the distribution of attacks for the 413 videos, each lasting $1 - 31$ seconds. The average length of videos in our dataset is around 8 seconds long. We also run an object detection model [9] to count the number of people/agents in every video (Figure 3.8b). More than 80% of the videos in our dataset contain at least one subject.

3.3.2 Experiments and Results

We elaborate on three experiments we perform to highlight the importance, novelty and use case of VIDEOSHAM. To begin with, we present the analysis of how well humans fair in detecting these attacks in Section 3.3.2.1, followed by an analysis of the performance of state-of-the-art deepfake detection methods and video forensic techniques in Section 3.3.2.2. Finally, in Section 3.3.2.3, we present some ideas and preliminary results for using interdisciplinary ideas for detecting such attacks.

[8] www.upwork.com.
[9] https://github.com/roboflow-ai.

3.3.2.1 Expt 1: How Well Do Humans Perform?

Setup: We first shortlist 60 videos from VIDEOSHAM. Out of these, 30 videos are real and the remaining 30 are manipulated (5 videos per attack). We recruit human participants from Amazon Mechanical Turk (AMT) and show each video to 20 participants. The participants are requested to watch the full video; followed by two questions. In the first question, the participants are asked to respond to the following prompt in either a yes or no - "Do you believe this video has been manipulated/edited to misrepresent facts?". And, in the second question, we ask them to explain in a sentence what they felt was manipulated with the following prompt- "If you answered YES above, what region or aspect of this video, do you believe is manipulated.". Note that participants are not informed whether videos are manipulated or not. They are also not informed about the set of attacks. We show the setup in Figure 3.9 which was used to collect a total of 1200 responses from AMT participants.

Study Analysis: We summarize the responses of the user study in Table 3.6. Both the real and manipulated videos receive 600 responses each. We observe that out of the 600 responses (corresponding to the 30 real videos), 342, i.e., 57% were correctly identified as real. Similarly, out of 600 responses for the manipulated videos, 389 were incorrectly identified as real, i.e., 35.2% of these responses correctly identified manipulated videos. Analyzing the responses by the type of attack, we observe that human participants are able to identify 45% of the videos manipulated using ATTACK 6. For the other attack types, the proportion of manipulated videos labeled as 'fake' ranges from 13-31%. Furthermore, we notice that human participants are able to more successfully identify manipulated videos that are modified using temporal attacks (ATTACK 5 and ATTACK 6) than spatial attacks (ATTACK $1-$ ATTACK 4). Moreover, we also received a number of responses from participants explaining their rationale behind reporting a manipulated video. From the responses received, there is no clear evidence that suggests that participants are able to identify the manipulated region/kind in case of spatial edits. But, they were somewhat

able to correctly identify the manipulated edit in case of temporal attacks. This would imply that a subset of our selection of attacks is indiscernible to the human eye.

Statistical Tests: Next we consider statistical tests to see if humans are able to tell a real video from a manipulated video. We consider the following quantities for this test, define $p_1 = P(\text{declaring video real}|\text{real video})$.
Also, let $p_2 = P(\text{declaring video real}|\text{manipulated video})$. If humans are able to tell real videos apart from manipulated videos, we expect p_1 to be larger than p_2. Hence, we test the one-sided statistical hypothesis:

$$H_0 : p_1 = p_2 \quad \text{against} \quad H_1 : p_1 > p_2.$$

We test this hypothesis with the test statistic $(p_1 - p_2)$[10]. In Table 3.6 we present the difference of proportions as well as the one-sided $p-$value of the test for each attack type. The first thing to note is that when combining across all attack types (last row), we see that even though p_1 is slightly bigger than p_2, this difference is not statistically significant ($p-$value of 0.177). This suggests that our edits are not discernible to human evaluators. When we break it down by attack type, we observe that only for ATTACK 6 (audio replacement), humans are more likely to declare such edits as manipulations ($p-$value < 0.001). For ATTACK 4 (text replaced or added), there is weak statistical evidence of humans detecting this manipulation ($p-$value of 0.097). For all other attacks, there is no statistical evidence that humans can tell when a video has been manipulated using that strategy. It is particularly telling that when an entity/subject is added or removed (ATTACKs 1 and 2), more of our human subjects declare such manipulated videos as real than they declare unedited videos. This shows how modern editing tools can be used to manipulate videos in a way that humans have no way of telling such edits just by looking at the video. This observation establishes the need to build high-quality video manipulation

[10] Given our sample size, we have an 86% statistical power of detecting a difference if the true values are $p_1 = 0.75$ and $p_2 = 0.74$.

Table 3.6: **Analysis of Human Performance on VideoSham Dataset:** We summarize how well the videos in the proposed VIDEOSHAM dataset decieve human participants. We observe that participants are unable to detect ATTACK 3 (26%) and ATTACK 4 (31%). Videos manipulated using ATTACK 5 are relatively easier to detect (75%)

(A) 1200 *responses* (20 participants × (30 real + 30 manipulated videos))										
GT		**#Resp (Total)**	**Rep Real**	**Rep Fake**	p_1	p_2	$p_1 - p_2$	$p-$ **value**	**CI(l)**	**CI(u)**
Real		600	454	146	0.757		0	–	–	–
Manipulated	1	100	87	13	0.757	0.87	−0.113	0.991	−0.182	1
	2	100	79	21	0.757	0.79	−0.033	0.725	−0.112	1
	3	100	74	26	0.757	0.74	0.016	0.408	−0.066	1
	4	100	69	31	0.757	0.69	0.066	0.097	−0.020	1
	5	100	75	25	0.757	0.75	0.006	0.493	−0.076	1
	6	100	55	45	0.757	0.55	0.207	$<$ **0.001**	0.114	1
		600	439	161	0.757	0.732	0.0250	0.177	−0.018	1

detection algorithms that can label manipulated videos at scale.

3.3.2.2 Expt 2: How Well Do Machines Perform?

To answer this question better, we evaluate state-of-the-art deepfake detection methods and video forensics techniques on VIDEOSHAM.

Deepfake Detection Methods: We evaluate Li et al. (Y. Li and Lyu 2019), XceptionNet (Rossler et al. 2019) and Mittal et al. (Mittal, Bhattacharya, et al. 2020a) on VIDEOSHAM. Deepfake videos generated using data-driven methods can only synthesize face images of a fixed size, and they must undergo an affine warping to match the configuration of the source's face. Due to resolution inconsistencies between warped face and background context, there are various artifacts on the synthesized faces. Li et al. (Y. Li and Lyu 2019) detects such artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network (CNN) model. On the other hand, XceptionNet (Rossler et al. 2019) is a transfer learning model which is also a CNN architecture, which was originally trained for the classical object detection task and later finetuned

Table 3.7: **Analysis of Machine Performance (Video Forensics Techniques and Deepfake Detection Methods) on VideoSham Dataset:** We evaluate 3 state-of-the-art deepfake detection methods and 2 video forensics techniques on VIDEOSHAM dataset. It is apparent that these algorithms do not perform well on VIDEOSHAMdataset, speaking to the complexity and diversity of the videos.

GT			#Vid.	Predicted									
				Deepfake Detection Methods						Video Forensics Techniques			
				Li et al (Y. Li and Lyu 2019)		MesoNet (Afchar et al. 2018)		Mittal et al. (Mittal, Bhattacharya, et al. 2020a)		Long et al. (Long et al. 2019)		Liu et al. (Y. Liu et al. 2021)	
				Predicted		Predicted		Predicted		Predicted		Predicted	
				Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Real			413	188	225	167	246	238	175	219	194	234	179
Manipulated	Attack	1	97	76	21	93	4	92	5	86	11	68	29
		2	97	63	34	84	13	66	31	84	13	46	51
		3	54	35	19	37	17	49	5	50	4	38	16
		4	45	32	13	34	11	42	3	39	6	34	11
		5	73	70	3	67	6	54	19	25	48	68	5
		6	47	45	2	44	3	31	16	38	9	41	6
			413	321	92	359	54	334	79	322	91	295	118

for deepfake detection on FaceForensics++ dataset. Finally, Mittal et al. propose an approach that simultaneously exploits the audio (speech) and video (face) modalities and also the perceived emotion features extracted from both modalities to detect any falsification or alteration in the input video. They use the correlation between the modalities to detect a fake video.

Table 3.8: **Preliminary Quantitative Results of Gaze and Affect Ideas on VideoSham Dataset (Expt 3):** For some preliminary analysis, we explore two ideas, *gaze* and *affect* of all agents involved. We observe that these two ideas in itself can effectively detect manipulations of the kind, ATTACK 1 and ATTACK 2.

GroundTruth	# videos	Reported Real	Reported Manipulated
Real	413	286	127
Attack 1	97	32	65
Attack 2	97	35	62

Video Forensics Techniques: We evaluate Long et al. (Long et al. 2019) and Liu et al. (Y. Liu et al. 2021) on VIDEOSHAM. Both of these methods are state-of-the-art methods in video forensics literature. While, Long et al. (Long et al. 2019) is specifically for detecting cases of frame duplications in a video, Liu et al. (Y. Liu et al. 2021) specifically focus on

detecting copy-move attacks. For all the methods, we use pretrained models and report the results when evaluated on VIDEOSHAM in Table 3.7.

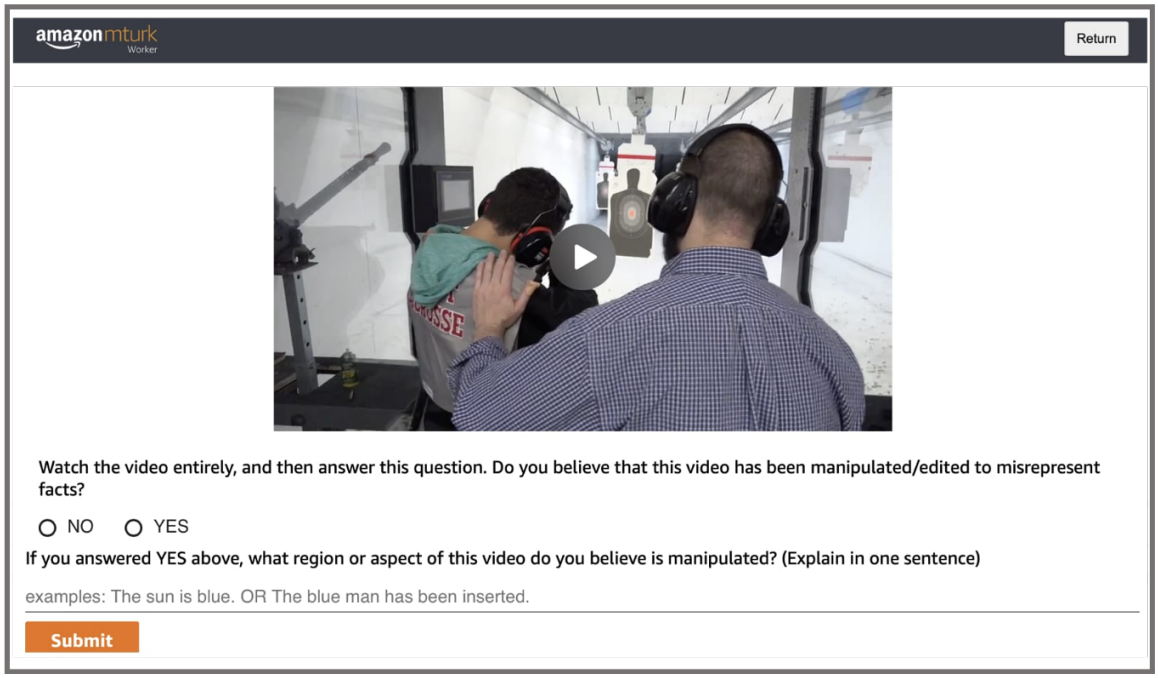


Figure 3.9: **User Study Setup Used to Access How Well Humans Fair on VideoSham Dataset:** We present the Amazon Mechanical Turk setup used (Section 3.3.2.1).

Study Analysis: All the five methods are less then 50% accurate on VIDEOSHAM dataset. This is not very surprising and quite expected, as all the deepfake methods (Li et al. (Y. Li and Lyu 2019), MesoNet (Afchar et al. 2018), and Mittal et al. (Mittal, Bhattacharya, et al. 2020a)) are trained specifically to look for manipulations in faces. Moreover, these method are not used to inferencing on videos with more or less than 1 person in the frame and with so much context information. Hence, we observe that these methods are only inferring based on artifacts caught near the face regions in the VIDEOSHAM videos. We also observe that Mittal et al. specifically are able to detect some of the temporal manipulations well; which is because the method is trained to look for a correlation between audio and visual modalities. Similarly, even the video forensics techniques are specifically performing well on attacks that they have been trained for, i.e. ATTACK 5 for Long et al. and ATTACK 1 and ATTACK 2 for Liu et al. (H. Liu et al. 2016). ATTACK 3 (color change) and ATTACK

4 (text replacement) tend to remain hard to be detected by most of these methods.

3.3.2.3 Expt 3: Beyond DeepFake Detection and Video Forensic Techniques

One can observe from the experiments in the previous section, that all the methods are largely dependent on the visual artifacts. However, given the diversity of attacks used to manipulate videos, we hypothesize the use of inter-agent and multimodal analysis models for detecting such manipulations. We show preliminary results in Figure 3.10.

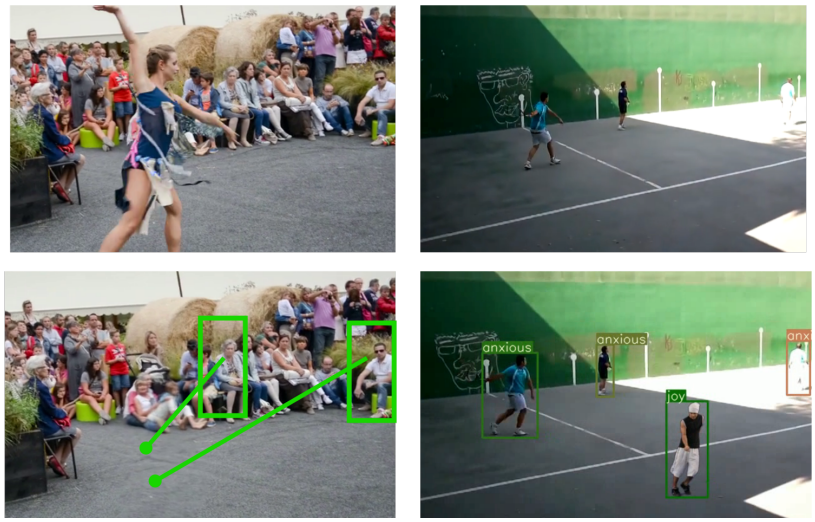


Figure 3.10: **Preliminary Experiment 3, Qualitative Results of Eye Gaze and Affect Ideas on the VideoSham Dataset for Detecting Video Manipulations:** We show the output of the automated techniques used to identify manipulated videos in VIDEOSHAM. *(Column 1)* In the first column, we remove the main subject from the foreground. We identify this as a manipulated image using a gaze tracking algorithm by noting that there is no object at the location of the crowd's gaze direction. *(Column 2)* Here, we manipulate an image by inserting the man in the black shirt. We use emotion recognition techniques to infer that this false subject has an affective state that is not in tune with those of the other players.

Strategy 1 (Gaze): To begin with, we believe that tracking gaze of subjects can be useful for detection experiments. Gaze following is a task in computer vision to identify objects and regions that the subject of interest is focusing on. The idea behind this strategy is to identify manipulated images by using gaze following to locate "absent" targets and/or "out-of-context" subjects in the video. To perform some preliminary analysis we deploy GazeFollow (Recasens 2016). More specifically, for each frame, we begin by obtain the spatial

coordinates of the subject's head's bounding box and pass this information as input to the gaze tracking algorithm, GazeFollow (Recasens 2016), which outputs the location of the subject's gaze. The final step in this strategy is to run an object detector to obtain a confidence score c_g corresponding to an object present at the gaze location. A low confidence score indicates a manipulated frame.

Strategy 2 (Affect): In this strategy we propose the use of affective cues. When we track and look for affective disparities in affective state of different subjects. Prior works in psychology (Kleinsmith and Bianchi-Berthouze 2013) and empirical works (Mittal, Guhan, et al. 2020) that subjects in social settings often share affective states. We use facial expressions, body postures and scene understanding to perceive the affective states of all subjects. We use the model EmotiCon (Mittal, Guhan, et al. 2020) trained on EMOTIC dataset (Kosti, Jose M Alvarez, et al. 2017a) to perceive these affective states and obtain an affective confidence score c_a. By empirically assigning a threshold, τ on the two confidence scores, we flag a video as manipulated. We observe that these two techniques help detect ATTACK 1 and ATTACK 2 significantly well. We add quantitative results for the same in Table 3.8. We show two qualitative results of these ideas in Figure 3.10.

Experiment 3 shows that, in addition to human assessment, specialized deepfake detection techniques, and video forensics, other approaches that are not intended for identifying manipulated videos can be used.

3.4 Conclusions, Limitations and Future Work

In this section, we first presented learning-based audio-visual method for detecting deepfake videos. We use the similarity between audio-visual modalities and the similarity between the affective cues of the two modalities to infer whether a video is "real" or "fake". We evaluated our method on two benchmark audio-visual deepfake datasets, DFDC, and DF-TIMIT. We also presented VIDEOSHAM dataset, a collection of 826 videos that contain videos manipulated by professional video editors using one of 6 spatial or temporal

attacks. We also performed 3 experiments to further analyze the proposed dataset. Our goal with the expt 1 and expt 2 was to understand how well humans can detect some of the manipulations that occur today circulated on social media. We also wanted to understand if the developments in the deepfake detection and video forensic literature match up to these manipulation attacks. Finally, through expt 3 we want to propagate the idea of using ideas beyond detection of visual artifacts for scalable models for video manipulation detection.

Our deepfake detection method has some limitations. First, our approach could result in misclassifications on both the datasets, as compared to the one in the real video. Given different representations of expressing perceived emotions, our approach can also find a mismatch in the modalities of real videos, and (incorrectly) classify them as fake. Furthermore, many of the deepfake datasets primarily contain more than one person per video. We may need to extend our approach to take into account the perceived emotional state of multiple persons in the video and come with a possible scheme for deepfake detection. From the analysis of the three experiments, we conclude from expt 1 and expt 2 that both humans and machines (5 methods shortlisted) struggle to detect these manipulations successfully. We believe that these are attacks of concern, as they are going undetected even by human participants. Moreover, we emphasize that these manipulations play a big role in many real-world video manipulations (Figure 3.1). More generally, we believe that computer vision algorithms perform almost comparable to humans in most of these ATTACKS. However, most methods are very attack-specific and do not generalize well to other attacks. Mostly every deepfake detection method fails to handle videos with more than 1 subject and hence has a very limited scope. Also, importantly most of the deepfake detection methods require huge amounts of training samples; and this is not a realistic assumption. It is important to build methods which can be less computationally intensive and at the same time are also able to generalize well. Similarly, methods in video forensics also are only able to handle very specific attacks. These are less dependent on

data, but computationally expensive as they are more or less, inference based methods.

In the future, fro detecting deepfake videos, we would like to look into incorporating more modalities and even context to infer whether a video is a deepfake or not. We would also like to combine our approach with the existing ideas of detecting visual artifacts like lip-speech synchronisation, head-pose orientation, and specific artifacts in teeth, nose and eyes across frames for better performance. Additionally, we would like to approach better methods for using audio cues.

We believe following are some knowledge gaps and research agendas that can help the society combat the increasing problem of misinformation, frauds and cybercrimes occurring due to manipulated media content shared online.

1. There is a need to build detection models focused on more diverse attacks or video manipulations. Through VIDEOSHAM, we attempted to include some of the attacks that have not been studied before owing to a lack of a dataset. We hope this dataset can be a step towards achieving better detection models for all the 6 attacks.

2. Moreover it is important to increase the scope of detection ideas being used currently for detecting manipulations. Current methods are extremely focused on visual perception. Our goal through experiment 3 was to show through very preliminary analysis that ideas based on inter-agent dynamics and multimodal cues can be a promising literature source. Another promising idea, is to include domain knowledge in detecting manipulations; as humans we have some contextual information which the detection models severely suffer from.

3. Largely all existing methods require a significant amount of training data to train the models. But, with newer manipulations and attacks on videos, it will become impossible to keep up with detection models for the same. We need to reduce the dependence on training data build detection models that are as generalizable as possible to potential attacks.

Chapter 4

Affective Analysis of Multimedia Content

In affective computing, perceiving the emotions conveyed in images and videos has found applications in digital content management (D. Joshi et al. 2014; Y. Wang and B. Li 2015), digital marketing (McDuff, El Kaliouby, Cohn, et al. 2014; Hussain, Mingda Zhang, X. Zhang, et al. 2017; Ye and Kovashka 2018), education (Downs and Strand 2008; Alqahtani and Ramzan 2019), and healthcare (Cohn et al. 2009). Such applications have resulted in automated ranking systems, indexing systems (H. L. Wang and Cheong 2006; Wiley 2003; Kimura et al. n.d.), and more personalized movie recommendation systems (E. Oliveira, Martins, and Chambel 2011).

Affective analysis of movies has been a problem of interest in the community (Quan, V.-T. Nguyen, and Tran 2018; Jin et al. 2017), along similar lines. In our work, we explore the problem of affective analysis of movies with the goal of understanding the emotions that the movies invoke in the audience.

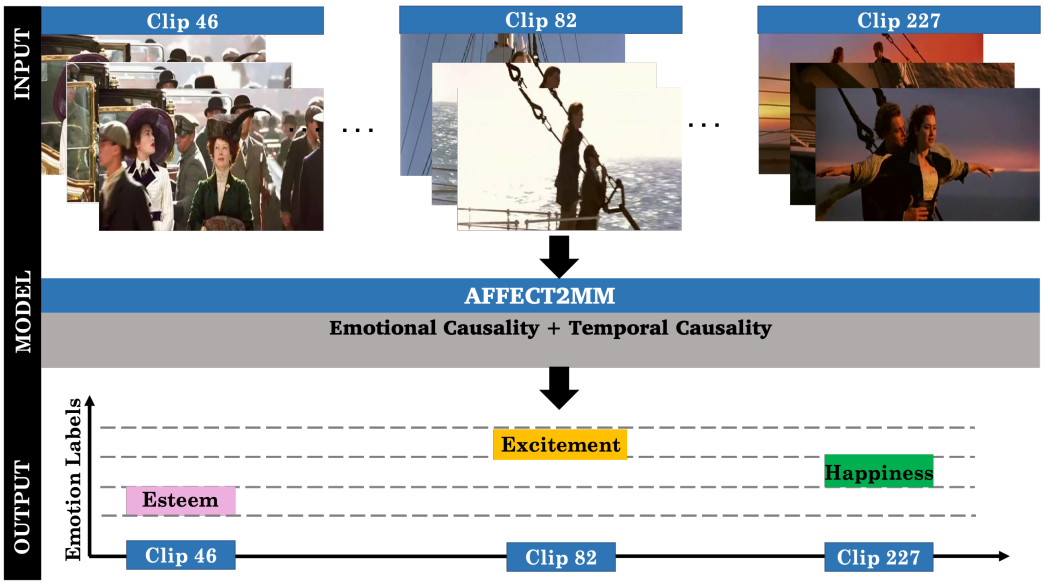


Figure 4.1: **Time-Series Emotion Perception Model:** We present Affect2MM, a learning model for time-series emotion perception for movies. We input multimedia content in the form of multiple clips and predict emotion labels for each clip. Affect2MM is based on the theory of emotion causation and also borrows the idea for temporal causality. We show some example clip-frames from the movie 'Titanic', a part of the MovieGraphs Dataset, and corresponding emotion labels.

There has been a growing interest (Desmond Ong et al. 2019) in dynamically modeling emotions over time ('time series emotion recognition' among the affective computing community. This underlying problem uses temporally continuous data (facial features, speech features, or other modality features) from multimedia content as input and predicts the emotion labels at multiple timestamps (clips) of the input. To aid in solving this time-series problem, several time-series emotion datasets have been proposed (McKeown et al. 2010; Hussain, Mingda Zhang, X. Zhang, et al. 2017; Trigeorgis et al. 2016; Kossaifi et al. 2019; Barros et al. 2018; Desmond Ong et al. 2019). While these datasets focus more on single-person emotional narratives recorded in controlled settings, multimedia datasets (movie databases) like LIRIS-ACCEDE (Baveye et al. 2015) and MovieGraphs (Vicol et al. 2018) (annotated for per-clip emotion labels) are also being explored for time-series emotion perception tasks.

There have been various efforts to understand how humans reason and interpret emotions resulting in various theories of emotion causation based on physiological, neurologi-

cal, and cognitive frameworks. One such theory is the "emotional causality" (Coëgnarts and Kravanja 2016) that has been developed from the Causal Theory of Perception (Hyman 1992) and Conceptual Metaphor Theory (Athanasiadou and Tabakowska 2010; Niemeier and Dirven 1997; Kövecses 2003). "Emotional Causality" refers to the understanding that an experience of emotion is embedded in a chain of events comprising of an (a) outer event; (b) an emotional state; and (c) a physiological response. Few works have explored such emotional causality for emotion perception in multimedia tasks.

Movies, as time-series multimedia content, model multiple human-centric situations and are temporally very long, but coherent sequences. To be able to reason about emotions invoked in various clips of the movie, it is important to develop a causal understanding of the story. Generic methods of handling such temporality include recurrent neural networks (Landis and Koch 1977; Soleymani, Asghari-Esfeden, et al. 2014), attention-mechanisms (T. Chen et al. 2014), graph modeling (R. Guo et al. 2020), and statistical methods like Granger causality (C. W. J. Granger 1969). Explicit modeling of causality in the context of time-series emotion perception has been relatively unexplored.

Emotion labels have been explored extensively, both as discrete (Kosti, Jose M. Alvarez, et al. 2017c) and continuous (Mehrabian and Russell 1974a), in affective analysis. The Valence-Arousal-Dominance (VAD) model (Mehrabian and Russell 1974a) is used for representing emotions in a continuous space on a 3D plane with independent axes for valence, arousal, and dominance values. The Valence axis indicates how pleasant (vs. unpleasant) the emotion is; the Arousal axis indicates how high (or low) the physiological intensity of the emotion is, and the dominance axis indicates how much the emotion is tied to the assertion of high (vs. low) social status. A combination of 3 values picked from each axis represents a categorical emotion like 'angry' or 'sad', much like how an (x, y, z) point represents a physical location in 3-D Euclidean space. Various transformations (Mehrabian 1996; Hoffmann et al. 2012) can be used to map discrete emotion labels to the VAD space. In this work, we work with both continuous emotion labels and discrete emotion labels.

Main Contributions: The following are the novel contributions of our work.

1. We present Affect2MM, a learning-based method for capturing the dynamics of emotion over time. Affect2MM aligns with the psychological theory of "emotional causality" to better model the emotions evoked by each clip of a movie.
2. To better model the temporal causality in movies for long-range multimedia content like movies, we use attention methods and Granger causality to explicitly model the temporal causality (between clips in movies). Our approach can be used for predicting both continuous emotion labels (valence and arousal) and also discrete class labels.

We evaluate our method on two movie datasets, MovieGraphs (Vicol et al. 2018) and the LIRIS-ACCEDE (Baveye et al. 2015) dataset. To showcase our method's generalizability, we also evaluate and compare our method on the SENDv1 (Desmond Ong et al. 2019) dataset, a single-person emotional narratives dataset.

The rest of the chapter is structured as follows: We discuss some prior work in multimedia analysis in Section 4.1. In Section 4.2 and Section 4.3 we formally state the research problem, go over some background concepts (Co-Attention mechanism and Granger Causality), followed by detailed explanation of Affect2MM's model architecture. We discuss implementation details in Section 4.4. We elaborate on quantitative, qualitative and ablation experiments performed to evaluate Affect2MM in Section 4.5. Finally we conclude with a discussion on some limitations of Affect2MM and future directions in Section 4.6.

4.1 Prior Work in Multimedia Analysis

In this section, we summarize prior work done in related domains. We first look into available literature in affective analysis of multimedia content and various applications in Section 4.1.1. In Section 4.1.2, we discuss the visual affective representations that

have previously been explored for related tasks. We also discuss conventional methods to model temporality in Section 4.1.3. In Section 4.1.4, we discuss "emotional causality" and other theories of emotion, as suggested in Psychology literature, and the need to align computation models with these theories.

4.1.1 Affective Analysis of Multimedia Content

Various approaches have explored understanding emotions evoked by multimedia content. Chen et al. (T. Chen et al. 2014), Ali et al. (Ali et al. 2017), Wei et al. (Wei et al. 2020) have performed affective multi-class classification on images collected from popular websites. Pilli et al. (Pilli et al. 2020), Hussain et al. (Hussain, Mingda Zhang, Xiaozhong Zhang, et al. 2017), and Zhang et al. (Huaizheng Zhang et al. 2020) studied predicting sentiments in image advertisements. Vedula et al. (Vedula et al. 2017) extended this idea and developed an advertisement recommendation system using sentiments in the advertisement content. Understanding the relationship between emotional responses to content has been learned by recording viewer responses in many controlled-use studies. Based on such studies, researchers have used facial responses (McDuff, El Kaliouby, Cohn, et al. 2014; Kassam 2010; Micu and Plummer 2010), facial electromyography (EMG) (Micu and Plummer 2010), electroencephalogram (EEG), pupillary response and gaze (Teixeira, R. Picard, and El Kaliouby 2014; Soleymani, Pantic, and Pun 2011), smile (McDuff, El Kaliouby, Demirdjian, et al. 2013; Teixeira, R. Picard, and El Kaliouby 2014; S. Yang et al. 2014). Similarly, Philippot (Philippot 1993) and Gross and Levenson (J. J. Gross and Levenson 1995) were the first ones to propose a small dataset of clips from films with participants' responses to them in controlled lab settings. McDuff et al. (McDuff, El Kaliouby, Cohn, et al. 2014) have recognized the constraint of collecting such data in controlled settings and have proposed collecting large-scale viewer data using webcams. Other movie-based datasets with some affective annotations that were used are HUMAINE (Douglas-Cowie et al. 2007), FilmStim (Schaefer et al. 2010), DEAP (Koelstra et al. n.d.), MAHNOB-HCI (Soleymani, Pantic, and Pun 2011), EMDB (Carvalho et al. 2012),

MovieGraphs (Vicol et al. 2018) and LIRIS-ACCEDE (Baveye et al. 2015). In our work, we evaluate our method on some of these datasets.

4.1.2 Visual Affective Rich Representation

Analyzing viewers response (face and body posture reactions, EEG, and ECG signals) for multimedia content to understand the affective content is not scalable due to a lack of data. Subsequent efforts are being made to perform the same analysis using cues directly from the content: images/videos/movies. Wei et al. (Wei et al. 2020), and Panda et al. (Panda et al. 2018) report that general visual feature extractions used for standard vision tasks (object recognition) do not scale up in terms of performance for emotion-related tasks. Scene and place descriptors extracted from the image/frames have been explored (Ali et al. 2017) to understand the affective component better. Researchers (Quan, V.-T. Nguyen, and Tran 2018; Jin et al. 2017; Batziou et al. 2018) have also focused on the visual aesthetics of the multimedia content to understand how they affect the evocation of emotions from viewers. Zhao et al. (Y. Zhao et al. 2019) have also used background music to analyze the affective state. In our work, we present a total of 6 features that we believe can help in a better understanding of multimedia content.

4.1.3 Modeling Temporality in Time-Series Models

While emotional causality guides us to study the emotional state of a clip, it becomes important to keep track of and model the temporality of emotions in long multimedia content like movies. While recurrent neural network architectures are inherently designed to keep track of such dependencies, explicitly modeling temporal causality has been a norm. Attention-based methods and their many variants (Cheng, Dong, and Lapata 2016; T. Chen et al. 2014) have been used before to help models learn important parts to "attend to" in time series data. Furthermore, most commonly, causality is frequently studied using graph structures and modeling (R. Guo et al. 2020; Glymour, K. Zhang, and Spirtes 2019). More

conventionally, statistical methods like Granger causality (C. W. J. Granger 1969) have been used to quantify the dependence of past events on future time series. More recently (Sindhwani, Minh, and Aurélie C. Lozano 2013; Tank et al. 2018), there have been multiple attempts to incorporate similar behaviors in neural networks.

4.1.4 Theory of Emotions: A Look Into Psychology

Some of the major theories of emotion that reason causation of emotion can be grouped into physiological, neurological, and cognitive theories. Physiological theories like the James-Lange Theory (JAMES 1884) and the Cannon-Bard Theory (Cannon 1927) suggest an external stimulus leads to a physiological reaction, and the emotional reaction is dependent on how the physical reaction is interpreted. Schachter-Singer Theory (Stanley Schachter and Jerome Singer 1962) propose a two-factor theory according to which a stimulus leads to a physiological response that is then cognitively interpreted and labeled, resulting in an emotion. Many more such theories attempt to understand how humans think about emotional states, also called "affective cognition" by (D. C. Ong, Zaki, and Goodman 2019). They also provide a taxonomy of inferences within affective cognition and also provide a model of the intuitive theory of emotion modeled as a Bayesian network. Another term scholars have used to understand this domain is "emotional causality" (Coëgnarts and Kravanja 2016).

4.2 Problem Formulation and Background Concepts

We formally state our problem in Section 4.2.1. We then give a brief overview of co-attention mechanism (Section 4.2.2) and Granger causality (Section 4.2.3).

4.2.1 Problem Statement

We consider multimedia content, which can be any image, video, or audio in the form of multiple *clips*. Each clip, $\mathcal{C}$, is a short sequence of frames that contains information

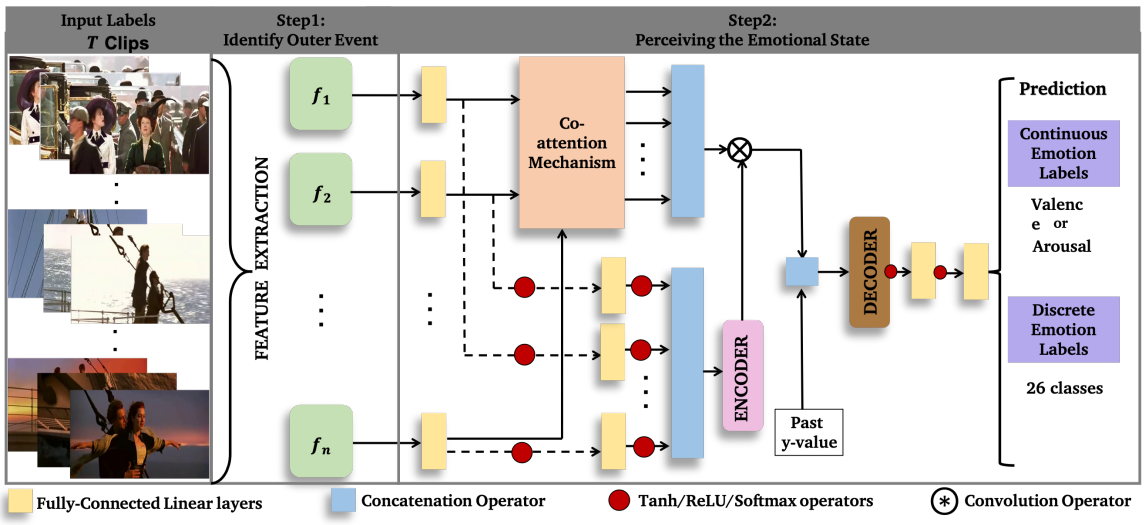


Figure 4.2: **Affect2MM Model Architecture:** We use the components of Emotional Causality to infer the emotional state depicted in the multimedia content. Given the input T clips, we first extract features, $f_1 \dots f_n$, to build the affective-rich representation and help identify the outer event/cause. We then pass these feature representations through a co-attention mechanism, cLSTM-based encoder, and decoder to perceive the emotional state in either continuous values (valence/arousal) or discrete labels. The networks consist of fully-connected layers (yellow), concat layers (blue), and non-linearity (red dots).

arising from multiple modalities such as facial expressions of the actors, their speech, the transcribed dialogues, the visual aesthetics of the scene, the description of the scene, and so on. Our goal is to predict the emotion label of each clip using the information contained in the corresponding frames. More formally,

Problem 4.2.1. *Given a set of clips spanning T time-steps, $\mathcal{C}^{1:T} = \{f_1^{1:T}, f_2^{1:T}, \dots f_p^{1:T}\}$, where $f_i^{1:T}$ denotes the i^{th} feature for each clip, we predict the emotion label, y, for each clip, denoted by $y^{1:T}$.*

Our formulation allows y to be general in that it may represent either categorical emotion labels ("happy", "sad", ...) or continuous real-valued labels, depending on the dataset. In this work, y can be either represent 26 categorical labels or one of the 2 real-valued labels - valence and arousal.

4.2.2 Co-attention Mechanism

Attention mechanisms (Bahdanau, Cho, and Bengio 2016) in neural networks have a long history in both the NLP and vision communities. The broad idea is to allow the learning model to only attend to relevant parts of the input rather than encoding the entire input. While intuitive, modeling it requires the computation of attention scores, a weighted matrix depicting the dependence of all outputs on each of the inputs. These attention scores are then used for making inferences. There have been many variants like self-attention (T. Chen et al. 2014), hierarchical and nested attention (Zichao Yang et al. 2016), and attention flow (Seo et al. 2018). One such variant is the co-attention mechanism (C. Xiong, Zhong, and Socher 2018).

The co-attention mechanism calculates the shallow semantic similarity between the two inputs and uses that as a reference. Given two inputs $(\{u_p^t\}_{t=1}^{t=T}, \{u_q^t\}_{t=1}^{t=T})$, where T is the timestep, co-attention aligns them by constructing a soft-alignment matrix S. Each (i, j) entry of the matrix S is the multiplication of the $\tanh$ activation for both the inputs.

$$S_{i,j} = \tanh(w_p u_p^i) \cdot \tanh(w_q u_q^j)$$

We feed these inputs through a single-layer neural network followed by a softmax function to generate the attention distribution α (Q. Zhang et al. 2018).

$$\begin{aligned} z &= \tanh\left(w_p u_p^i \oplus w_q u_q^j\right) \\ \alpha &= \mathsf{softmax}\left(w_\alpha z\right) \end{aligned} \tag{4.1}$$

We use $\oplus$ to denote the concatenation of the two features. Based on the attention distribution, the attended/relevant parts can be obtained as follows:

$$\hat{u}_q^j = \sum_i \alpha_{i,j} \cdot u_p^i$$

4.2.3 Granger Causality (GC)

Granger causality (C. W. J. Granger 1969) is a way of quantifying the extent to which the past activity of some time-series data is predictive of another time-series data. In our case, we use the term "time-series data" to refer to an element of $\mathcal{C}^{1:T} = \{f_1^{1:T}, f_2^{1:T}, \dots f_p^{1:T}\}$. The p features may include facial action units of the actors involved, the audio features, embeddings of the text/transcript, scene descriptions, action or situation descriptions, visual aesthetics of the clips, etc. In other words, the sequence of, say, facial features across T clips, $f_1^{1:T}$ is a time-series modality data. We use GC to reason the causality between different modalities connected across time. If there exists a causality between $f_1^{1:T}$ and $f_2^{1:T}$, then we say that f_1 *Granger-causes* f_2.

In our approach, we first explore the existence of Granger causality between time-series modalities in temporal clips. If GC is found to exist between two modalities, then we show that it can be used to improve the accuracy of any general emotion recognition system.

Existence of GC in LSTMs: The most widely-used model to estimate GC in linear systems is the Vector AutoRegressive (VAR) model (Lütkepohl 2005; Aurelie C Lozano et al. 2009). However, solutions for the linear models do not scale to non-linear time series data (Teräsvirta, Tjøstheim, and C. Granger 2011; Sindhwani, Minh, and Aurélie C. Lozano 2013). Tank et al. (Tank et al. 2018) address non-linear GC in neural network architectures like Multi-Layer Perceptron (MLP) and LSTMs by introducing a new LSTM architecture called component-wise LSTM (cLSTM), which models each time-series modality through independent LSTM networks. More formally, consider the input

$$X_{1:T} = \Pi \left(\begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,T} \\ f_{2,1} & f_{2,2} & \dots & f_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ f_{p,1} & f_{p,2} & \dots & f_{p,T} \end{bmatrix} \right) \tag{4.2}$$

where the j^{th} row of $X_{1:T}$ is denoted as $X^{(j)}$ and linear transformation Π is applied row-wise (more details in Section 4.3.3.1). To check the existence of GC, we begin by passing $X_{1:T}$ to the cLSTM which is made up of p separate LSTMs, $L^{(j)}$, $j = 1, 2, \ldots, p$, each operating on the input $X_{1:T}$. In our system, this stage is executed using Equation 4.6. Then, following the theoretical framework put forth in Tank et al. (Tank et al. 2018), we check for GC by solving the following optimization problem using line search (Armijo 1966):

$$W^{*(j)} = \arg\min_{W^{(j)}} \sum_{t} (x_{i,t} - L^{(j)}(X_{1:T}))^2 + \lambda \sum_{k=1}^{p} \left\| W_{:k}^{(j)} \right\|_2, \tag{4.3}$$

If $\left\| W_{:k}^{*(j)} \right\|_2 = 0$, then the cLSTM identifies the k^{th} time series modality to be *Granger non-casual* to the j^{th} time-series modality. Otherwise, it is Granger-causal. In Equation 4.3, $W^{(j)}$ denotes the weights of the LSTM $L^{(j)}$ and $W_{:k}^{(j)}$ denotes the k^{th} column of $W^{(j)}$.

Table 4.1: **Summary of Time-Series Emotion Recognition Datasets (Long-Form Multimedia Content):** We summarise details of the three long-form multimedia content datasets used to study time-series emotion recognition. We use all the three datasets to evaluate the proposed Affect2MM model. We provide details about the dataset splits, features used and the evaluation metrics. We also mention emotion labels (C:continuous and D:discrete) available and used for training.

		Affective Features						Labels	Evaluation
Dataset	Train/Val/Test	Facial	Audio	Textual	VA	Scene	Situation	(C/D)	Metrics
SENDv1 (D. C. Ong, Zaki, and Goodman 2019)	144/39/41 videos	✓	✓	✓	×	×	×	Valence (C)	CCC
MovieGraphs (Vicol et al. 2018)	33/7/10 movies	✓	✓	✓	✓	✓	✓	26 classes (D)	Accuracy
LIRIS-ACCEDE (Baveye et al. 2015)	44/10/12 movies	✓	✓	×	✓	✓	×	Valence/Arousal (C)	MSE

4.3 Affect2MM: Our Approach

We first give an overview of our method and some notations used in Section 4.3.1. This is followed by a description of each of the individual components of Affect2MM (Section 4.3.3.2 – Section 4.3.3.4.

4.3.1 Overview

We present an overview of our time-series emotion prediction model for multimedia content, Affect2MM in Figure 4.2. Our approach draws on the theory of Emotional Causality to infer the emotional state that is conveyed in the multimedia content. The theory of Emotion Causality (Coëgnarts and Kravanja 2016) consists of the following main events:

(a) Identifying Outer Event: This stage refers to a stimulus that is contained in the multimedia content that causes an emotion in the consumer of that multimedia content. Such emotion-causing stimuli in multimedia commonly include affective cues such as facial expressions and speech of the actors, but often also include the visual aesthetic features as well as the context of the multimedia content.

We work along these lines and extract these affective cues and visual aesthetic features from the multimedia content and build an affective-rich representation, described further in Section 4.3.2.

(b) Perceiving the Emotional State: This event refers to the formation of an emotional state in the consumer of the multimedia content upon receiving the stimuli contained in the latter. We develop a novel co-attention-based deep neural network that predicts a perceived emotion conveyed through multimedia.

Lastly, the theory discussed thus far assumes a notion of causality between the two events. That is, the "outer event" *causes* the "perceived emotional state". In order to computationally model this causality, we investigate the causality between the affective cues using Granger causality (C. W. J. Granger 1969). In the following sections, we describe the computational details of the two events of Emotion Causality.

4.3.2 Building the Affective-Rich Representation

Our goal here is to build a representation of features that are capable of inciting a perceived emotional state in the audience. We extract features that can contribute to an affective-

rich representation from the input $\mathcal{C}^{1:T}$ clips of the multimedia content. For each clip, we extract at most p feature representations, including but not limited to– the facial action units of the actors involved, the audio features, embeddings of the text/transcript, scene descriptions, action or situation descriptions and visual aesthetics of the clips,

$$f_i^{1:T} = \mathcal{F}_i(\mathcal{C}^{1:T}) \tag{4.4}$$

where $f_i^{1:T} \in \mathbb{R}^{p \times |\text{feature size}|_i \times T}$ for $i = 1, 2, \ldots, p$ is the i^{th} feature representation obtained using the extractor, $\mathcal{F}_i$. We describe the feature extractor $\mathcal{F}$ in Section 4.4.2. Affect2MM can work with any subset of the 6 features representations mentioned. This is also shown in our results.

4.3.3 Perceiving the Emotional State

We use the features, $f_1^{1:T}, f_2^{1:T} \ldots f_p^{1:T}$, generated from the first event to predict perceived emotions in the consumer of the multimedia content. Our approach consists of a deep neural network that uses co-attention to learn and, ultimately, to be able to focus on the useful elements of the features at different time instances. The key intuition here is that the relevance of features varies during the length of the multimedia. For example, the scene-setting and the visual aesthetics may stand out towards the beginning of a movie, as the viewers are not acquainted with the actors, rather are trying to build the scene in their minds. But later, the facial expressions and the speech of the actors may develop a stronger presence as the viewers get to know the actors and their stories. Co-attention helps capture the time-varying nature of the pairwise interdependent affective-rich features, implicitly handling transitivity amongst related modalities.

We begin by simultaneously encoding the feature representation (Eqn. 4.6) using recurrent neural network architectures (cLSTMs in our case) and computing the co-attention between pairs of features (Eqn. 4.7). The results from these operations are convolved to

obtain a final context vector (Eqn. 4.8). The final perceived emotion label is computed by passing the context vector through an LSTM decoder, followed by a combination of linear and non-linear matrix multiplication operations (Eqn. 4.10).

4.3.3.1 cLSTM Encoder

We encode the p time-series features using a cLSTM. As described in Section 4.2.3, we first compute $X_{1:T}$ using Equation 4.2. More formally, the transformation, $\Pi(f_i^{1:T})$ is given by,

$$\Pi(\cdot) = \text{softmax}(\phi(\cdot))$$

Then, the transformed feature representation of each time series is computed as,

$$x_i^{1:T} = \Pi(f_i^{1:T}), \tag{4.5}$$

where ϕ is a linear operation with suitable weights. We can then obtain $X^{1:T}$ by row-wise stacking $x_i^{1:T}$ as follows,

$$X^{1:T} = x_1^{1:T} \oslash x_2^{1:T} \oslash \cdots \oslash x_p^{1:T},$$

where $\oslash$ is a row-wise stacking operator. The stacked inputs are encoded using the cLSTM encoder as defined in Eq. 4.3.

$$h_{\text{enc}} = \text{cLSTM}(X^{1:T}) \tag{4.6}$$

4.3.3.2 Co-attention Scores

We learn the interactions between the different features $f_1^{1:T}, f_2^{1:T} \ldots f_p^{1:T}$ by aligning and combining modalities pairwise using Eq. 4.1. We obtain m values of α_k, where $k = 1, 2, \ldots, m$ and $m = \binom{p}{2}$, corresponding to each pairwise co-attention operation.

$$\alpha_k = \text{Co-attention}(\phi(f_{k_1^{1:T}}), \phi(f_{k_2}^{1:T})) \tag{4.7}$$

where k_1, k_2 are indices denoting one of the p feature vectors, $\phi(.)$ are linear layer operators with appropriate weights and $\oplus$ is the concatenation operator. We obtain a final α as,

$$\alpha = \alpha_1 \oplus \alpha_2 \oplus \cdots \oplus \alpha_m$$

4.3.3.3 Decoder

Finally, once we have computed h_{enc} and α, we can obtain the 'context vector'd, by convolving the attention weights (α) with the encoded feature representation, h_{enc},

$$d = h_{\text{enc}} \otimes \alpha \tag{4.8}$$

The idea is to retain more of the corresponding values of h_{enc} where the attention weights α are high and less information from the parts where the attention weights are low. Finally, the decoder uses the 'context vector', d and the past emotion values y', concatenated together. We simply concatenate these two vectors and feed the merged vector to the decoder. This returns $\hat{y}$ the predicted labels.

$$h_{\text{dec}} = \text{LSTM}(d \oplus y') \tag{4.9}$$

$$\hat{y} = \phi(\text{ReLU}(\phi(h_{\text{dec}}))) \tag{4.10}$$

4.3.3.4 Vector Auto-Regressive (VAR) training of Shared cLSTM Encoder

The cLSTM encoder is shared in a multitask learning fashion to regress future values of input multimodal time-series data $\{f_1^{1:T}, f_2^{1:T}, \ldots f_p^{1:T}\}$ through vector autoregressive training (Tank et al. 2018) as shown in Equation 4.3. The VAR training can be viewed as a secondary task to the primary emotion prediction task, involving shared encoder layers.

The group lasso penalty applied to the columns of the $W_{:k}^{(j)}$ matrix forces the cLSTM to predict the future values of k^{th} modality without relying on the past values of j^{th} modality. The ridge regularization penalty (r), the parameter for non-smooth regularization (l), and the learning rate of VAR training determine the sparsity of the Granger Causal relationship matrix (Tank et al. 2018) to mitigate the problem of multicollinearity amongst the multivariate time series input.

Table 4.2: **Generating Affective Labels for MovieGraphs Dataset:** We list the attributes of Moviegraphs used for all clips for grouping them into 26 discrete emotion labels. These generated labels for each clips are used for training and evaluating Affect2MM.

Class Id	Emotion Labels	Attribute Labels available in MovieGraphs
0	Affection	loving, friendly
1	Anger	anger, furious, resentful, outraged, vengeful
2	Annoyance	annoy, frustrated, irritated, agitated, bitter, insensitive, exasperated, displeased
3	Anticipation	optimistic, hopeful, imaginative, eager
4	Aversion	disgusted, horrified, hateful
5	Confident	confident, proud, stubborn, defiant, independent, convincing
6	Disapproval	disapproving, hostile, unfriendly, mean, disrespectful, mocking, condescending, cunning, manipulative, nasty, deceitful, conceited, sleazy, greedy, rebellious, petty
7	Disconnection	indifferent, bored, distracted, distant, uninterested, self-centered, lonely, cynical, restrained, unimpressed, dismissive
8	Disquietment	worried, nervous, tense, anxious, afraid, alarmed, suspicious, uncomfortable, hesitant, reluctant, insecure, stressed, unsatisfied, solemn, submissive
9	Doubt/Conf	confused, skeptical, indecisive
10	Embarrassment	embarrassed, ashamed, humiliated
11	Engagement	curious, serious, intrigued, persistent, interested, attentive, fascinated
12	Esteem	respectful, grateful
13	Excitement	excited, enthusiastic, energetic, playful, impatient, panicky, impulsive, hasty
14	Fatigue	tire, sleepy, drowsy
15	Fear	scared, fearful, timid, terrified
16	Happiness	cheerful, delighted, happy, amused, laughing, thrilled, smiling, pleased, overwhelmed, ecstatic, exuberant
17	Pain	pain
18	Peace	content, relieved, relaxed, calm, quiet, satisfied, reserved, carefree
19	Pleasure	funny, attracted, aroused, hedonistic, pleasant, flattered, entertaining, mesmerized
20	Sadness	sad, melancholy, upset, disappointed, discouraged, grumpy, crying, regretful, grief-stricken, depressed, heartbroken, remorseful, hopeless, pensive, miserable
21	Sensitivity	apologetic, nostalgic
22	Suffering	offended, hurt, insulted, ignorant, disturbed, abusive, offensive
23	Surprise	surprise, surprised, shocked, amazed, startled, astonished, speechless, disbelieving, incredulous
24	Sympathy	kind, compassionate, supportive, sympathetic, encouraging, thoughtful, understanding, generous, concerned, dependable, caring, forgiving, reassuring, gentle
25	Yearning	jealous, determined, aggressive, desperate, focused, dedicated, diligent
26	None	-

Table 4.3: **Affect2MM Hyperparameter Details:** We summarize the values of hyperparameters used to train and evaluate Affect2MM for SENDv1, MovieGraph and LIRIS-ACCEDE datasets.

	Dataset		
Hyperparameters	SENDv1	MovieGraphs	LIRIS-ACCEDE
Dropout Ratio	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam
Embedding Dimension (Facial Expression)	32	204	204
Embedding Dimension (Visual Aesthetics)	N/A	41	317
Embedding Dimension (Audio)	88	300	1584
Embedding Dimension (Action/Situation)	N/A	300	N/A
Embedding Dimension (Scene)	N/A	300	4096
Embedding Dimension (Textual)	300	300	N/A
Hidden Dimension (Linear Layers)	512	1024	512
Hidden Dimension (cLSTM Encoder)	512	1024	512
Hidden Dimension (LSTM Decoder)	512	1024	512
Number of hidden layers	1	1	1
Epochs	10	10	20
Batch Size	1	1	1
Learning Rate (Affect2MM model)	1e-4	1e-4	1e-4
Learning Rate (Multivariate VAR)	0.001	0.001	0.001
Activation Function of Linear layers	LeakyReLU	LeakyReLU	LeakyReLU
Dimension of FCN Layers	[(512 x 4), (4 x 1)]	[(1024 x 4), (4 x 27)]	[(512 x 4), (4 x 2)]

4.4 Implementation Details

We give an overview of the datasets (SENDv1, LIRIS-ACCEDE, and MovieGraphs) used for evaluating Affect2MM in Section 4.4.1. In Section 4.4.2, we mention the features extracted for training. Finally, we discuss the training hyperparameters in Section 4.4.3.

4.4.1 Datasets

Here we discuss details of the three datasets we used to evaluate and benchmark Affect2MM. For further readability, we have summarized these details in Table 4.1.

SENDv1 Dataset: The dataset consists of video clips of people recounting important and emotional life stories unscripted. The videos have been recorded in a face-centered setting with no background. Valence ratings collected are provided for every 0.5 seconds of the video.

Evaluation Metrics: Most previous works in this dataset have reported the Concordance Correlation Coefficient (CCC) (Lawrence and K. Lin 1989) for validation and test splits along

with the standard deviation. The CCC captures the expected discrepancy between the two vectors, compared to the expected discrepancy if the two vectors were uncorrelated and can be calculated as follows:

$$\text{CCC}(Y, \hat{Y}) = \frac{2\ \text{Corr}(Y, \hat{Y})\sigma_Y \sigma_{\hat{Y}}}{\sigma_Y^2 \sigma_{\hat{Y}}^2 + (\mu_y - \mu_{\hat{Y}})^2}$$

where $\text{Corr}(Y, \hat{Y})$ is the Pearson correlation between the groundtruth (Y) and predicted valence values ($\hat{Y}$) for T clips/timestamps of a video, and the μ and σ are the mean and standard deviation predictions.

LIRIS-ACCEDE Dataset: The dataset contains videos from a set of 160 professionally made and amateur movies. The movies are in various languages including English, Italian, Spanish, and French.

Evaluation Metrics: Consistent with prior methods we report the Mean Squared Error (MSE) for the test splits. Given predicted valence value, $\hat{y}$ and true valence value y for T clips of a movie, we compute the MSE as follows:

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

MovieGraphs Dataset: This dataset provides detailed graph-based annotations of social situations depicted in movie clips for 51 popular movies. Each graph consists of several types of nodes to capture the emotional and physical attributes of actors, their relationships, and the interactions between them. The dataset was collected and manually annotated using crowd-sourcing methods. We process all clips in every movie and extract the available 'emotional' attributes and group them in 26 discrete emotion labels as described by Kosti et al. (Kosti, Jose M. Alvarez, et al. 2017c). To create the discrete 26 emotion labels, we provide the attribute values we used as "emotional keywords" (Table 4.2). These are then used for training the MovieGraphs dataset.

Evaluation Metrics: Because the labels in MovieGraphs are discrete, detailing 26 emo-

tion classes, we report and compare against the Top-1 accuracy.

4.4.2 Feature Extraction

To build up an affective rich representation of the multimedia content, we work with a total of 6 features: facial, audio, textual, visual aesthetics, scene, and situation descriptors. We summarize the features available for training Affect2MM in Table 4.1. Below we provide details for extracting features from each of the three datasets.

SENDv1 Dataset: We used the facial features, audio features, and text embeddings as input for SENDv1. We used the extracted features for the three modalities as explained by Ong et al. (Desmond Ong et al. 2019). To summarize, for audio features they used openSMILE v$2.3.0$ (Eyben et al. 2013) to extract the extended GeMAPS (eGeMAPS) set of 88 parameters for every 0.5-second window. For text features, they provide third-party commissioned professional transcripts for the videos. The transcript was then aligned (every 5 seconds) and a 300-dimensional GloVe word embeddings (Pennington, Socher, and Manning 2014) was used. For the facial features, they provide 20 action points (R. Ekman 1997) extracted using the Emotient software by iMotions [1] for each frame (30 per second).

LIRIS-ACCEDE Dataset: Like mentioned in Table 4.1, we used the facial features, audio features, scene descriptors, and visual aesthetic features. While we used the already available features for audio and visual aesthetics, we extract the facial features and scene descriptors ourselves. The audio features provided were extracted using the openSMILE toolbox [2], which compute a $1,582$ dimensional feature vector. For the visual aesthetics, the authors provide the following: Auto Color Correlogram, Color and Edge Directivity Descriptor, Color Layout, Edge Histogram, Fuzzy Color, and Texture Histogram, Gabor, Joint descriptor joining CEDD and FCTH in one histogram, Scalable Color, Tamura, and Local Binary Patterns extracted using the LIRE [3] library. We extracted the face features

[1] https://imotions.com/emotient/
[2] http://audeering.com/technology/opensmile/
[3] http://www.lire-project.net/

ourselves using Bulat et al. (Bulat and Tzimiropoulos 2017). These result in 68 action units with the 3D coordinates. For the scene descriptors we use Xiao et al.'s (Xiao et al. 2018) 4096 dimensional intermediate representation.

Table 4.4: **Affect2MM Evaluation on SENDv1 Dataset:** for time-series emotion perception. We report CCC values on both the validation and test sets for comparisons with state-of-the-art methods.

Method	Modalities							
	FA		AT		FT		FAT	
	Val	Test	Val	Test	Val	Test	Val	Test
LSTM (Desmond Ong et al. 2019)	.100	.160	.800	.090	.280	.400	.140	.150
VRNN (Desmond Ong et al. 2019)	.140	.170	.320	.350	.240	.300	.170	.240
SFT (D. Ong et al. 2019)	.150	.160	.080	.080	.320	.350	.120	.140
MFT (D. Ong et al. 2019)	.060	.080	.360	.330	.400	.360	.420	.440
B3-MFN (D. Ong et al. 2019)	0.220	.090	.370	.330	.330	.310	.340	.280
Human (Desmond Ong et al. 2019)	-	-	-	-	-	-	.470	.500
Ours	.557	.582	.592	.601	.556	.567	.599	.597

Table 4.5: **Affect2MM evaluation on MovieGraphs Dataset:** for time-series emotion perception for time-series emotion perception. We report top-1 accuracy for comparisons with state-of-the-art methods.

Method	Validation (Top-1 Acc)	Test (Top-1 Acc)
EmotionNet (Wei et al. 2020)	35.60	27.90
Ours	39.88	30.58

MovieGraphs Dataset: For the MovieGraphs dataset as summarized in Table 4.1, we use all the features except the audio as the audios were not provided in the dataset. We now explain below how we retrieve the features. We extracted the face features ourselves using Bulat et al. (Bulat and Tzimiropoulos 2017). These result in 68 action units with the 3D coordinates. For the transcript, we used the 300-dimensional GloVe word embeddings (Pennington, Socher, and Manning 2014) to obtain the feature representation. For visual aesthetic features, we extracted various features for color, edges, boxes, and segments using Peng et al. (Peng and JEMMOTT III 2018). For the scene and situation descriptors, we used the provided text in the dataset and used the 300-dimensional GloVe word embeddings again to make them into feature representations.

4.4.3 Training Hyperparameters

All our results were generated on an NVIDIA GeForce GTX1080 Ti GPU. Hyper-parameters for our model were tuned on the validation set to find the best configurations.We used RMSprop for optimizing our models with a batch size of 1. We experimented with the range of our model's hyperparameters such as: number of hidden layers $(1, 2, 3)$, size of hidden layers (cLSTM, LSTM Decoder, Dense), dropout $\{0.2, 0.3, 0.4, 0.5.0.6\}$, hidden dimension $\{64, 128, 256, 512\}$, and embedding dimension $\{64, 128, 256, 512\}$. The ridge penalty (r) and non-smooth regularization parameter (l) of VAR training of the cLSTM was kept constant at $1e^{-4}$ and 0.001, respectively.The learning rate of both the tasks - emotion prediction and VAR were in this range - $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}\}$. More specific details on model-specific training hyperparameters are summarized in Table 4.3.

Table 4.6: **Affect2MM evaluation on LIRIS-ACCEDE (MediaEval2018 Dataset):** for time-series emotion perception. We report MSE for comparisons with state-of-the-art methods (lower the better).

Method	Valence (MSE)	Arousal (MSE)
CERTH-ITI (Batziou et al. 2018)	0.117	0.138
THUHCSI (Jin et al. 2017)	0.092	0.140
Quan et al. (Quan, V.-T. Nguyen, and Tran 2018)	0.115	0.171
Yi et al. (Y. Yi, H. Wang, and Q. Li 2018)	0.090	0.136
GLA (Sun, T. Liu, and Prasad 2019)	0.084	0.133
Ko et al. (Ko et al. 2018)	0.102	0.149
Zhao et al (Y. Zhao et al. 2019)	0.071	0.137
Ours	0.068	0.128

Table 4.7: **Ablation Experiments on Affect2MM:** We perform ablation experiments to understand the importance of co-attention and Granger Causality for modeling temporal causality.

Experiment	SENDv1	MG	LIRIS-ACCEDE	
	(CCC)	(Acc)	Valence (MSE)	Arousal Arousal(MSE)
Affect2MM w/o (co-attn.& GC)	.570	28.290	0.122	0.143
Affect2MM w/o GC	.585	29.450	0.095	0.135
Affect2MM	.597	30.580	0.068	0.128

4.5 Experiments and Results

We first discuss our quantitative results in Section 4.5.1, where we compare the performance of our method with SOTA methods on the three datasets. We then go over the ablation experiments performed in Section 4.5.2. Finally, in Section 4.5.3, we present some qualitative results for Affect2MM.

4.5.1 Quantitative Results

We compare Affect2MM with SOTA methods on the three datasets.

SENDv1 Dataset: For this dataset we summarise the results in Table 4.4. We provide CCC score for every combination of the three features/modalities used and also the combined results. Two prior works (D. C. Ong, Zaki, and Goodman 2019; D. Ong et al. 2019) have experimented with various network architectures and ideas and reported CCC values. We list and compare our performance from their best-performing methods in Table 4.4.

MovieGraphs Dataset: The dataset has not been previously tested for any similar task. We trained a recently proposed SOTA method, EmotionNet (Wei et al. 2020), for affective analysis of web images on the MovieGraphs dataset and compared our performance with this approach. We summarize this result in Table 4.5.

LIRIS-ACCEDE Dataset: For this dataset we summarise the results in Table 4.6. To be consistent with prior methods, we report Means Squared Error (MSE) and Pearsons' Correlation Coefficient (PCC) for our method. We compare against 7 existing SOTA methods evaluated on the same dataset. Some of these listed methods were a part of the MediaEval2018 Challenge.

4.5.2 Ablation Experiments

We perform a small ablation study to analyze the importance of each of the two components we incorporate for modeling the temporal causality in Affect2MM. We report

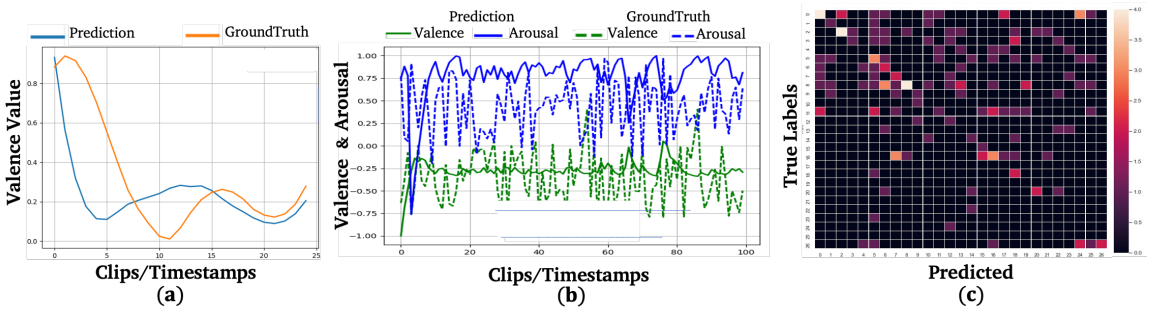


Figure 4.3: **Affect2MM's Qualitative Plot Analysis:** *(a)* We show the valence values learned for a single sample on the test set in the SENDv1 Dataset, and in *(b)*, we show both valence and arousal values from the LIRIS-ACCEDE Dataset. *(c)* We show the confusion matrix for the MovieGraphs dataset. Brighter colors on the diagonal indicate more correct classifications.

the performance without co-attention between the features and also without Granger causality. These results are summarized in Table 4.7. As reported, we see a performance improvement of about $4-5\%$ across all datasets and another $2-3\%$ with the addition of Granger causality.

4.5.3 Qualitative Results

Time-Series Emotion Predictions: We present qualitative results for Affect2MM in Figures 4.3 and 4.4. In Figure 4.3(a), we show the valence values (range $\in [0, 1]$) learned for a single sample from the test set for SENDv1 Dataset. In Figure 4.3(b), we show the predicted valence and arousal values for a sample from LIRIS-ACCEDE Dataset (range $\in [-1, 1]$) along with ground-truth labels. In Figure 4.3(c), we plot the 27×27 confusion matrix for all the test points in the MovieGraphs dataset. The horizontal axis represents the predicted class labels while the vertical axis represents the true labels. The discrete classes are in alphabetical order with 27th class as 'None' (Table 4.2). Brighter colors on the diagonal indicate more correct classifications.

Interpreting GC matrix: We visualize the Granger causality plots in Figure 4.4 (*(a), (b) and (c)*) for all the three datasets. The dimensions of the GC matrix is |features $\times$ features|. Hence, it is $|3 \times 3|$ for SENDv1 dataset, $|4 \times 4|$ for LIRIS-ACCEDE dataset and $|6 \times 6|$ for Moviegraphs dataset. To interpret the GC matrix, we read the features appearing on

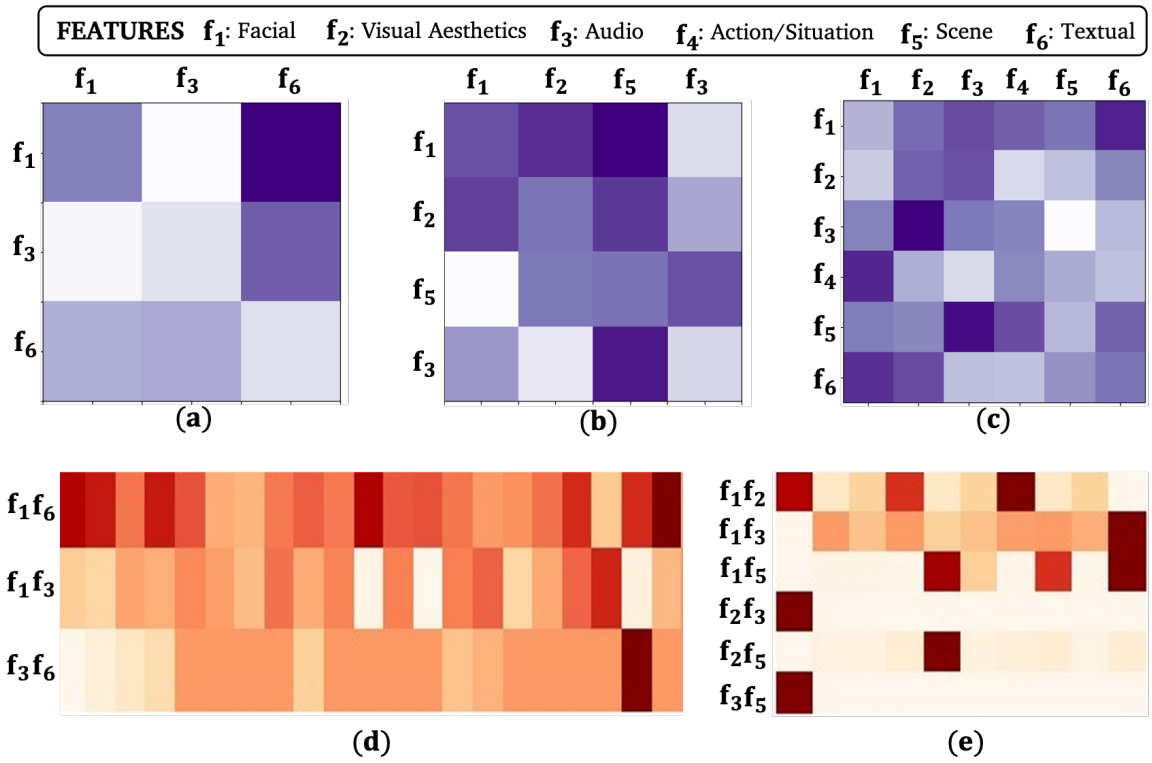


Figure 4.4: **Affect2MM Analysis (GC Matrices and Co-attention Weights):** *(Top)* We show the GC matrices for all three datasets (Desmond Ong et al. 2019; Baveye et al. 2015; Vicol et al. 2018). Darker shades indicate the existence of a Granger causality between corresponding features.*(Bottom)* We plot normalized co-attention weights for one sample each from the (d) SENDv1 and (e) LIRIS-ACCEDE datasets with time-stamps on the x-axis and pairs of features on the y-axis. Darker shades imply a higher co-attention between the pair of features and more relevance to predictions of the current and the following clips.

the horizontal axis and query whether they Granger-cause (or non-cause) the features on the vertical axis. Darker shades of purple indicate causality while lighter shades indicate non-causality. So for example, in (a) f_3 Granger-non-causes f_1 for SENDv1 dataset, and in (b), f_5 Granger-causes f_1 with a higher certainty.

Causality Qualitative Analysis: Here we further analyse co-attention and GC qualitatively. We present the analysis and reasoning of GC and co-attention for one such video from SENDv1 dataset (ID123vid1 video), for which the corresponding GC matrix and co-attention weights are shown in Figures 4.4(a) and (d).

Interpreting Co-attention Weights: We visualize the co-attention weights for one sample each from SENDv1 and LIRIS-ACCEDE in Figure 4.4 (*(d), (e)*). We plot the

clips/timestamps on the x-axis and pairs of all the features on the y-axis. For every clip, darker shades of red indicate a higher co-attention weight between a pair of modalities.

1. **Text GC Face:** In Figure 4.4(a), the text modality strongly Granger-causes the facial modality because at 2:20, when the subject says, "... he had a sailor's mouth ...", the subject then moments later motions to his face and changes his facial expression to mimic a sailor face.
2. **Text GC Speech:** In Figure 4.4(a), the text Granger-causes speech which is most clearly evidenced when at 2:05, the subject says, "... feel really lucky, really blessed ...", then subject then audibly sighs with relief.
3. **Speech correlation with text:** In Figure 4.4(d), we see a weak correlation between speech and text in the beginning because when the subject is remembering the good things about his grandfather, the text contains positive stories about the grandfather while the tone in his speech is still sad. However, the correlation between text and speech is stronger at the end of the video when the subject mentions the word, "death", and then takes a long pause signifying great sadness and loss.

4.6 Conclusion, Limitations and Future Work

We present Affect2MM, a learning method for time-series emotion prediction for multimedia content. We use Emotional Causality via co-attention and Granger causality for modeling temporal dependencies in the input data and building an affective-rich representation to understand and perceive the scene. We evaluated our method on three benchmark datasets and achieved state-of-the-art results.

There are some limitations to our approach. We need to make the feature extraction process online. Furthermore, our approach does not currently work with scenes containing single actors. In the future, we would like to extend our approach to solve Step 3 of the Emotional Causality Theory to predict the physiological response of the viewer to the movie. We would also like to explore the causality and correlation of modalities in single-

actor scenes to automate the analysis of debates and group discussions. Finally, we would like to incorporate the ideas presented in this paper with recommendation systems.

Chapter 5

Analyzing User Behavior on Social Media Platforms

The popularity of social media has increased dramatically in recent years. According to a 2021 survey by Pew Research Center (Center 2021), 81% of U.S. adults use social media, with younger people being the most frequent users. A survey conducted by Common Sense Media in 2020 (Media 2020) found that 84% of U.S. teenagers reported using social media, with over one-third using it almost constantly. Additionally, the American Academy of Pediatrics also recognizes the significance of social media in adolescent development and socialization, stating in a 2021 policy statement (Pediatrics 2021) that "the use of social media has become a central aspect of adolescent development."

Consequently, social media has become a rich source for researchers to address a variety of societal problems. Some of these research problems include understanding the spread of misinformation and fake news on social media platforms (Kaliyar, Goswami, and Narang 2021; Islam et al. 2020), flagging inappropriate content (Papadamou et al. 2020), and, detecting hateful and abusive content (Mozafari, Farahbakhsh, and Crespi 2020; Ribeiro et al. 2018). Social media has also been used to study social networks and relationships, including how they

influence health behaviors (Valente 2015), political opinions (Center 2018), and psychological well-being (Verduyn et al. 2017). While a lot of these problem statements focus on the *content* aspect of social media, given the increased popularity it is also important to understand the impact social media can have on the *users*.

Some interesting characteristics of social media usage is as follows:

(a) Users often have little control over the content they consume on social media feeds. Does this influence the behavior of users?

(b) All conversations on social media are $1 : n$, and so it is important to understand the impact of a such large number of $1 : n$ conversations.

Recent studies (n.d.) have exposed negative effects of social media due to users' lack of control over the content they consume. Users tend to post content that presents a false image of themselves to gain social validation (S. Zhao, Grasmuck, and J. Martin 2008; Walther et al. 2008; Jinyoung Kim and Ahn 2013). Users also often change their behavior on social media to portray a positive impression of themselves on these platforms (Sleeper et al. 2013; Rainie, Lenhart, and A. Smith 2012). All these cause issues such as body image concerns, anxiety, and mental health problems, particularly in teenagers, as a consequence of negative social comparison (Jang et al. 2016; Spitzer, Crosby, and Witte 2022). Additionally, sharing content on social media can elicit emotions that transfer to other users, leading to similar emotional experiences, particularly in image/video-based applications (Qian et al. 2013; Crandall and Snavely 2012), further damaging users' well-being.

To minimize the negative effects of such content, understanding creators' intent and educating their social network about this intent is crucial (Saldias and Rosalind W Picard 2019). To this end, various efforts have been made to comprehend the intent behind sharing multimodal social media content (Jia et al. 2021; Kruk et al. 2019; n.d.; Yen 2017; Xu et al. 2022a).

Understanding human intent behind multimodal social media content is challenging due to the lack of a standard intent taxonomy for this type of data. Prior works have pro-

posed diverse taxonomies and scrape images from various social media platforms (Twitter, Unsplash, Instagram, Weibo) (Jia et al. 2021; Kruk et al. 2019; D. Zhang et al. 2021; Xu et al. 2022b). Additionally, intent prediction goes beyond visual recognition tasks and requires understanding human cognition and behavior. Existing intent prediction models for such data are limited to image and text modalities.

A recent study led by Facebook (Kramer, Guillory, and Hancock 2014) highlighted one of the most subtle and least combated problems of digital content on social media–emotion contagion, which is defined as follows (Goldenberg and J. Gross 2019; Ferrara and Zeyao Yang 2015):

Emotion contagion *(EC) is a diffusion of emotions (positive or negative) and opinions over users in a social network such that the emotions and opinions of a "perceiver" become more similar to those of the "expressor" as a result of exposure to them.*

Emotion contagion can occur as a result of any type of exposure to the emotions of others. This can be broadly classified into non-digital (face-to-face or telephonic) and digital (social media) conversations. We now formally define Digital Emotion Contagion:

Digital Emotion Contagion *(DEC) is when EC occurs by sharing and expressing opinions on online platforms via multimodal digital content such as posts on Reddit and Facebook, tweets on Twitter, etc.*

While in both non-digital and digital emotion contagion, the emotions and opinions of "perceivers" change as a result of exposure to "expressors", the exposure is a lot more intense and frequent on digital media platforms as all interactions on social media platforms are $1:n$ opposed to $1:1$ conversations in the non-digital world. As already discussed, users have little control over the content they consume on online social media platforms, putting them at risk of this emotion contagion. Moreover, social media platforms are known to incentivize emotion-rich content, leading to a self-reinforcing loop of enhanced emotion contagion (Goldenberg and J. Gross 2019). Goldenberg et al. (Goldenberg and J. Gross 2019) note that because of this it is important to note that this is *mediated digital emotion*

contagion.

However, detecting the occurrence of contagion on social media platforms is challenging. Firstly, users can have similar emotional responses to a similar situation without any contagion, but differentiating such cases of similar emotional responses from contagion is hard. Moreover, to do so requires tracking user activity on social media platforms over a time span which is very sensitive data to publicly release. Additionally, while it is often possible to collect human-annotated data to build datasets for furthering research; it is not so straightforward here as by its very nature contagion is a hidden phenomenon. Consequently, prior work in emotion contagion research has been restricted to proving its existence on social media platforms (Ferrara and Zeyao Yang 2015; Kramer, Guillory, and Hancock 2014; Fan et al. 2014) like Facebook, Twitter, and Weibo. While conducting such controlled experiments has been controversial in the past as we are interfering with social media users, prior works have presented various hypotheses (R. Lin and Utz 2015; S. He et al. 2016; Coviello et al. 2014; Bhullar 2012; Gruzd, Doiron, and Mai 2011) about factors responsible for causing emotion contagion on social media. As suggested, researchers in the community agree that since detecting the occurrence of contagion as a research problem is in nascent stages; it's time to shift focus on understanding various factors that would help understand when contagion can be *stronger* or *weaker* on specific social media platforms.

Main Contributions: Towards this goal of understanding and making users emotionally aware of the content they consume on social media, we make the following contributions.

1. We propose a learning-based model called INTENT-O-METER (Figure 5.1(a)), a human intent prediction model for multimodal social media posts. In addition to visual (image) and textual (caption) features, INTENT-O-METER leverages Theory of Reasoned Action (TRA) factoring in (i) the creator's attitude towards sharing a post, and (ii) the social norm or perception towards the post in determining the creator's intention. We also integrate this intent prediction model into a web application interface (similar to Instagram) to understand users' feedback on the

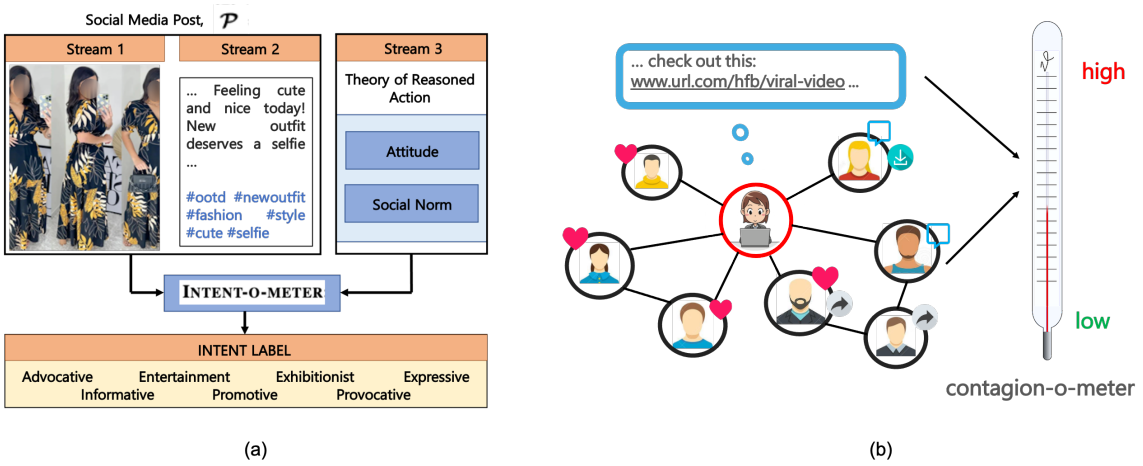


Figure 5.1: **Understanding User Behavior and Increasing Awareness about the Content Consumed on Social Media**: We study two characteristics of how social media can impact users. In (a), we present a model, INTENT-O-METER, a human intent prediction model for multimodal social media posts. In (b), we attempt to understand additional factors that could lead to stronger or weaker contagion on Instagram.

use of such intent labels along with social media posts. We try to understand if users will be open to trying this for their own awareness of social media content.

2. To understand the impact of contagion better, given a social media post *m* and the user *u*, we perform analysis to understand what factors could lead to stronger or weaker contagion on Instagram (Figure 5.1(b)).

The rest of the chapter is structured as follows: We discuss some prior work and discuss prior work in various social media research problems, intent recognition datasets, and digital emotion contagion in Section 5.1. In Section 5.2 we discuss our work in understanding intent behind the content the users share on social media platforms and INTENTGRAM dataset. In Section 5.3, we elaborate on our analysis and findings about factors that can lead to stronger or weaker contagion on Instagram. Finally we conclude with a discussion on some limitations of and future directions in Section 5.4.

Table 5.1: **Summary of Intent Taxonomy from Prior Literature:** We summarize the 7-label taxonomy we adopt for INTENTGRAM (borrowed from Kruk et al.) and the number of samples per label.

Label	# Samples	Interpretation
advocative	$9,293$	advocate for a figure, idea, movement.
entertainment	$8,938$	entertain using art, humor, memes etc.
exhibitionist	$5,327$	create a self-image reflecting the person.
expressive	$9,800$	express emotion at an external entity.
informative	$7,964$	information regarding a subject or event.
promotive	$4,661$	promote events, products, organizations.
provocative	$9,289$	directly attack an individual or group.
Total	$55,272$	

5.1 Prior Work in Understanding User Behavior on Social Media Platforms

In this section, we discuss previous works in related domains. To begin, we first go over the impact of social media on the mental well-being of users (Section 5.1.1). We then explain the various interpretations of the word "Intent" and the need to infer the intent of social media content in Section 5.1.2. Then in Section 5.1.3, we summarize various datasets and models that have been proposed in the recent past for inferring intent for social media content. We also provide an understanding of the Theory of Reasoned Action and our motivation for using this for our model in Section 5.1.4. In Section 5.1.5, we go over the theory of emotion contagion and then we specifically focus on digital emotion contagion (Section 5.1.6) and discuss the challenges of existing research directions on emotion contagion in social media.

5.1.1 Social Media's Impact on Mental Well-Being

Social media sites like Instagram, Facebook, and Twitter have become an important part of our daily lives, especially for young adults (Lakhiwal and Kar 2016; Van Dijck and Poell 2013). The pressure to publish "socially acceptable" and "socially likable" content often results in

a depiction of a false narrative on social media; more specifically image/video-based platforms like Instagram. Sophisticated editing tools and filters add to this false narrative. The impact of such content on young people is of grave concern. They often compare themselves to others (what they see) to assess their opinions and abilities, and such comparison has been known to lead to depression (Keles, McCrae, and Grealish 2020). Such comparisons can have a serious impact on physical and mental well-being. Young people also quantify their social acceptance in terms of the number of likes/comments/shares/follows (Tiggemann et al. 2018) which again traps them in a vicious circle.

5.1.2 Interpretations of 'Intent'

The term "intent" can have various meanings in different contexts, such as representing the next steps of an agent (Wooldridge and Jennings 1995; Bratman 1988) or actions (Ignat et al. 2021), emotions, and attitudes (Xu et al. 2022b). However, answering the question of why people post on social media platforms requires a specific intent taxonomy. Previous works (Purohit et al. 2015; Xu et al. 2022a; Kruk et al. 2019; Jia et al. 2021; D. Zhang et al. 2021) have proposed different taxonomies, but there is little consensus among them. We summarize the various datasets and taxonomies for intent prediction for social media data in Section 5.2.3.2. Social pressure to present likable and well-edited (filters) content can result in a false narrative on social media, which can affect young people's self-esteem and social acceptance. Educating young adults about the intent of content creators can help them be less vulnerable to what they see on social media.

5.1.3 Social Media Intent Recognition Models

While there aren't many intent prediction models for social media data, we briefly go over the ideas presented in some recent works. Kruk et al. (Kruk et al. 2019) and Zhang et al. (D. Zhang et al. 2021) use both visual (image) and textual (captions) modalities to predict an author's intent for their Instagram posts. Jia et. al. (Jia et al. 2021) focus more on predicting

intent labels based on the amount of object/context information and use hashtags as an auxiliary modality to help with the better intent prediction. The scope of these works is limited to just the visual and textual features of the data. Understanding human intent, however, is a psychological task (Talevich et al. 2017), extending beyond standard visual recognition. Therefore, we conjecture that additional cues from social media psychology literature are needed to improve the state-of-the-art in intent prediction.

Table 5.2: **Summary of Characteristics of SOTA Intent Prediction Datasets:** We compare INTENTGRAM Dataset with state-of-the-art intent prediction datasets. See Section 5.2.4.3 for a detailed discussion on a comparison between these datasets. I: image, V: video, C: caption, and H: hashtag. † Not Available Publicly.

Datasets	Features				#Labels	Size	Source
	I	V	C	H			
MDID (Kruk et al. 2019)	✓	✗	✓	✓	7	1,299	Instagram
Intentonomy (Jia et al. 2021)	✓	✗	✗	†✓	28	14,455	Unsplash
MET-Meme (Xu et al. 2022a)	✓	✗	✓	✗	5	10,045	Twitter, Weibo, Google, Baidu
†**Purohit et al.** (Purohit et al. 2015)	✗	✗	✓	✗	3	4,000	Twitter
†**MultiMET** (D. Zhang et al. 2021)	✓	✗	✓	✗	4	6,109	Twitter, Facebook
MIntRec (Hanlei Zhang et al. 2022)	✗	✓	✓	✗	20	2,224	TV Series
WHYACT (Ignat et al. 2021)	✗	✓	✓	✗	24	1,077	YouTube Videos
Intentgram	✓	✗	✓	✓	7	55,272	Instagram

5.1.4 Social Media and Theory of Reasoned Action

The Theory of Reasoned Action (TRA) (Fishbein and Ajzen 1977) assumes that people make rational choices when they engage in a specific behavior (*e.g. posting content on social media*), and that behavior is driven by *intentions*. Furthermore, TRA lays out the following two factors that determine *intention*: (i) attitude toward the behavior and (ii) the subjective norms associated with the behavior. Attitudes toward the behavior refer to the overall evaluations of the performance of the behavior in question, and subjective norms refer to perceived pressure or opinion from relevant social networks. Generally, individuals who have more favorable attitudes and perceive stronger subjective norms regarding behavior are more likely to show greater intentions to perform a behavior. Prior research (S. Kim, Joonghwa Lee, and D. Yoon 2015; X. Lin, Featherman, and Sarker 2013; Peslak, Ceccucci, and Sendall 2012) has used TRA to reason and develop an understanding of what

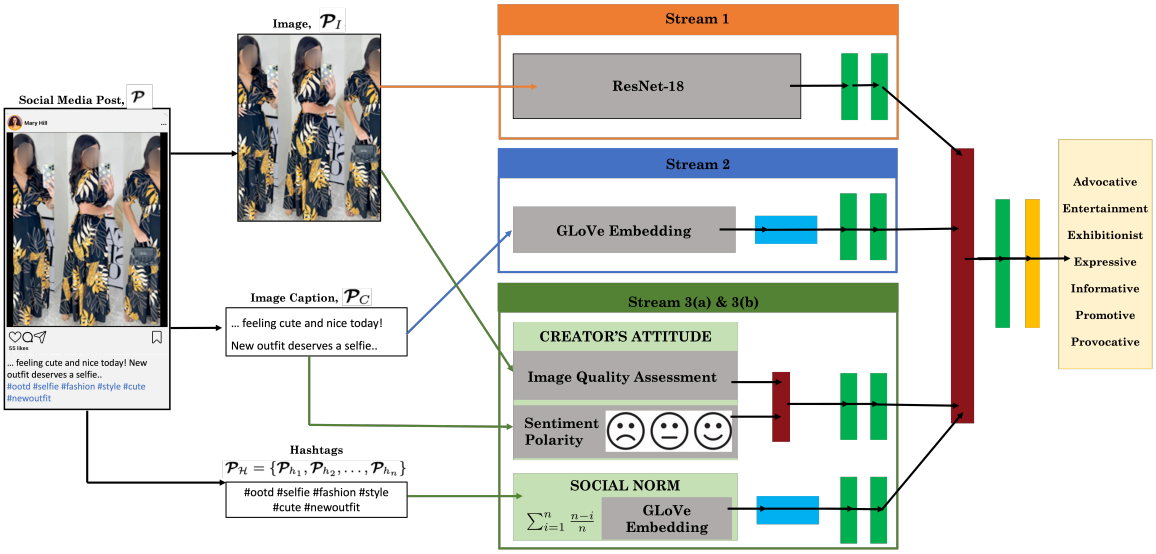


Figure 5.2: **Intent-o-meter Model Architecture**: Given as input a social media post, $\mathcal{P} = \{\mathcal{P}_I, \mathcal{P}_C, \mathcal{P}_{\mathcal{H}}\}$, which has three components (an image, $\mathcal{P}_I$, with an associated caption, $\mathcal{P}_C$, and a set of hashtags, $\mathcal{P}_{\mathcal{H}} = \{\mathcal{P}_{h_1}, \mathcal{P}_{h_2}, \ldots, \mathcal{P}_{h_n}\}$), our goal is to predict the *intent label* for $\mathcal{P}$. INTENT-O-METER has three streams. In the first stream (orange), we encode the visual features of the image, in the second stream (blue) we encode the captions, and finally, in the third stream (green) we model the Theory of Reasoned Action; both *attitude of the author/creator* and the *social norm of the kind of post*, $\mathbf{P}$. We then fuse the three streams (dark red) to make the final intent prediction. The networks consist of fully-connected layers (light green), an LSTM layer (blue), a concatenation operation (dark red), and a softmax layer (yellow).

motivates social media users to share information online. They confirm that TRA can be used as a model for social networking behavior. They also find that both intention and subjective norm are positively associated with the intention to use social media (Tarkiainen and Sundqvist 2005; H.-K. Bang et al. 2000). While these studies, however, confirm TRA and its role in modeling user intent on social media, no work so far uses TRA to *predict* user intent

5.1.5 Theory of Emotion Contagion

Prior works have suggested that humans instinctively tend to align with the emotional states they perceive around them (P. Ekman, Levenson, and Friesen 1983; Hatfield, Cacioppo, and Rapson 1992; Barger and Grandey 2006). Various studies have concluded that emotions can be contagious (Schachter and JE Singer 1962), as a response to which individuals show behav-

ioral, attentional, and emotional synchrony (Hatfield, Cacioppo, and Rapson 1992). Prior literature has also associated emotion contagion with feelings of empathy and sympathy (Hatfield, Cacioppo, and Rapson 1992; Dellarocas, Xiaoquan Zhang, and Awad 2007) and emotional arousal (Mehrabian and Russell 1974b; Russell 2003; Mehrabian 1980). The study of emotional contagion has been the focus of various disciplines because different types of interactions, such as commercial transactions, team communication, and human–robot interactions, can transfer emotions (Jingjing Li, Jian Zhang, and Zhiguo Yang 2017; Q. Chen et al. 2019; Kuang et al. 2019; Manera, Grandi, and Colle 2013; Matsui and Yamada 2019). Marketing research on emotional contagion has focused on understanding how positive or negative emotions converge in positive or negative consumer behavior (Dellarocas, Xiaoquan Zhang, and Awad 2007; Kramer, Guillory, and Hancock 2014; Fox et al. 2018; Cowley 2014). More recently, emotion contagion through social media has been of heightened interest because of the high engagement on these platforms.

5.1.6 Digital Emotion Contagion

Most prior works (Kramer, Guillory, and Hancock 2014; Ferrara and Zeyao Yang 2015; Fan et al. 2014; Coviello et al. 2014) have conducted controlled experiments on social media platforms and confirmed the presence of emotional contagion and its manipulative effects on individuals. Similarly, (Tromholt 2016) and (Hunt et al. 2018) show that the content we consume on social media affects not only the emotions that we express on these platforms but also our general well-being. As discussed in prior literature (Goldenberg and J. Gross 2019), contagion can occur due to three mechanisms: (i) mimicry, (ii) activation, and (iii) social appraisal. More specifically, digital media platforms are known to incentivize competition for attention and positive reinforcement in the forms of likes or shares (Brady, Gantman, and Van Bavel 2020; Brady, Crockett, and Van Bavel 2020), and expressing emotions is an extremely useful way to attract attention. As a result, such emotion-rich digital activities lead to self-reinforcing loops that enhance emotion contagion over time. (Saldias and Rosalind W Picard

2019) developed *Tweet Moodifier*, a Google Chrome extension that enables Twitter users to filter and visually mark emotional content in their Twitter feed to make them aware of and reflect on the emotion-rich content being consumed.

5.2 Intent-o-meter: Understanding Users Intent to Share Online

In this section, we go over the proposed intent prediction model, INTENT-O-METER for multimodal social media posts. We formulate the problem in Section 5.2.1. We then explain the model architecture in Section 5.2.2. We introduce our dataset, INTENT-GRAM, the taxonomy, and data collection process and compare it with the state-of-the-art datasets intent prediction datasets in Section 5.2.3. In Section 5.2.4, we go over the quantitative experiments and ablation experiments to evaluate INTENT-O-METER on state-of-the-art methods and datasets. We explain the user evaluation we perform to further evaluate in Section 5.2.5.

5.2.1 Problem Formulation

Problem 5.2.1. *Given as input a social media post, $\mathcal{P} = \{\mathcal{P}_I, \mathcal{P}_C, \mathcal{P}_\mathcal{H}\}$, which has three components: an image, $\mathcal{P}_I$, with an associated caption, $\mathcal{P}_C$, and a set of hashtags, $\mathcal{P}_\mathcal{H} = \{\mathcal{P}_{h_1}, \mathcal{P}_{h_2}, \ldots, \mathcal{P}_{h_n}\}$, our goal is to predict the intent label for $\mathcal{P}$.*

We present an overview of our intent prediction model, INTENT-O-METER, in Figure 5.2. As our input is multimodal, we refer to multimodal deep learning literature and extract both the visual features from the input image $\mathcal{P}_I$ as well as the textual features from the associated caption. For the former, we use a state-of-the-art visual feature extraction backbone network, the ResNet architecture family (Section 5.2.2.1) while for the latter, we leverage the GLoVe word embeddings with a recurrent neural network (Section 5.2.2.2). In addition, we also extract features that model the Theory of Reasoned

Action; the *attitude of the creator* and the *social norm of the kind of post*, $\mathcal{P}$ (Section 5.2.2.3). We concatenate the three features in late fusion to make the final intent prediction (Section 5.2.2.4).

5.2.2 Our Approach

In the following sections, we describe each component in more detail.

5.2.2.1 Stream 1: Visual Modality

The dominant modality for such social media platforms is often the visual modality, i.e., images and videos. To be consistent with prior work, we use the ResNet-18 network pre-trained on the ImageNet dataset (Jia Deng et al. 2009) to encode the visual features (K. He, Xiangyu Zhang, et al. 2016). We use the output of the second-to-last layer for the image representation ($\mathbb{R}^{N\times 512}$). To fine-tune this, we then add two trainable fully-connected layers (ϕ) with ReLU non-linearity and 0.5 dropout, to finally get f_{VISUAL}.

$$f_{\text{VISUAL}} = \mathcal{S}_1\Big(\text{RESNET}_{18}(\mathcal{P}_I)\Big) \tag{5.1}$$

Table 5.3: **Ablation Experiments on Intentgram Dataset (Benefit of TRA in Intent Prediction):** We highlight the importance of using TRA in addition to visual and textual features by ablating INTENT-O-METERand analyzing each component in isolation. – indicates the absence of hashtag information in the dataset. (u) indicates uniform weighting for hashtag embeddings. *stream* 1: visual, *stream* 2: textual, *streams* 3*(a)* and 3*(b)*: TRA.

Dataset	Metric	Experiments (Streams)			
		1 + 2	1 + 2 + 3(a)	1 + 2 + 3(b)	1 + 2 + 3(a) + 3(b) INTENT-O-METER
Intentonomy	F1	32.72	40.68	-	**40.68**
MET-Meme	F1	38.89	47.74	-	**47.74**
MDID	Acc.	54.29	55.58	57.12/55.92(u)	**58.20**
Intentgram	Acc.	50.21	52.36	53.73/51.23(u)	**54.01**
	AUC	73.58	76.86	75.51/74.87(u)	**79.48**

5.2.2.2 Stream 2: Textual Modality

Prior work in multimodal learning shows that visual information is often not enough to recognize human intent (Mittal, Bhattacharya, et al. 2020b; A. B. Zadeh et al. 2018). We use the user-generated captions, $\mathcal{P}_C$, of the images as a complementary cue. To encode these captions we leverage pre-trained GLoVe word embeddings (Pennington, Socher, and Manning 2014) to encode caption words in 50 dimensions. We use an LSTM layer, followed by two fully connected layers (ϕ) with ReLu non-linearity and 0.5 dropout to get f_{TEXTUAL}.

$$f_{\text{TEXTUAL}} = \mathcal{S}_2\Big(\text{LSTM}(\text{GLoVe}(\mathcal{P}_C))\Big) \tag{5.2}$$

5.2.2.3 Stream 3: Modeling TRA

As discussed in Section 5.1.4, according to the Theory of Reasoned Action (TRA), individuals who have more favorable attitudes and perceive stronger subjective norms regarding a behavior (in this case, posting particular content) are more likely to show greater intentions to execute that behavior. Many studies (S. Kim, Joonghwa Lee, and D. Yoon 2015; X. Lin, Featherman, and Sarker 2013; Peslak, Ceccucci, and Sendall 2012) have validated the influence of TRA on users while posting content on social media, but no method exists that computationally models both these components from a post, $\mathcal{P}$. We describe this below.

Stream 3(a) Attitude: In TRA, a user's attitude indicates how strongly the creator believes in the post they are sharing online. Since "belief" in a post is subjective, we refer to social media psychology literature where studies have correlated engagement and frequency with social media use and in particular, one such study (Scott et al. 2017) states, *"highly engaged youth participated on social media platforms often and in diverse ways: messaging friends, reacting to and circulating others' posted content, and generating their own."* We model such engagement in two ways. The first is via caption sentiments. Kruk et al. (Kruk et al. 2019) show that two different captions for the same Instagram image can completely change the overall meaning of the image-caption pair.

With this intuition, we compute the polarity of the sentiments expressed in the captions. We use the VADER (Hutto and Gilbert 2014) library to compute these features.

$$f_{\text{SENTIMENT}} = \text{VADER}(\mathcal{P}_C) \tag{5.3}$$

The second way in which we model user engagement and frequency on social media is via the editing and filters applied on the images before they are posted on the various social media platforms. Doing so is reflective of the time spent by the creator in preparing the post and indicative of the attitude the creator has towards the image they are sharing. To help our model learn this, we compute k image quality or visual aesthetic features, $q_1, q_2, \ldots, q_k$ which correspond to image blurriness, dominant color, brightness, contrast, exposure, average color, and so on.

$$f_{\text{QUALITY}} = \text{IMAGE\_QUALITY}(\mathcal{P}_I) = \left[q_1, q_2 \ldots q_k\right]^\top \tag{5.4}$$

We concatenate the features and use fully connected layers and non-linearity to compute f_{3_a}.

$$f_{\text{ATTITUDE}} = \mathcal{S}_{3_a}\left(\left[f_{\text{SENTIMENT}}; f_{\text{QUALITY}}\right]^\top\right) \tag{5.5}$$

Stream 3(b) Social Norm: The goal here is to understand how well the content posted is perceived socially. In the absence of any formal definition of social perception of content, we use hashtags, associated with a post $\mathcal{P}$, as a proxy for social acceptance of the post. Hashtags for social media posts are used to categorize content to make it more visible. Moreover, prior work (Martín, Lavesson, and Doroud 2016; X. Chen et al. 2020) has shown that hashtags are directly correlated to growing one's social network and expanding their audience. We assume that the most influential hashtags appear first in the set of available hashtags, $\mathcal{P}_\mathcal{H} = \{\mathcal{P}_{h_1}, \mathcal{P}_{h_2}, \ldots, \mathcal{P}_{h_n}\}$. This is a reasonable assumption due to the auto-suggest feature in most devices. Assuming a linear piece-wise weighting scheme, with a

weight of $\frac{n-1}{n}$, for the hashtags, we use pre-trained GLoVe word embeddings (Pennington, Socher, and Manning 2014) to encode the words as $50-$dimensional features. We use an LSTM layer, followed by two fully connected layers with non-linearity and dropout to get f_{SOCIAL}.

$$f_{\text{SOCIAL}} = S_{3_b}\left(\sum_{i=1}^{n} \frac{n-i}{n} \mathcal{P}_{h_i}\right) \tag{5.6}$$

We conclude this section by emphasizing that our current TRA model, based on caption sentiments, image aesthetics, and hashtag embeddings, is heuristic and may be one of several possible way alternatively modeling TRA. It should, accordingly, not be presumed as a gold standard way of computationally modeling TRA–that remains an open research question–and we hope this work is a stepping stone towards further research in this area.

5.2.2.4 Fusion: Inferring the Intent Label

To fuse the 4 features/encodings we have computed, $f_{\text{VISUAL}}, f_{\text{TEXTUAL}}, f_{\text{ATTITUDE}}$, and f_{SOCIAL} from the three streams, we concatenate these features before making any individual intent inferences.

$$\begin{aligned} f_{\text{CONCAT}} &= \left[f_{\text{VISUAL}}, f_{\text{TEXTUAL}}, f_{\text{ATTITUDE}}, f_{\text{SOCIAL}}\right]^{\top} \\ f_{\text{FUSE}} &= \mathcal{S}_{\text{FUSE}}\left(f_{\text{CONCAT}}\right) \end{aligned} \tag{5.7}$$

We use two fully-connected layers followed by a softmax layer. This output is used for computing the loss and back-propagating the error back to the network.

5.2.3 Introducing Intentgram Dataset

We present our intent taxonomy and data collection procedure for INTENTGRAM in Section 5.2.3.1 followed by a comparison with other social media intent datasets in Sec-

tion 5.2.3.2. We depict samples in INTENTGRAM dataset corresponding to one of the 7 intent classes in Figure 5.4.

Table 5.4: **Intent-o-meter's Evaluation on the MDID Dataset:** We summarize the experiment results on MDID dataset here. We report top-1 accuracy and AUC score for comparisons. There are a total of 7 intent labels.

Method	top-1 Accuracy	AUC
Random	28.10	50.00
Gonzaga et al. (Gonzaga, Murrugarra-Llerena, and Marcacini 2021)	54.50	84.40
Kruk et al. (Kruk et al. 2019)	56.70	85.60
Intent-o-meter	**58.20**	**89.70**

Table 5.5: **Intent-o-meter's Evaluation on the MET-Meme Dataset:** We summarize the experiments for MET-Meme dataset here. We report top-1 accuracy and AUC score for comparisons. There are a total of 7 intent labels.

Method	Validation Micro F1	Test Micro F1
Random	23.20	22.32
Kruk et al. (Kruk et al. 2019)	36.36	38.89
Xu et al. (Xu et al. 2022a)	37.64	41.65
Intent-o-meter	**41.33**	**47.74**

Table 5.6: **Intent-o-meter's Evaluation on the Intentonomy dataset:** We present experiments for intent prediction on the Intentonomy dataset. We report Micro F1 Score and Macro F1 Scores for comparisons. There are a total of 28 intent labels.

Method	Micro F1	Macro F1
Random	7.18	6.94
Kruk et al. (Kruk et al. 2019)	32.72	28.57
Jia et al. (Jia et al. 2021)	38.49	31.12
Intent-o-meter	**40.68**	**34.71**

5.2.3.1 Taxonomy, Collection and Pre-processing

7-label Taxonomy: We follow the intent taxonomy used by Kruk et al. (Kruk et al. 2019), as they also define the labels on Instagram data. We summarize this further in Table 5.1.

Table 5.7: **Intent-o-meter's Evaluation on our Proposed Dataset, Intentgram:** We summarize evaluations on INTENTGRAM here. We report accuracy, AUC scores and Micro F1 Score for comparisons. There are a total of 7 intent labels.

Method	top-1 Accuracy	AUC	Micro F1
Random	28.10	50.00	–
Kruk et al. (Kruk et al. 2019)	50.21	73.58	49.15
Intent-o-meter	**54.01**	**79.48**	**53.54**

Scraping Instagram Posts: We used the Apify scraper to collect Instagram posts from publicly available profiles, similar to Kruk et al. (Kruk et al. 2019). As a first step, we begin by scraping Instagram posts belonging to the 7 categories (Table 5.1) using hashtags provided by Kruk et al. We initially collected and clustered a large number of Instagram content to understand and identify popular hashtags.

Based on the frequency of usage, we choose top-10 hashtags for each of the intent labels (added these hashtags in Table 5.9).

Dataset Pre-processing: With an aim to curate a large-scale collection of publicly available Instagram posts we scrape 2000 samples for all the hashtags under consideration. Thus after the initial phase, we end up getting $1,40,000$ posts in total. The Apify platform provides a mirror of the original Instagram posts (viable only for a short time) to download them. We then apply pre-processing and cleaning as described below to get the final dataset consisting of $55,272$ posts. For fair evaluation, we restrict ourselves to a total of $10,053$ samples (equally distributed across all 7 categories) for the purpose of training, validation, and testing. We will release the entire dataset to facilitate further research by the community.

Intentgram Cleaning and Processing: We list down the various hashtags used to scrape the public posts in Table 5.9. Because this is a dataset scraped from Instagram, it was quite noisy and required cleaning. We mainly filter the data points in the following aspects-

1. **Duplicate Removal:** We observed that during the scraping process a lot of

duplication posts got scrapped due to the modus operandi of the Apify platform. The first thing was to remove the duplicate posts.

Another interesting observation made was many of the scraped posts had 2 or more of the hashtags under consideration thereby getting scraped more than once. For the purpose of this study, we restrict ourselves to considering posts belonging to a single category. Hence the duplicates were ignored resulting in a further reduction in sample size.

2. **Language:** In this work, we limited ourselves to posts with English captions only and therefore, discarded posts written in other languages.
3. **Non-Textual Characters:** We clean the captions of emoticons, special characters, unnecessary punctuation marks, etc.
4. **Multimodal Posts:** Since we are developing a multimodal intent prediction model, we also remove posts without hashtags and captions.

Table 5.8: **Summary and Characteristics of Intentgram Dataset Statistics:** We summarize some insights from our dataset, INTENTGRAM; average number of hashtags, average number of likes and the average length of caption for Instagram posts per Intent class label.

	Avg No. of Hashtags	Avg No. of Likes	Avg Caption Length
Advocative	16.71	101	198
Entertainment	19.29	156	151
Exhibitionist	16.07	53	122
Expressive	15.22	34	163
Informative	17.08	45	215
Promotive	14.33	331	199
Provocative	15.58	174	306

Detailed Analysis of Intentgram Dataset: We summarize some insights of our dataset, INTENTGRAM in Table 5.8. We observe that the number of hashtags used is more or less consistent across all the 7 categories. However, the interesting thing to note is the average number of likes in the *Promotive* class is significantly more than in the other classes. This might be attributed to the fact that a lot of people tend to get more influenced by such content over social media. Similarly, the average caption length is

considerably higher in the Provocative class. This might be due to the fact that creators use more textual content in their posts to make their case even more strong.

Explanation of Intent Taxonomies: As mentioned in Section 5.1.2, there is no consensus on the intent taxonomy for social media posts. We summarize the various taxonomies that have been used for annotating social media posts with intent labels in the recent past in Table 5.10. We list the various datasets and their source of social media posts too. Our decision to stick with the 7-label intent taxonomy as proposed by Kruk et al. was driven by the fact that their source of social media posts was similar to INTENTGRAM's source, Instagram.

Table 5.9: **Building and Scraping Intentgram Dataset (Hashtags Used):** We summarize the hashtags used to scrape Instagram posts for the 7 Intent labels.

Intent Label	Hashtags Used to Scrape Instagram Posts
Advocative	#pride, #maga, #gay, #trump, #lgbt, #love, #usa, #freedom, #insta-gay, #conservative
Entertainment	#meme, #earthporn, #fatalframes, #earthpix, #wanderlust, #nature, #earthfocus, #naturelovers, #naturegram, #traveldiaries
Exhibitionist	#selfie, #ootd, #fashion, #style, #picoftheday, #beautiful, #cute, #photography, #follow, #instalike
Expressive	#lovehim, #merrychristmas, #christmas, #happy, #christmastree, #christmasdecor, #christmastime, #xmas, #winter, #photooftheday
Informative	#news, #Noticias, #hiphop, #technology, #instadaily, #podcast, #reels, #viral, #Business
Promotive	#ad, #NYCC22, #funkogram, #funkocollector, #Collectible, #FunkoNews, #Funkos, #Loungefly, #FPN #FunkoPOP
Provocative	#antifa, #redpill, #eattherich, #socialism, #antifascist, #anticapitalism, #anticapitalist, #capitalismkills, #antiracist

Table 5.10: **Summary of SOTA Social Media Intent Prediction Datasets:** We summarize the various social media intent taxonomies proposed in the recent past.

Dataset	# Intent labels	Labels
MDID (Kruk et al. 2019)	7	advocative, entertainment, exhibitionist, expressive, informative, promotive, provocative
MET-Meme (Xu et al. 2022a)	5	entertaining, expressive, interactive, offensive, other
Purohit et al. (Purohit et al. 2015)	3	seeking, offering, none
MultiMET (D. Zhang et al. 2021)	4	persuasive, descriptive, expressive, others
multirow1*NYT Survey (n.d.)	5	entertaining, self-fulfillment, promotive, grow relationships, define ourselves
(Jia et al. 2021) **Intentonomy**	28	Attractive, BeatCompete, Communicative, CreativeUnique, CuriousAdventurousExcitingLife, EasyLife, EnjoyLife, FineDesignLearnArt-Arch, FineDesignLearnArt-Art, FineDesignLearnArt-Culture, GoodParentEmocloseChild, Happy, HardWorking, Harmony, Health, InLove, InLoveAnimal, InspirOthers, ManagableMakePlan, NatBeauty, PassionAbSmthing, Playful, ShareFeelings, SocialLifeFriendship, SuccInOccupHavGdJob, TeachOthers, ThngsInOrdr, WorkILike
Intentgram	7	advocative, entertainment, exhibitionist, expressive, informative, promotive, provocative

Dataset Statistics: We also collect relevant metadata for each post such as caption, hashtags, number of likes, and number of comments. Due to privacy concerns, we release only the ResNet-18 features of the images in Instagram posts.

5.2.3.2 Comparing Intentgram with SOTA Datasets

Table 5.2 compares our proposed dataset, INTENTGRAM, with state-of-the-art intent classification datasets. INTENTGRAM uses the 7-label taxonomy (*advocative, entertainment, exhibitionist, expressive, informative, promotive, provocative*) borrowed from MDID dataset, which is based on Goffman and Hogan's prior work (Goffman 2021; Hogan 2010) for Instagram data. INTENTGRAM is the most diverse in terms of available modalities and features consisting of images, captions, and hashtags. The MDID dataset (Kruk et al. 2019) also uses Instagram as the source data but is $40\times$ smaller than INTENTGRAM. In fact, INTENTGRAM is the largest dataset containing approximately $55K$ data points. Finally, we note that while the MDID, Intentonomy, MET-Mete, MultiMET and the dataset proposed by Purohit et al. are specifically intended for intent classification and social media analysis, the MIntRec and the WHYACT are in fact action prediction datasets.

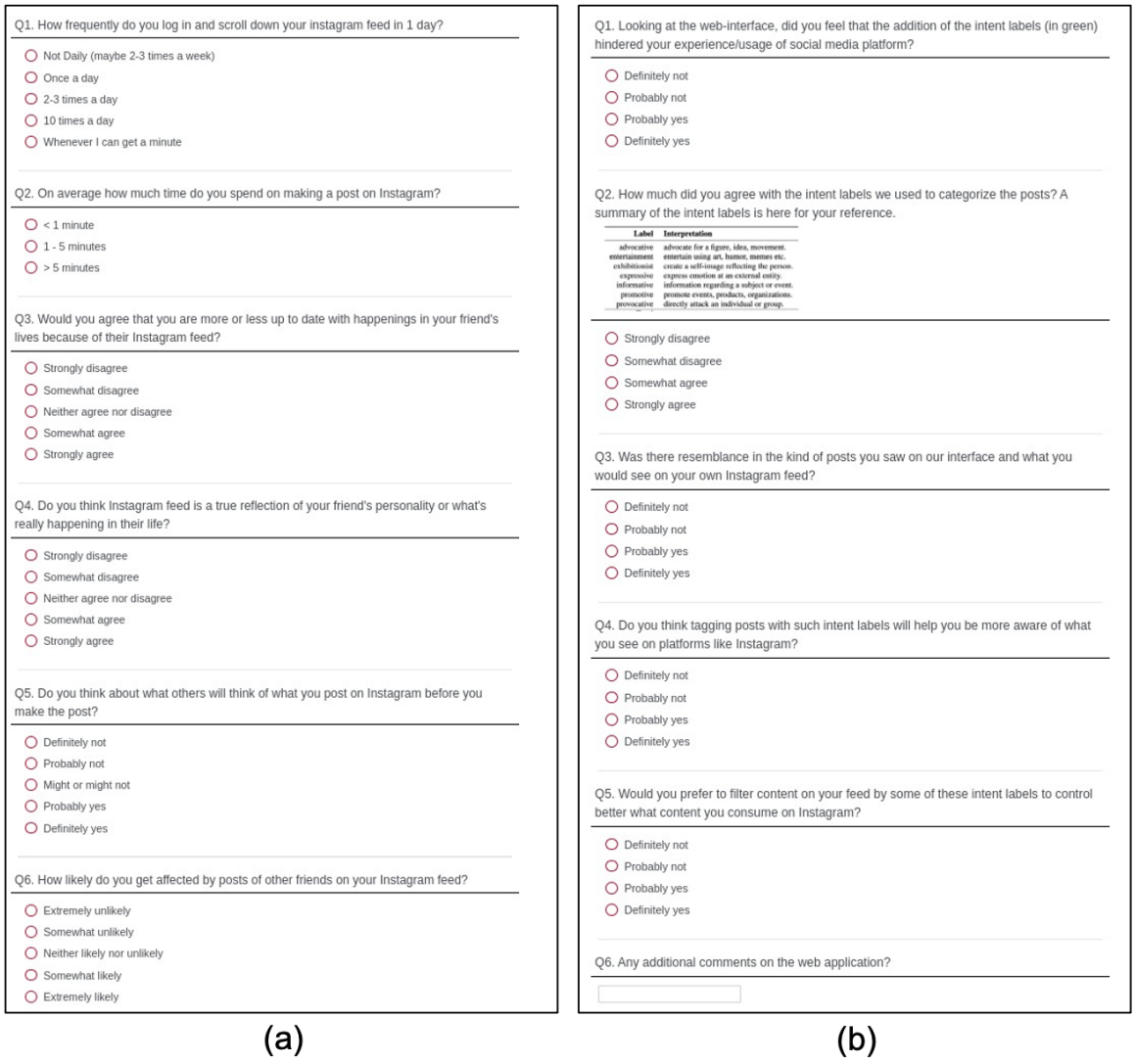


Figure 5.3: **Screenshots of the two Questionnaires Used for Evaluating Intent-o-meter Model and Intentgram Dataset**: We show a screenshot of the two questionnaires used as a part of the userstudy. In (a), we show the 6 questions participants are asked before they see the web interface. In (b), we show a screenshot of the 6 questions participants are asked after they see the web interface.

5.2.4 Experiments and Results

Our experiments answer the following two questions: (i) Does modeling TRA result in better intent prediction in social media posts? and (ii) How does INTENT-O-METER compare to state-of-the-art (SOTA) methods?

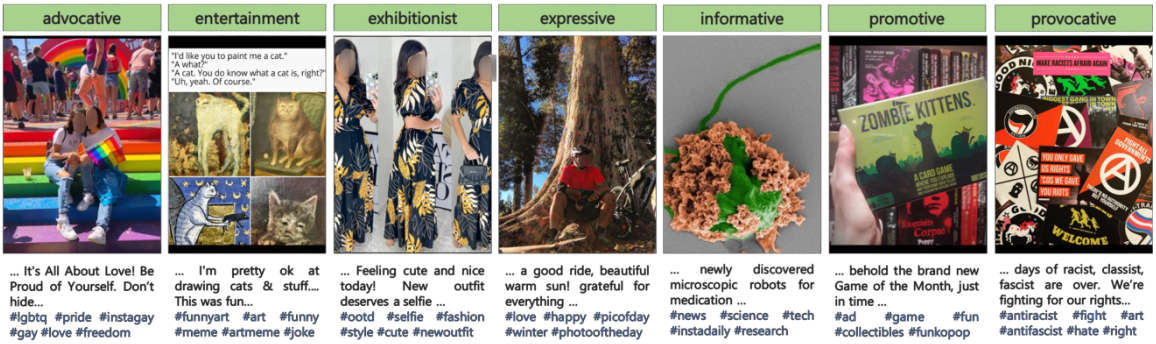


Figure 5.4: **IntentgramDatasets's Qualitative Examples:** We depict a single Instagram post from each of the 7 intent labels from our dataset along with a portion of the caption and a subset of the hashtags used with the post. For privacy concerns, we will blur the human faces.

5.2.4.1 Experimental Setup

Dataset Splits: We use four intent prediction datasets: Intentonomy (Jia et al. 2021), MDID (Kruk et al. 2019), and MET-Meme (Xu et al. 2022a), and INTENTGRAM. We used the original splits provided by the authors for Intentonomy, MDID, and MET-Meme datasets. For the purpose of experiments, we sample $10,053$ posts from INTENTGRAM ($1,443$, $1,154$, $1,415$, $1,576$, $1,475$, $1,420$, and, $1,570$ posts respectively for the 7 intent label) and we split training, validation, and testing sets in the ratio $60:20:20$, resulting in $6,031$, $2,011$, and $2,011$ samples for train, validation, and test sets, respectively.

Evaluation Metrics: Different datasets have used different metrics for evaluation. The Intentonomy dataset uses Micro F1 score and Macro F1 score. Similarly, MDID reports accuracy and AUC metric. For the MET-Meme dataset, we have reported and compared against both validation and test F1 scores. For our dataset, INTENTGRAM we report Accuracy, AUC metric, and Micro-F1 score.

Training Details: All our results were generated on an NVIDIA GeForce GTX1080 Ti GPU. Hyper-parameters for our model were tuned on the validation set to find the best configurations. We used Adam optimizer for optimizing our models with a batch size of 50. We experimented with the range of our model's hyperparameters such as: dropout $\{0.2, 0.3, 0.4, 0.5, 0.6\}$, learning rate $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$, number of epochs

$\{50, 75, 100, 125\}$, and the hidden dimension of LSTM layers $\{32, 24, 16\}$.

5.2.4.2 Benefits of TRA in Intent Prediction

In Table 5.3, we highlight the benefit of modeling TRA, in addition to leveraging the visual and textual features obtained from images, captions, and hashtags. Specifically, we ablate INTENT-O-METER on all four datasets and report the F1 score, accuracy, and the AUC. In particular, we compare the results in the first column ("$1 + 2$") with the last column ("INTENT-O-METER"). Our results show that leveraging TRA improves the F1 score by 7.96% and 8.85% on the Intentonomy and MET-Meme, results in higher accuracy by 4% each on MDID and INTENTGRAM, and increases AUC by 5.9 points on INTENTGRAM.

We also perform additional tests where we individually analyze the individual effect of embedding the caption sentiments and image aesthetics as well as associated hashtags. In particular, the column under ("$1+2+3(a)$") highlights the benefit of modeling caption sentiment. And in the ("$1 + 2 + 3(b)$") column, we analyze Equation 5.6 by comparing linear piece-wise weighting with uniform weighting with each weight set to 1, and conclude that weighting, in some form, is better. Future work involves exploring more sophisticated weighting schemes including transformer-based attention.

In addition to the above ablation experiment, we can also draw further evidence for TRA from our experiments comparing INTENT-O-METER with state-of-the-art intent prediction methods that solely rely on visual and textual features, which we describe below.

5.2.4.3 Comparing Intent-o-meter with SOTA

We summarize our comparisons with SOTA methods on the MDID (Table 5.4), Intentonomy (Table 5.6), MET-Meme (Table 5.5), and our dataset INTENTGRAM (Table 5.7).

Performance on MDID dataset: We compare against the prediction model proposed

by Kruk et al. (Kruk et al. 2019)[1] and Gonzaga et al. (Gonzaga, Murrugarra-Llerena, and Marcacini 2021). While Kruk et al. propose the use of image and captions for predicting intent labels, Gonzaga et al. create a transductive graph learning method. We observe that our model outperforms these methods by up to 3.7% in top-1 accuracy and 5.3 AUC points.

Performance on Intentonomy dataset: We compare against the prediction model proposed by Jia et al. (Jia et al. 2021) who propose the use of hashtags as an auxiliary modality for predicting intent labels. We observe that our model outperforms their method by up to 3.59% in F1 score.

Performance on MET-Meme dataset: We compare against the baseline prediction model proposed by Xu et al. (Xu et al. 2022a) who only use image modality to predict intent labels and Kruk et al.(Kruk et al. 2019). We observe that our model outperforms these methods by up to 6.9% in F1 score.

Performance on our dataset, INTENTGRAM: We compare against the intent prediction model proposed by Kruk et al. We observe that our model outperforms these methods by 4% in top-1 accuracy and F1, as well as by 6 AUC points.

Conflating our results obtained from the ablation experiment in the previous section with our comparison results with SOTA methods that do not use TRA on 4 standard datasets, we find strong evidence that modeling TRA significantly improves intent prediction in terms of F1 score, top-1 accuracy, and AUC.

[1] Code replicated by us due to unavailability.

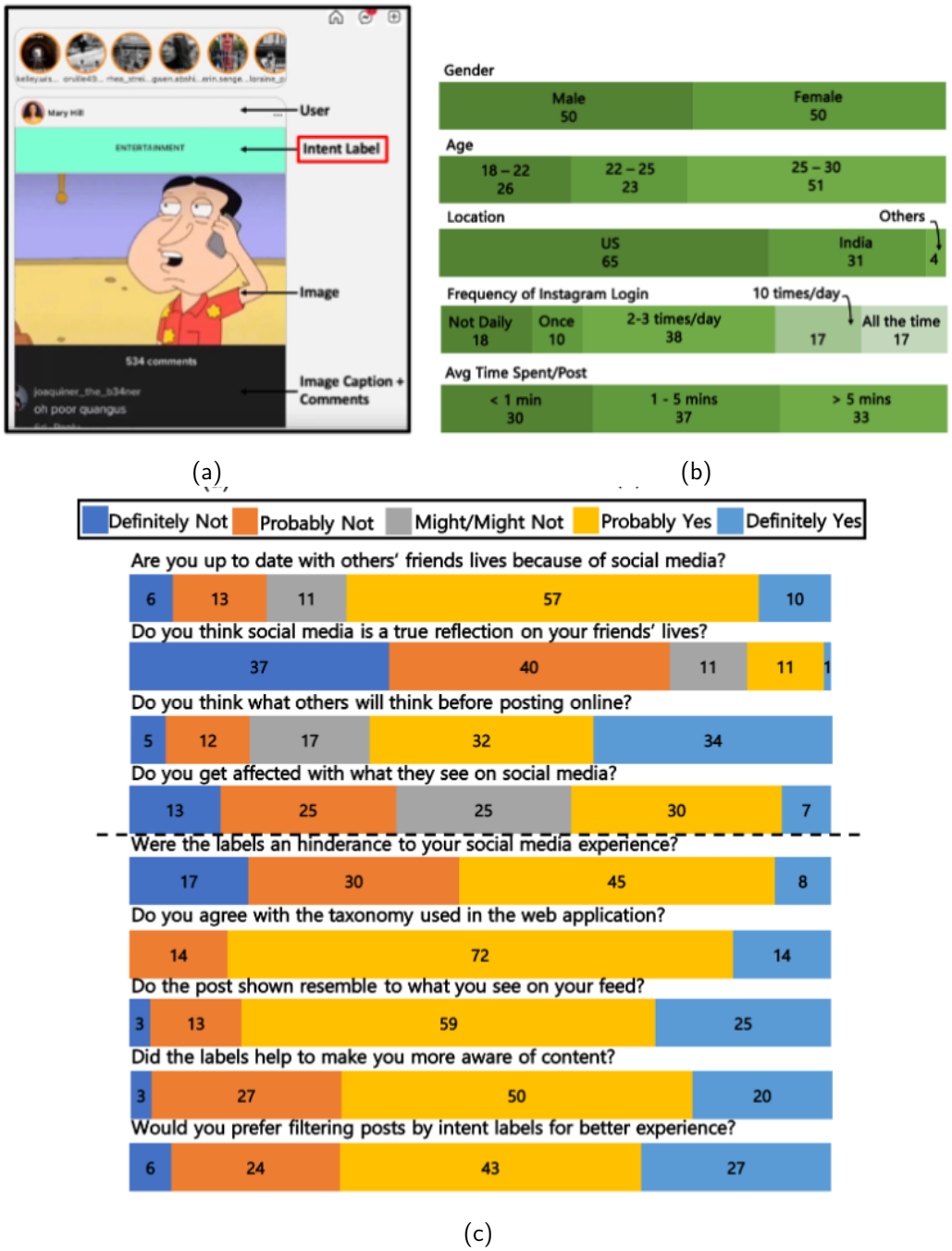


Figure 5.5: **User Study Setup and Analysis Used for Evaluating Intentgram Dataset and Intent-o-meter Model**: We summarize our user study setup and findings here. In (a), we show a screenshot with various components highlights, in (b) we report the background of the 100 participants recruited for the user study and, finally in (c) we report the answers to the questions of the pre-questionnaire and post questionnaire.

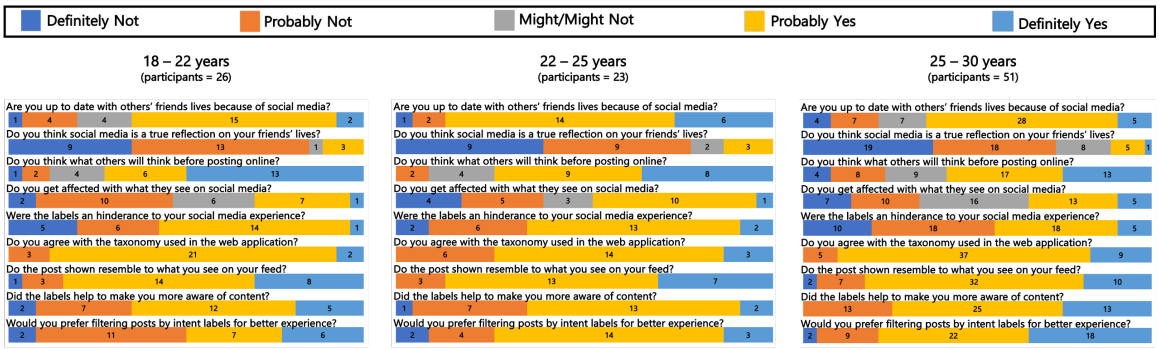


Figure 5.6: **Analyzing User Study Responses Based on Age Groups:** We analyze the responses for the three age groups, $18-22$ *years*, $22-25$ *years*, and $25-30$ *years*.

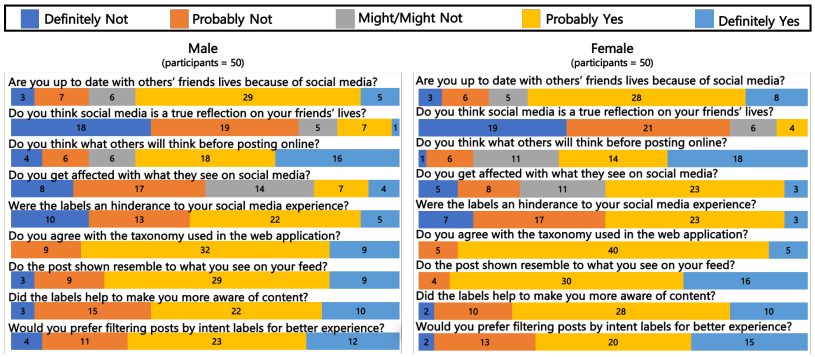


Figure 5.7: **Analyzing User Study Responses Based on Gender:** We analyze the responses for the 50 *male participants* and 50 *female participants* separately.

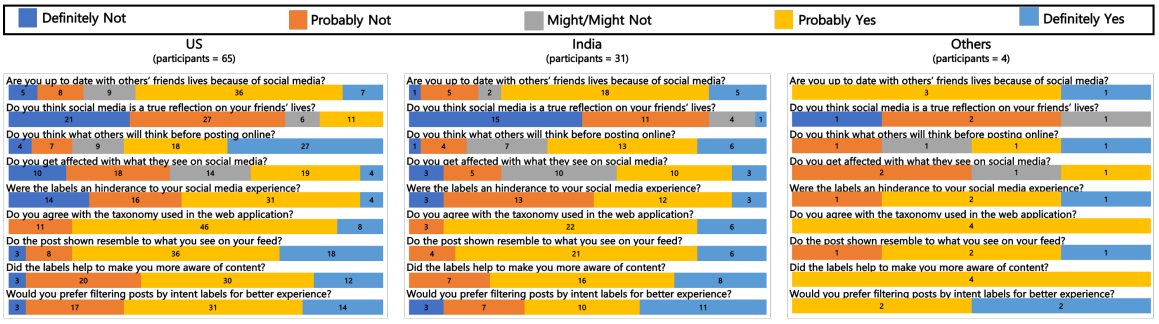


Figure 5.8: **Analyzing User Study Responses Based on Location:** We analyze the responses of the users based on their residing location. We have 65 users based in *US*, 31 based in *India*, and 4 from the *rest of the world*.

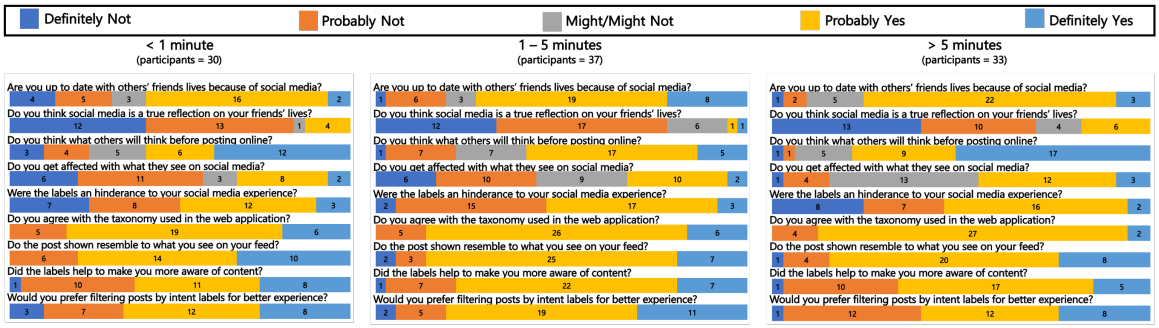


Figure 5.9: **Analyzing User Study Responses Based on Average Time Taken to make a Post:** We analyze the responses of the users based on the time they spend to make a post on Instagram. We group the 100 users into three groups, ≤ 1 minute, $1-5$ minutes, and ≥ 5 minutes.

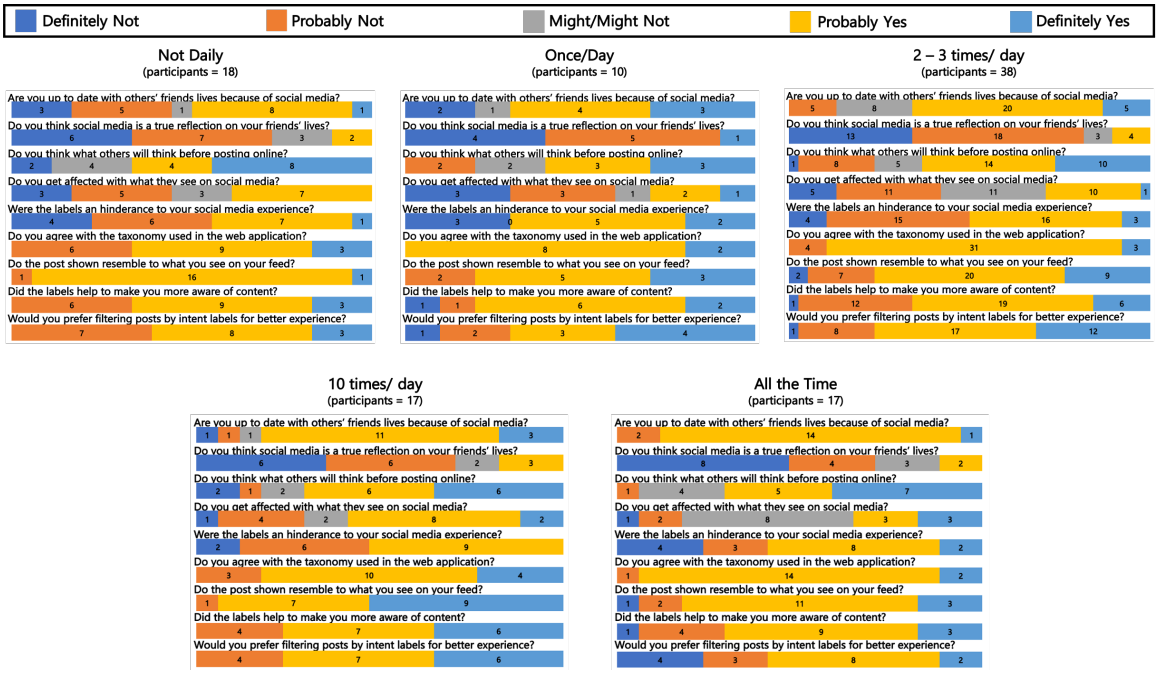


Figure 5.10: **Analyzing User Study Responses Based on Frequency of Logins to Instagram:** We analyze the responses of the users based on the time they spend to make a post on Instagram. We group the 100 users into five groups; *not daily*, *once per day*, $2-3$ *times a day*, 10 *times per day*, and *all the time*.

5.2.5 Understanding Human Preference

To understand human preferences and their reaction to intent labels displayed alongside social media posts, we conducted a user study, similar to T-Moodifier (Saldias and Rosalind W Picard 2019), to answer two questions: (i) do intent labels on posts make users more

aware of the content they consume? and (ii) would they prefer to have their content filtered by such labels? We describe the user study setup in Section 5.2.5.1 and analyze the results of the study in Section 5.2.5.2.

5.2.5.1 User Study Setup

The study consists of a web application where users interact with an "Instagram-like" interface in which the posts are taken from INTENTGRAM. For each post, users also see an intent label for that post (highlighted in green on top in Figure 5.5(a)). We instruct participants to scroll through the feed for $5 - 10$ minutes to experience the interface.

Prior to interacting with the interface, we ensure that (a) participants are between the ages of 18 and 30 and (b) they sign a consent form. In addition, we request them to answer a pre-study questionnaire which consists of 6 questions (Appendix Figure 5.3(a)) based on their current usage of Instagram. We also provide a screen recording of our web application to the users in case they have issues accessing the web application[2]. Finally, after the task, we ask participants to answer a post-study questionnaire, that consists of another 6 questions to collect their feedback on our web application (Appendix Figure 5.3(b)).

The aim of building this application was to make active social media users aware of an experience they could have if they had access to information about the intent behind the posts available on their social media news feed. We also wished to understand how receptive users would be to such a design. The main challenge here was to build a web application that closely resembles a platform that most participants of the study would be familiar with. Therefore, we chose to build from scratch, a web application that resembles the popular social media platform Instagram as much as possible. The UI of the web application was built entirely using React JS. For the posts, we used images, captions, and hashtags from our dataset INTENTGRAM. The web application was hosted using

[2]Web Application screen recording shown to participants.

firebase[3]. Firebase provided us with the feature of having a real-time database. It is a cloud-hosted database wherein data was stored as JSON. Any new updates made either to the data or web design were conveyed to the users instantly.

5.2.5.2 User Study Analysis

We recruit 100 participants for our user study (50 identify as female and 50 as male). We summarise statistics about the participants age and geographical locations in Figure 5.5(b) (*rows 2,3*). We also gather information about their amount of usage of social media application, Instagram. In Figure 5.5(b) (*row 4*), we report the frequency of social media logins and in Figure 5.5(b) (*row 5*), we record the average time taken to publish a post by participants.

In addition to statistics about the participants, we also gather information about the role of social media in their lives. In Figure 5.5(c), 67% lean towards believing they are up to date with their friends lives because of social media and 77% participants also believe that social media is not a true reflection of their friends' lives. Similarly, 37% participants report getting affected by what they see online, while 25% unsure if they are getting affected. As a testimony to our web interface, roughly half participants, 53% reported that the display of the intent labels was not a hindrance to their social media application experience; 86% participants seem to in agreement with the taxonomy of intent labels used to tag posts; and 84% participants also report a resemblance to the posts shown and the posts they see on their own personal social media feeds. And finally, 70% participants reported both that the intent labels helped them become aware of the content they are consuming on social media and that they would prefer filtering the content based on such intent labels.

We also asked participants for optional suggestions, comments, and feedback on the web application. A common theme among the suggestions was the presentation of the

[3]https://firebase.google.com

intent labels. One participant suggested color-coding intent labels, and another suggested making intent labels optional and letting users control if they would want to view posts with labels or without labels. Some participants appreciated the green highlighting that distinguished the labels whereas others mentioned preferring a more subtle appearance *e.g* in a corner in a smaller font. Another suggestion was to provide a feedback mechanism for users to report an incorrect intent label, and one participant suggested extending this to a multi-label classification as some posts seemed relevant for multiple labels.

We filtered user study responses by age, gender, location, frequency of use, and frequency of posting (Figures 5.6-5.9). Although most trends seem consistent across the various parameters, we make some interesting observations. First, the percentage of people open to filtering social media content by intent labels increases as people grow older. The question asking if users are affected by social media revealed that females are more affected by other users' posts than men. Perhaps the most interesting observation is that the same percentage of participants worry about what others will think in each posting frequency category.

5.3 What causes Emotion Contagion on Instagram?

In this section, we go over the analysis performed to discover new factors that can help reason digital emotion contagion on Instagram. We formulate the problem in Section 5.3.1. We then explain the dataset used and the experimental setup in Section 5.3.2. We then explain the analysis and insights in Section 5.3.3 and Section 5.3.4.

5.3.1 Problem Formulation

In a social media network of N users, $u^1, u^2, \ldots, u^N$, suppose the i^{th} user makes a post p^i. Suppose further that j attributes associated with each user are denoted by $a_1, a_2, \ldots, a_j$ and k attributes associated with each post are denoted by $b_1, b_2, \ldots, b_k$. Let $u^i(a_m)$ represent the m^{th} attribute value of u^i. Similarly, $p^i(b_n)$ represent the n^{th} attribute value

Table 5.11: **Summary of Factors Causing Emotion Contagion:** We summarize factors suggested by prior literature that are known to cause stronger emotion contagion on various social media platforms.

Aspect	Factors	References
Homophily	Connection Strength	(R. Lin and Utz 2015)
	Age, gender, demographics	(S. He et al. 2016)
Causality	Time Gap b/w Content Consumed and Action Taken	(Coviello et al. 2014)
Interference	Sentiment	(Coviello et al. 2014; Bhullar 2012; Gruzd, Doiron, and Mai 2011)

of p^i.

The probability that p^i causes emotion contagion is $P\left(\zeta|\left(u^i, p^i\right)\right)$. However, computing this probability is intractable; as prior research is only limited to proving that contagion exists. There have been various hypotheses that reason few user profile attributes, $u^i(a_m)$ and post attributes $p^i(b_s)$ that tend to cause *stronger* or a *weaker* contagion. Prominent studies (Goldenberg and J. Gross 2019; Ferrara and Zeyao Yang 2015) have propounded factors that can indicate strong or weak emotion contagion on online digital platforms. These factors are also summarized in Table 5.11. Various studies suggest content high in emotional content is more prone to cause more contagion than other content (Coviello et al. 2014; Gruzd, Doiron, and Mai 2011). It has also been shown that stronger ties between the expression and perceiver lead to stronger contagion (R. Lin and Utz 2015). The strength of ties on social media networks can be computed by either the number of mutual connections, reciprocity between these connections, and also the "follower ratio", i.e. the ratio of the users' followers vs. who they follow on the social media platform. On the other hand, perceivers' personalities (M. Cao et al. 2017) (easily influenced/agreeable), their online activities (Del Vicario et al. 2016), and their demographic features like age, gender, and culture (S. He et al. 2016) have proven to influence the degree of emotion contagion online. Therefore, based on some of these user profile attributes, $u^i(a_m)$ and post attributes $p^i(b_s)$, we are able to model the relation $P\left(\zeta|\left(u^i(a_m), p^i(b_n)\right)\right) \succeq P\left(\zeta|\left(u^j(a_m), p^j(b_n)\right)\right)$, is typically determined, which suggests that the strength of contagion caused by the u^i users' p^i post will be more than u^j users' p^j post. Here, $x \succeq y$ represents strength in a qualitative sense

and not in a quantitative sense.

Problem 5.3.1. *We want to build on the factors listed in Table 5.11 and find more user profile attributes* $(a_1, \ldots a_M)$ *and post attributes* $(b_1, \ldots b_N)$*, that would lead to stronger or weaker contagion.*

5.3.2 Dataset and Experiment Analysis Setup

We sample $2,000$ data points from INTENTGRAM to study Instagram-specific user attributes and profile attributes to identify new factors that lead to contagion on Instagram, a popular social media platform. As discussed in Section 5.2.3.1, INTENTGRAM is a collection of Instagram posts scraped using the Apify API. For each of these data points, we have the corresponding information:

- ***userId:*** This is the identification number associated with the Instagram user.
- ***postImage:*** The image associated with the Instagram post.
- ***hashtags:*** The list of the hashtags used along the image posted.
- ***caption:*** The caption used along with the image posted.
- ***# comments:*** This is the number of comments the post got on Instagram.
- ***# likes:*** This is the number of likes the post got on Instagram.
- ***intentCategory:*** Each of the datapoint is associated with one of the 7 intent labels (*advocative*, *entertainment*, *exhibitionist*, *expressive*, *informative*, *promotive*, *provocative*).

Additionally, we use Instagrapi library [4] to further retrieve the following information corresponding to each of the data points:

[4] https://pypi.org/project/instagrapi/

- ***followerCount:*** The number of users who follow the user who made the corresponding Instagram post.
- ***followingCount:*** The number of users the user corresponding to the Instagram post follows.
- ***# posts/year:*** This signifies a user activity; and is the number of posts the user makes on average in a year.
- ***isVerified:*** Is a boolean variable indicating if the account is verified or not by Instagram.
- ***isPrivate:*** Another boolean variable indicating if the user account is private or public.

Furthermore, we compute two more additional pieces of information for each of the data points:

- ***# human subjects:*** We used a face detection algorithm (Jiankang Deng et al. 2020) to detect the number of human subjects in the image posted by the user.
- ***Edits/Filters Used:*** We obtain a metric to associate the quality of the image posted (C. Chen and Mo 2022) as a proxy for measuring the number of edits or filter used in the image.

Table 5.12: **Summary of Correlation Values for Reasoning Contagion:** We summarize the Pearson correlation values between our hypothesized Instagram-specific user attributes ($a_2 - a_4$) and post attributes ($b_2 - b_8$) and already established factors that cause contagion (a_1 *and* b_1). Significant correlations are shown in bold for $p < 0.05$. Also, for features of nominal values, we use a variation of Pearson Correlation, a.k.a Point Biserial Correlation Test.

	FFR (a_1)	Sentiment Polarity (b_1)
User Factors ($a_1 \dots a_M$)		
account usage (a_2)	**0.264**	-0.027
verified account? (a_3)	0.008	**0.192**
private account? (a_4)	**-0.161**	0.050
Post Factors ($b_1 \dots b_N$)		
# comments (b_2)	0.091	0.029
# likes (b_3)	-0.093	**0.186**
# subjects (b_4)	**0.194**	-0.084
# hashtags (b_5)	-0.013	-0.044
caption length (b_6)	0.134	0.097
edits/filters (b_7)	**0.203**	**0.364**
intent label (b_8)	**0.288**	**0.303**

5.3.3 User Profiles and Instagram Posts and Contagion Correlation

Based on the features described above we now explain the first analysis we performed. As already discussed, our goal is to discover newer user-specific attributes ($a_1 \dots a_M$) and post-specific attributes ($b_1 \dots b_N$) that can reason digital emotion contagion on social media platforms. If we had access to such 'contagion' labels (ζ), our experiments and analysis would have been a lot more direct; making it possible to even develop learning models to infer the level of contagion of such Instagram data (more generally, social media data). However, considering we do not have access to the contagion values, ζ, we proxy it with our existing knowledge about factors causing contagion (summarized in Table 5.11). The first user attribute (a_1), reflected in Table 5.11 (row 1), the strength of ties between the expressor and the perceiver can affect the level of contagion. Prior empirical studies have confirmed that the stronger the connection; the more will the contagion be. While there is more than one way to compute this strength of connection in social networks, we

model it as the "follower ratio" (FFR), which is defined as follows/

$$\text{FFR} = \frac{\text{followerCount}}{\text{followingCount}}$$

Similarly, the first post-specific attribute (b_1), reflected in Table 5.11 (last row), is the sentiment polarity of the content being shared. It has been empirically shown that content that is more susceptible to invoking stronger emotions in the perceiver leads to more contagion. To this end, we compute the sentiment polarity of the corresponding content (Hutto and Gilbert 2014).

Our Hypotheses: Now, given access to a_1 and b_1, which act as a proxy to contagion ζ values, we hypothesize that the following user-specific factors ($a_2 - a_4$) and the following post-specific factors ($b_2 - b_8$) could be useful in reasoning contagion too.

- a_2, ***user account activity:*** We believe how frequently a user shares posts on their account could indicate how much contagion their posts will cause.
- a_3, ***verified account:*** We believe the fact that the user account is verified or not can also be an important factor to consider when understanding contagion.
- a_4, ***private/public account:*** Similar to the previous point, whether the user account from which the post is made is public or private can also be an important indicator.
- b_2, ***# comments:*** We believe an important factor could be how many comments an Instagram post receives.
- b_3, ***# likes:*** And, similarly there could be a correlation between the number of likes an Instagram post has and the contagion.
- b_4, ***# human subjects:*** An interesting factor to study is the number of human subjects in the image.

- b_5, ***# hashtags used:*** We also believe the number of hashtags users use with their posts could have some correlation with contagion.

- b_6, ***length of caption:*** More expressive long captions rather than short captions could also be a determinant factor causing contagion.

- b_7, ***edited/filtered Images*** The number of edits or filters used before the image is posted, in other words, the time spent on creating a post can have a relation to the contagion caused by the post.

- b_8, ***intent labels:*** Finally given the digital emotion contagion we are studying is 'mediated', it becomes important to understand the users' intent behind sharing a post. Here, we want to understand if certain intent tends to be more contagious.

Study Analysis: We study the correlation between a_1, b_1 with the target attributes, $a_2 - a_4, b_2 - b_8$. We summarize the Pearson correlation values in Table 5.12 where we highlight both strong and weak correlations and mark the statistically significant results in bold. We discuss some of these interesting relationships below.

We have discovered a positive correlation between user activity and a higher "follower ratio" ($p = -0.152$). This implies that users who are more active tend to have more followers compared to the number of users they follow. This finding can be rationalized by considering social media accounts of public figures, where their higher level of activity and engagement with their followers may result in a higher follower ratio. Furthermore, we have observed that verified accounts tend to have content that is more polar in sentiment ($p = -0.224$). Similarly, public accounts have a higher follower-to-following ratio (FFR) compared to private accounts. These results indicate that verified accounts, which are often associated with public figures or influencers, tend to generate more polarized sentiment in their content. Moreover, public accounts, which have a wider reach, tend to have a higher FFR compared to private accounts, suggesting differences in follower engagement and interaction based on account visibility settings.

We observed weak correlations between a_1 and b_1 with several key factors, such as the number of comments (b_2), hashtags (b_5), and caption length (b_6). Interestingly, there is a positive correlation between edited/filtered images (b_7) and both a_1 and b_1, with p-values of -0.153 and -0.246 respectively. This can be attributed to the findings of our proposed intent prediction model, which revealed that image quality features play a significant role. Furthermore, we have observed statistically significant positive correlations between a_1, b_1, and a_8, i.e., intent labels, with p-values of 0.003 and -0.193 respectively. We attribute this to the strong association between specific intent labels and a_1 and b_1. The positive correlation with sentiment polarity is also in line with our previous findings, which showed that incorporating emotions enhances the accuracy of intent prediction models.

In summary: our analysis has revealed interesting correlations between user activity, account verification status, sentiment polarity, and follower engagement on social media. Specifically, we found that more active users tend to have a higher follower ratio, which could be explained by the engagement dynamics of public figures. Verified accounts tend to generate more polarized sentiment in their content, and public accounts tend to have a higher follower-to-following ratio compared to private accounts, indicating differences in follower engagement based on account visibility settings.

Additionally, our findings suggest that image quality features play a significant role in the correlations between user activity and intent labels, as well as sentiment polarity. Given these positive correlations, we believe that some of these attributes are essential and can be used as determinants for contagion on social media platforms.

Table 5.13: **Analysis of Contagion:** We use thresholding on a_1 and b_1 to compute proxy ζ values. We then analyze b_4 and b_8 on a per-class basis.

			Contagion (ζ)		
			Strong	**Medium**	**Weak**
# Subjects	(b_4)	0	88	135	211
		1	483	169	105
		$2-3$	281	306	96
		>3	54	49	23
Intent Labels	(b_8)	Advocative	51	109	131
		Entertainment	123	71	54
		Exhibitionist	148	94	70
		Expressive	135	89	62
		Informative	85	139	49
		Promotive	77	113	69
		Provocative	211	93	27

5.3.4 More Deeper Analysis

In the previous section, we analyzed the correlations between a_1 and b_1, with the other factors that we hypothesized could be also used as determinants of emotion contagion on social media platforms. In other words, we used a_1 and b_1 as proxy markers of contagion. In this section, we go one step ahead and threshold values of a_1 and b_1 to create three discrete contagion, ζ values. To create these ζ values, we do the following-

- $\zeta =$ **strong contagion**: $0.7 < a_1 \leq 1.3$ and $b_1 > 0.7$
- $\zeta =$ **medium contagion**: $0.4 < a_1 \leq 0.7$ or $1.3 < a_1 \leq 1.6$ and $0.4 < b_1 \leq 0.7$
- $\zeta =$ **weak contagion**: $a_1 \leq 0.4$ or $1.3 \geq a_1$ and $0.4 \leq b_1$

We present a comprehensive analysis of two crucial factors, namely b_4 and b_8, in Table 5.13, summarizing our nuanced insights. Our findings reveal compelling patterns in the contagion dynamics of social media posts.

First, we observe that posts with one or more human subjects in the content exhibit a robust contagion effect, ranging from strong to medium levels. Remarkably, 63.8% of

posts containing only one subject and 42.5% of posts containing more than three subjects demonstrate strong contagion. Furthermore, posts with two subjects exhibit both strong (41.1%) and medium (44.8%) contagion. Intriguingly, posts without any subjects display the weakest contagion effect, with 43.8% indicating contagion. These findings challenge the conventional belief that posts with more subjects would trigger stronger contagion, suggesting that the presence of fewer human subjects in a post may actually lead to more intense contagion.

Second, we delve into the role of intent labels in contagion dynamics. As expected, posts labeled as "Provocative" and "Entertainment" show the highest potential for contagion, with 63.7% and 49.5% of posts demonstrating strong contagion, respectively. Notably, posts labeled as "Exhibitionist" and "Expressive" also exhibit strong contagion, albeit to a lesser degree. In contrast, posts labeled as "Advocative", "Informative", and "Promotive" demonstrate weak to medium contagion. These findings highlight that an "X factor" element, such as being provocative or entertaining, is essential to trigger contagion, whereas posts aimed at imparting information may not have the intended effect.

In summary, our analysis reveals intriguing insights into the contagion dynamics of social media posts, shedding light on the impact of factors such as the number of human subjects and intent labels. These findings contribute to a deeper understanding of the complex mechanisms underlying social media contagion and have potential implications for various domains, including social media marketing, public health communication, and online information dissemination.

5.4 Conclusion, Limitations, and, Future Work

We proposed INTENT-O-METER, an intent prediction model for social media posts using visual and textual modalities, along with the Theory of Reasoned Action. We evaluated our model on the Intentonomy, MDID, and MET-Meme datasets. We introduced INTENTGRAM, a dataset of $55K$ social media posts scraped from public Instagram profiles.

We also developed a web application with intent labels displayed on the posts and test it with existing Instagram users. While the literature on digital emotion contagion is quite in its nascent stages, we perform statistical analysis and present new insights. We show that attributes like the quality of the image posted on Instagram, and the intent label with which a post was made by the expressor have positive correlations with other factors that have been known to lead to a stronger contagion. Moreover, as shown by prior studies, we confirm that quantitative metrics like the number of comments and likes to a post do not necessarily impact how contagious the content can be.

We recognize that the Theory of Reasoned Action (TRA) is not the only method for modeling psychological and cognitive cues in social media posts. Employing other theories, such as the Theory of Planned Behavior (Ajzen 1991), could aid in comprehending human intent. Furthermore, we aim to enhance existing features by identifying additional ones. For example, we plan to create improved user profiles, analyze a user's social network, and assess their social media activity to better capture a person's motives. Our study on users shows that tagging posts with intent tags improves awareness of consumed content, and users are willing to try filtering content based on the tags. Nevertheless, the extent to which this tagging can alleviate the negative impacts of social media remains an unsolved research question. To further investigate this, we intend to conduct more user experiments in our future work. We also think that our approach can be applied to other social media platforms, which could offer new feature sources and improved feature design.

The lack of datasets and user-annotated labels for digital emotion contagion are definitely a roadblock; as also pointed by prior literature. However, like suggested we explored and presented interesting insights specifically for Instagram data that could potentially lead to higher contagion. To further our understanding, we hope to see how generic features like this are cross social media platforms and also further brainstorm for more such attributes.

Chapter 6

Conclusion

This dissertation advanced the capabilities of existing human emotion perception systems and created new connections between emotion perception and multimedia analysis, social media analysis, and multimedia forensics. Specifically, two novel algorithms were introduced in this dissertation to improve human emotion perception models. These algorithms were subsequently applied to various domains, such as detecting fabricated multimedia, comprehending human behavior and psychology on social media networks, and extracting emotions elicited by movies.

The initial segment of this dissertation centered around enhancing emotion perception models via two distinct approaches. The first approach proposed a new technique to fuse multiple modalities for multimodal emotion perception models. The second approach involved leveraging contextual information, such as the background scene, multiple modalities of the human subject, and socio-dynamic inter-agent interactions, to predict perceived emotion. This led to the development of context-aware human emotion perception models.

The subsequent segment of this dissertation delved into three distinct domains of AI applications. These included video manipulations and deepfake detection, multimedia content analysis, and user behavior analysis on social media platforms. We showed that

the solutions proposed for these applications can be enriched by using ideas from human behavior and emotion understanding models.

To substantiate our discoveries, we conducted evaluations on cutting-edge datasets using the most advanced techniques for the relevant research problems. Moreover, we introduced three new datasets named GroupWalk, VIDEOSHAM, and, INTENTGRAM to enhance the research in this field.

We believe that with the extensive research conducted in this dissertation, we have contributed new findings to the field of emotion perception. By introducing novel algorithms and incorporating ideas from emotion perception into various domains of AI applications, we have broadened the understanding of how ideas of human behavior and emotions can be utilized in different contexts. Moreover, by exploring three very diverse applications we have shown the potential of integrating human emotion perception in vastly different domains. Furthermore, by demonstrating the effectiveness of emotion perception in these applications, such that it helps them in combating fake content and protecting themselves on social media platforms, this will enable individuals to become more comfortable with the role and potential of emotion perception.

Bibliography

(N.d.). https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

(N.d.). https://foundationinc.co/wp-content/uploads/2018/12/NYT-Psychology-Of-Sharing.pdf.

Afchar, Darius et al. (2018). "Mesonet: a compact facial video forgery detection network". In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, pp. 1–7.

AI, Google (n.d.). *Google AI Blog: Contributing Data to Deepfake Detection Research*. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. (Accessed on 02/16/2020).

Ajzen, Icek (1991). "The theory of planned behavior". In: *Organizational behavior and human decision processes* 50.2, pp. 179–211.

Akputu, Kingsley Oryina, Kah Phooi Seng, and Yun Li Lee (2013). "Facial emotion recognition for intelligent tutoring environment". In: *2nd International Conference on Machine Learning and Computer Science (IMLCS'2013)*, pp. 9–13.

Aldao, Amelia (2013). "The future of emotion regulation research: Capturing context". In: *Perspectives on Psychological Science* 8.2, pp. 155–172.

Ali, Afsheen Rafaqat et al. (2017). "High-level concepts for affective understanding of images". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 679–687.

Allen, Chris T, Karen A Machleit, and Susan Schultz Kleine (1992). "A comparison of attitudes and emotions as predictors of behavior at diverse levels of behavioral experience". In: *Journal of consumer research* 18.4, pp. 493–504.

Allen, Jennifer et al. (2020). "Evaluating the fake news problem at the scale of the information ecosystem". In: *Science Advances* 6.14, eaay3539.

Alqahtani, Fehaid and Naeem Ramzan (2019). "Comparison and efficacy of synergistic intelligent tutoring systems with human physiological response". In: *Sensors* 19.3, p. 460.

Amerini, Irene et al. (2011). "A sift-based forensic method for copy–move attack detection and transformation recovery". In: *IEEE transactions on information forensics and security* 6.3, pp. 1099–1110.

Anderson, Katie (July 2018). "Getting acquainted with social networks and apps: combating fake news on social media". In: *Library Hi Tech News* 35. DOI: 10.1108/LHTN-02-2018-0010.

Andrew, Galen et al. (2013). "Deep canonical correlation analysis". In: *ICML*, pp. 1247–1255.

Angus, Jean (Mar. 2022). "Happify Health and Zuellig Pharma Partner to Commercialize Prescription Digital Therapeutics in Asia". In: *Business Wire*. URL: https://www.businesswire.com/news/home/20220314005181/en/Happify-Health-and-Zuellig-Pharma-Partner-to-Commercialize-Prescription-Digital-Therapeutics-in-Asia.

Armijo, Larry (1966). "Minimization of functions having Lipschitz continuous first partial derivatives." In: *Pacific J. Math.* 16.1, pp. 1–3. URL: https://projecteuclid.org:443/euclid.pjm/1102995080.

Athanasiadou, Angeliki and Elzbieta Tabakowska (2010). *Speaking of emotions: Conceptualisation and expression*. Vol. 10. Walter de Gruyter.

Aviezer, Hillel, Yaacov Trope, and Alexander Todorov (2012). "Body cues, not facial expressions, discriminate between intense positive and negative emotions". In: *Science* 338.6111, pp. 1225–1229.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL].

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2017). "Multimodal Machine Learning: A Survey and Taxonomy". In: *CoRR* abs/1705.09406. arXiv: 1705.09406. URL: http://arxiv.org/abs/1705.09406.

Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency (2016). "Openface: an open source facial behavior analysis toolkit". In: *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 1–10.

Bang, Hae-Kyong et al. (2000). "Consumer concern, knowledge, belief, and attitude toward renewable energy: An application of the reasoned action theory". In: *Psychology & Marketing* 17.6, pp. 449–468.

Barger, Patricia B and Alicia A Grandey (2006). "Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms". In: *Academy of management journal* 49.6, pp. 1229–1238.

Barrett, Lisa Feldman, Batja Mesquita, and Maria Gendron (2011). "Context in emotion perception". In: *Current directions in psychological science* 20.5, pp. 286–290.

Barrett, Lisa Feldman, Batja Mesquita, and Eliot R Smith (2010). "The context principle". In: *The mind in context* 1, p. 2.

Barros, Pablo et al. (2018). "The omg-emotion behavior dataset". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–7.

Bashar, M et al. (2010). "Exploring duplicated regions in natural images". In: *IEEE Transactions on Image Processing*.

Batziou, Elissavet et al. (2018). "Visual and Audio Analysis of Movies Video for Emotion Detection@ Emotional Impact of Movies Task MediaEval 2018." In: *MediaEval*.

Bauer, Andrea et al. (2009). "The autonomous city explorer: Towards natural human-robot interaction in urban environments". In: *International journal of social robotics* 1, pp. 127–140.

Baur, Tobias et al. (2013). "A job interview simulation: Social cue-based interaction with a virtual character". In: *2013 International Conference on Social Computing*. IEEE, pp. 220–227.

Baveye, Yoann et al. (2015). "Liris-accede: A video database for affective content analysis". In: *IEEE Transactions on Affective Computing* 6.1, pp. 43–55.

Belkin, Mikhail and Partha Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6, pp. 1373–1396.

Bhattacharya, Uttaran et al. (2020). "Step: Spatial temporal graph convolutional networks for emotion perception from gaits". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02, pp. 1342–1350.

Bhullar, Naureen (2012). "Self-ratings of love and fear on emotional contagion scale depend on the environmental context of rating". In: *Current Research in Social Psychology* 2.

Bigham, Jeffrey P et al. (2010). "VizWiz: nearly real-time answers to visual questions". In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342.

Botha, Johnny and Heloise Pieterse (2020). "Fake news and deepfakes: A dangerous threat for 21st century information security". In: *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*, p. 57.

Bougiatiotis, Konstantinos and Theodoros Giannakopoulos (2018). "Enhanced movie content similarity based on textual, auditory and visual information". In: *Expert Systems with Applications* 96, pp. 86–102.

Brady, William J, MJ Crockett, and Jay J Van Bavel (2020). "The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online". In: *Perspectives on Psychological Science* 15.4, pp. 978–1010.

Brady, William J, Ana P Gantman, and Jay J Van Bavel (2020). "Attentional capture helps explain why moral and emotional content go viral." In: *Journal of Experimental Psychology: General* 149.4, p. 746.

Bratman, Michael E. (1988). "Intention,–Plans,–And–Practical–Reason". In: *Mind* 97.388, pp. 632–634.

Brunner, Jim (June 2020). *Fox News runs digitally altered images in coverage of Seattle's protests, Capitol Hill Autonomous Zone | The Seattle Times*. Link.

Bulat, Adrian and Georgios Tzimiropoulos (2017). "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)". In: *International Conference on Computer Vision*.

Busso, Carlos et al. (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4, p. 335.

Cai, Linqin et al. (2019). "Audio-Textual Emotion Recognition Based on Improved Neural Networks". In: *Mathematical Problems in Engineering* 2019.

Cannon, Walter B (1927). "The James-Lange theory of emotions: A critical examination and an alternative theory". In: *The American journal of psychology* 39.1/4, pp. 106–124.

Cao, Mengxiao et al. (2017). "A method of emotion contagion for crowd evacuation". In: *Physica A: Statistical Mechanics and its Applications* 483, pp. 250–258.

Cao, Zhe et al. (2017). "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Carvalho, Sandra et al. (2012). "The emotional movie database (EMDB): A self-report and psychophysiological study". In: *Applied psychophysiology and biofeedback* 37.4, pp. 279–294.

Castellano, Ginevra, Loic Kessous, and George Caridakis (2008). "Emotion recognition through multiple modalities: face, body gesture, speech". In: *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, pp. 92–103.

Center, Pew Research (2018). "Social media use in 2018". In: URL: https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/.

— (2021). "Social Media Fact Sheet". In: URL: https://www.pewresearch.org/internet/fact-sheet/social-media/.

Chandra, Rohan et al. (2019). "Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs". In: *Proceedings of the 3rd ACM Computer Science in Cars Symposium*, pp. 1–9.

Chen, Chaofeng and Jiadi Mo (2022). *IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment*. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch.

Chen, Qishan et al. (2019). "How leaders' psychological capital influence their followers' psychological capital: social exchange or emotional contagion". In: *Frontiers in psychology* 10, p. 1578.

Chen, Tao et al. (2014). "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks". In: *arXiv preprint arXiv:1410.8586*.

Chen, Xi et al. (2020). "Event popularity prediction using influential hashtags from social media". In: *IEEE Transactions on Knowledge and Data Engineering*.

Cheng, Jianpeng, Li Dong, and Mirella Lapata (2016). *Long Short-Term Memory-Networks for Machine Reading*. arXiv: 1601.06733 [cs.CL].

Chernykh, Vladimir and Pavel Prikhodko (2017a). "Emotion recognition from speech with recurrent neural networks". In: *arXiv preprint arXiv:1701.08071*.

— (2017b). "Emotion recognition from speech with recurrent neural networks". In: *arXiv preprint arXiv:1701.08071*.

Choi, Woo Yong, Kyu Ye Song, and Chan Woo Lee (2018). "Convolutional attention networks for multimodal emotion recognition from speech and text data". In: *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, pp. 28–34.

Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman (2018). "Voxceleb2: Deep speaker recognition". In: *arXiv preprint arXiv:1806.05622*.

Clavel, Chloé et al. (June 2008a). "Fear-type emotion recognition for future audio-based surveillance systems". In: *Speech Communication* 50, pp. 487–503. DOI: 10.1016/j.specom.2008.03.012.

— (June 2008b). "Fear-type emotion recognition for future audio-based surveillance systems". In: *Speech Communication* 50, pp. 487–503. DOI: 10.1016/j.specom.2008.03.012.

Coëgnarts, Maarten and Peter Kravanja (2016). "Perceiving Causality in Character Perception: A Metaphorical Study of Causation in Film". In: *Metaphor and Symbol* 31.2, pp. 91–107. DOI: 10.1080/10926488.2016.1150762. eprint: https://doi.org/10.1080/10926488.2016.1150762. URL: https://doi.org/10.1080/10926488.2016.1150762.

Cohn, Jeffrey F et al. (2009). "Detecting depression from facial actions and vocal prosody". In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, pp. 1–7.

Coviello, Lorenzo et al. (2014). "Detecting emotional contagion in massive social networks". In: *PloS one* 9.3, e90315.

Cowley, Elizabeth (2014). "Consumers telling consumption stories: word-of-mouth and retrospective evaluations". In: *Journal of business research* 67.7, pp. 1522–1529.

Cozzolino, Davide, Giovanni Poggi, and Luisa Verdoliva (2015). "Efficient dense-field copy–move forgery detection". In: *IEEE Transactions on Information Forensics and Security* 10.11, pp. 2284–2297.

Crandall, David and Noah Snavely (2012). "Modeling people and places with internet photo collections". In: *Communications of the ACM* 55.6, pp. 52–60.

Davidson, RJ, KR Scherer, and HH Goldsmith (2003). "The role of affect in decision making". In: *Handbook of affective sciences* 3, pp. 619–642.

Del Vicario, Michela et al. (2016). "Echo chambers: Emotional contagion and group polarization on facebook". In: *Scientific reports* 6.1, pp. 1–12.

Dellarocas, Chrysanthos, Xiaoquan Zhang, and Neveen F Awad (2007). "Exploring the value of online product reviews in forecasting sales: The case of motion pictures". In: *Journal of Interactive marketing* 21.4, pp. 23–45.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Deng, Jiankang et al. (2020). "Retinaface: Single-shot multi-level face localisation in the wild". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212.

Denham, Susanne A et al. (2000). "Prediction of externalizing behavior problems from early to middle childhood: The role of parental socialization and emotion expression". In: *Development and psychopathology* 12.1, pp. 23–45.

Dhall, Abhinav, Roland Goecke, Jyoti Joshi, et al. (2016). "Emotiw 2016: Video and group-level emotion recognition challenges". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, pp. 427–432.

Dhall, Abhinav, Roland Goecke, Simon Lucey, et al. (2012). "Collecting large, richly annotated facial-expression databases from movies". In: *IEEE multimedia* 19.3, pp. 34–41.

Dolhansky, Brian et al. (2019). "The Deepfake Detection Challenge (DFDC) Preview Dataset". In: *arXiv preprint arXiv:1910.08854*.

Douglas-Cowie, Ellen et al. (2007). "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data". In: *International conference on affective computing and intelligent interaction*. Springer, pp. 488–500.

Downs, Andrew and Paul Strand (2008). "Effectiveness of emotion recognition training for young children with developmental delays." In: *Journal of Early and Intensive Behavior Intervention* 5.1, p. 75.

Ekman, Paul (1971). "Universals and cultural differences in facial expressions of emotion." In: *Nebraska symposium on motivation*. University of Nebraska Press.

— (1978). "Facial action coding system: A technique for the measurement of facial movement". In: *Consulting Psychologists Press: Palo Alto*.

— (1993). "Facial expression and emotion." In: *American psychologist* 48.4, p. 384.

Ekman, Paul, Wallace V Freisen, and Sonia Ancoli (1980). "Facial signs of emotional experience." In: *Journal of personality and social psychology* 39.6, p. 1125.

Ekman, Paul, Robert W Levenson, and Wallace V Friesen (1983). "Autonomic nervous system activity distinguishes among emotions". In: *science* 221.4616, pp. 1208–1210.

Ekman, Rosenberg (1997). *What the face reveals: sic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Eyben, Florian et al. (2013). "Recent developments in opensmile, the munich open-source multimedia feature extractor". In: *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835–838.

Fan, Rui et al. (2014). "Anger is more influential than joy: Sentiment correlation in Weibo". In: *PloS one* 9.10, e110184.

Fernández-Dols, José-Miguel and Maria-Angeles Ruiz-Belda (1995). "Expression of emotion versus expressions of emotions". In: *Everyday conceptions of emotion*. Springer, pp. 505–522.

Ferrara, Emilio and Zeyao Yang (2015). "Measuring emotional contagion in social media". In: *PloS one* 10.11, e0142390.

Figueira, Alvaro and Luciana Oliveira (Dec. 2017). "The current state of fake news: challenges and opportunities". In: *Procedia Computer Science* 121, pp. 817–825. DOI: 10.1016/j.procs.2017.11.106.

Fishbein, Martin and Icek Ajzen (1977). "Belief, attitude, intention, and behavior: An introduction to theory and research". In: *Philosophy and Rhetoric* 10.2.

Fox, Alexa K et al. (2018). "The face of contagion: Consumer response to service failure depiction in online reviews". In: *European Journal of Marketing*.

Fukui, Hiroshi et al. (2019). "Attention branch network: Learning of attention mechanism for visual explanation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10705–10714.

Garg, Samanyou (2019). "Group emotion recognition using machine learning". In: *arXiv preprint arXiv:1905.01118*.

Giannakopoulos, Theodoros (2015). "pyaudioanalysis: An open-source python library for audio signal analysis". In: *PloS one* 10.12.

Glymour, Clark, Kun Zhang, and Peter Spirtes (2019). "Review of Causal Discovery Methods Based on Graphical Models". In: *Frontiers in Genetics* 10, p. 524. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00524. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00524.

Goffman, Erving (2021). *The presentation of self in everyday life*. Anchor.

Goldenberg, Amit and James Gross (2019). "Digital Emotion Contagion". In.

Gonzaga, Victor Machado, Nils Murrugarra-Llerena, and Ricardo Marcacini (2021). "Multimodal intent classification with incomplete modalities using text embedding propagation". In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pp. 217–220.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

Greenaway, Katharine H, Elise K Kalokerinos, and Lisa A Williams (2018). "Context is everything (in emotion research)". In: *Social and Personality Psychology Compass* 12.6, e12393.

Gross, James J and Robert W Levenson (1995). "Emotion elicitation using films". In: *Cognition & emotion* 9.1, pp. 87–108.

Gruzd, Anatoliy, Sophie Doiron, and Philip Mai (2011). "Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics". In: *2011 44th Hawaii International Conference on System Sciences*. IEEE, pp. 1–9.

Guan, Haiying et al. (2019). "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation". In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, pp. 63–72.

Güera, David and Edward J Delp (2018). "Deepfake video detection using recurrent neural networks". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, pp. 1–6.

Gunes, Hatice and Massimo Piccardi (2007). "Bi-modal emotion recognition from expressive face and body gestures". In: *Journal of Network and Computer Applications* 30.4, pp. 1334–1345.

Guo, Ruocheng et al. (July 2020). "A Survey of Learning Causality with Data: Problems and Methods". In: *ACM Comput. Surv.* 53.4. ISSN: 0360-0300. DOI: 10.1145/3397269. URL: https://doi.org/10.1145/3397269.

Guo, Shengnan et al. (2019). "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 922–929.

Gurban, Mihai et al. (2008). "Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition". In: *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 237–240.

Haber, Nick, Catalin Voss, and Dennis Wall (Mar. 2020). "Upgraded Google Glass Helps Autistic Kids See Emotions". In: *IEEE Spectrum*. URL: https://spectrum.ieee.org/upgraded-google-glass-helps-autistic-kids-see-emotions.

Harris, Douglas (2018). "Deepfakes: False pornography is here and the law cannot protect you". In: *Duke L. & Tech. Rev.* 17, p. 99.

Hatfield, Elaine, John T Cacioppo, and Richard L Rapson (1992). "Primitive emotional contagion." In.

He, Kaiming, Georgia Gkioxari, et al. (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

He, Kaiming, Xiangyu Zhang, et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Saike et al. (2016). "Exploring entrainment patterns of human emotion in social media". In: *PloS one* 11.3, e0150630.

He, Yinan et al. (2021). "ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4360–4369.

Hebden, Sophie (Oct. 2015). "Emotionally literate tech could help treat autism". In: *Horizon Magazine*. URL: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/emotionally-literate-tech-help-treat-autism.

Heidicker, Paul, Eike Langbehn, and Frank Steinicke (2017). "Influence of avatar appearance on presence in social VR". In: *2017 IEEE symposium on 3D user interfaces (3DUI)*. IEEE, pp. 233–234.

Helbing, Dirk and Peter Molnar (1995). "Social force model for pedestrian dynamics". In: *Physical review E* 51.5, p. 4282.

Hochschild, Arlie Russell (1979). "Emotion work, feeling rules, and social structure". In: *American journal of sociology* 85.3, pp. 551–575.

Hodosh, Micah, Peter Young, and Julia Hockenmaier (2013). "Framing image description as a ranking task: Data, models and evaluation metrics". In: *Journal of Artificial Intelligence Research* 47, pp. 853–899.

Hoffmann, Holger et al. (2012). "Mapping discrete emotions into the dimensional space: An empirical approach". In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 3316–3320.

Hogan, Bernie (2010). "The presentation of self in the age of social media: Distinguishing performances and exhibitions online". In: *Bulletin of Science, Technology & Society* 30.6, pp. 377–386.

Hotelling, Harold (1936). "Relations Between Two Sets of Variates". In: *Biometrika* 28.3/4, pp. 321–377. ISSN: 00063444. URL: http://www.jstor.org/stable/2333955.

Hunt, Melissa G et al. (2018). "No more FOMO: Limiting social media decreases loneliness and depression". In: *Journal of Social and Clinical Psychology* 37.10, pp. 751–768.

Hussain, Zaeem, Mingda Zhang, X. Zhang, et al. (2017). "Automatic Understanding of Image and Video Advertisements". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1100–1110.

Hussain, Zaeem, Mingda Zhang, Xiaozhong Zhang, et al. (2017). "Automatic understanding of image and video advertisements". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1705–1715.

Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 216–225.

Hyman, John (1992). "The causal theory of perception". In: *The Philosophical Quarterly (1950-)* 42.168, pp. 277–296.

Ignat, Oana et al. (Nov. 2021). "WhyAct: Identifying Action Reasons in Lifestyle Vlogs". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4770–4785. DOI: 10.18653/v1/2021.emnlp-main.392. URL: https://aclanthology.org/2021.emnlp-main.392.

Islam, Md Rafiqul et al. (2020). "Deep learning for misinformation detection on online social networks: a survey and new perspectives". In: *Social Network Analysis and Mining* 10, pp. 1–20.

Jakobs, Esther, Antony SR Manstead, and Agneta H Fischer (2001). "Social context effects on facial activity in a negative emotional setting." In: *Emotion* 1.1, p. 51.

JAMES, WILLIAM (Apr. 1884). "II.—WHAT IS AN EMOTION ?" In: *Mind* os-IX.34, pp. 188–205. ISSN: 0026-4423. DOI: 10.1093/mind/os-IX.34.188. eprint: https://academic.oup.com/mind/article-pdf/os-IX/34/188/9278514/os-IX_34_188.pdf. URL: https://doi.org/10.1093/mind/os-IX.34.188.

Jang, Jin Yea et al. (2016). "Teens engage more with fewer photos: temporal and comparative analysis on behaviors in instagram". In: *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pp. 71–81.

Jasper, James M (1998). "The emotions of protest: Affective and reactive emotions in and around social movements". In: *Sociological forum*. Vol. 13. Springer, pp. 397–424.

Jeon, Hyeonseong, Youngoh Bang, and Simon S Woo (2019). "FakeTalkerDetect: Effective and Practical Realistic Neural Talking Head Detection with a Highly Unbalanced Dataset". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0.

Jia, Menglin et al. (2021). "Intentonomy: a dataset and study towards human intent understanding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12986–12996.

Jiang, Liming et al. (2020). "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection". In: *arXiv preprint arXiv:2001.03024*.

Jin, Zitong et al. (2017). "THUHCSI in MediaEval 2017 Emotional Impact of Movies Task." In: *MediaEval* 17, pp. 13–17.

Joshi, Dhiraj et al. (2014). "On aesthetics and emotions in scene images: A computational perspective". In: *Scene vision: making sense of what we see*, p. 241.

Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang (2021). "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach". In: *Multimedia tools and applications* 80.8, pp. 11765–11788.

Kassam, Karim Sadik (2010). *Assessment of emotional experience through facial expression*. Harvard University.

Keles, Betul, Niall McCrae, and Annmarie Grealish (2020). "A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents". In: *International Journal of Adolescence and Youth* 25.1, pp. 79–93.

Kettenring, Jon R (1971). "Canonical analysis of several sets of variables". In: *Biometrika* 58.3, pp. 433–451.

Khalid, Hasam, Shahroz Tariq, and Simon S Woo (2021). "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset". In: *arXiv preprint arXiv:2108.05080*.

Khelifi, Fouad and Ahmed Bouridane (2017). "Perceptual video hashing for content identification and authentication". In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.1, pp. 50–67.

Kim, Dahun et al. (2019). "Deep Video Inpainting". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5792–5801.

— (2020). "Recurrent Temporal Aggregation Framework for Deep Video Inpainting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.5, pp. 1038–1052.

Kim, Jinyoung and June Ahn (2013). "The show must go on: the presentation of self during interpersonal conflict on Facebook". In: *Proceedings of the American Society for Information Science and Technology* 50.1, pp. 1–10.

Kim, Soojung, Joonghwa Lee, and Doyle Yoon (2015). "Norms in social media: The application of theory of reasoned action and personal norms in predicting interactions with Facebook page like ads". In: *Communication Research Reports* 32.4, pp. 322–331.

Kim, Yelin, Honglak Lee, and Emily Mower Provost (2013). "Deep learning for robust feature generation in audiovisual emotion recognition". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 3687–3691.

Kimura, Kenta et al. (n.d.). "Emotional State of Being Moved Elicited by Films: A Comparison With Several Positive Emotions". In: *Frontiers in Psychology* (). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.01935. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.01935.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kleinsmith, Andrea and Nadia Bianchi-Berthouze (2012). "Affective body expression perception and recognition: A survey". In: *IEEE Transactions on Affective Computing* 4.1, pp. 15–33.

— (2013). "Affective Body Expression Perception and Recognition: A Survey". In: *IEEE Transactions on Affective Computing* 4, pp. 15–33.

Knapp, R Benjamin, Jonghwa Kim, and Elisabeth André (2010). "Physiological signals and their use in augmenting emotion recognition for human–machine interaction". In: *Emotion-oriented systems: The Humaine handbook*. Springer, pp. 133–159.

Ko, Tobey H et al. (2018). "Towards Learning Emotional Subspace." In: *MediaEval*.

Koelstra, Sander et al. (n.d.). "Deap: A database for emotion analysis; using physiological signals". In: ().

Korshunov, Pavel and Sébastien Marcel (2018a). "Deepfakes: a new threat to face recognition? assessment and detection". In: *arXiv preprint arXiv:1812.08685*.

— (2018b). "Speaker inconsistency detection in tampered video". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 2375–2379.

Korshunova, Iryna et al. (2017). "Fast face-swap using convolutional neural networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3677–3685.

Kossaifi, Jean et al. (2019). "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kosti, Ronak, Jose M Alvarez, et al. (2017a). "EMOTIC: Emotions in Context dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 61–69.

— (2017b). "Emotion recognition in context". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kosti, Ronak, Jose M Alvarez, et al. (2019). "Context based emotion recognition using emotic dataset". In: *IEEE transactions on pattern analysis and machine intelligence* 42.11, pp. 2755–2766.

— (July 2017c). "EMOTIC: Emotions in Context Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Kövecses, Zoltán (2003). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Kramer, Adam DI, Jamie E Guillory, and Jeffrey T Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24, pp. 8788–8790.

Krishnamurthy, Gangeshwar et al. (2018). "A deep learning approach for multimodal deception detection". In: *arXiv preprint arXiv:1803.00344*.

Kruk, Julia et al. (2019). "Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4622–4632.

Kuang, Beibei et al. (2019). "Universality vs. cultural specificity in the relations among emotional contagion, emotion regulation, and mood state: An emotion process perspective". In: *Frontiers in psychology*, p. 186.

Kwon, Patrick et al. (2021). "KoDF: A Large-scale Korean DeepFake Detection Dataset". In: *arXiv preprint arXiv:2103.10094*.

Labbe, Aaron (Nov. 2022). "Emotion AI: Why It's the Future of Digital Health". In: *Forbes*. URL: https://www.forbes.com/sites/forbestechcouncil/2022/11/23/emotion-ai-why-its-the-future-of-digital-health/?sh=41063f346516.

Lakhiwal, Akshat and Arpan Kumar Kar (2016). "Insights from Twitter analytics: modeling social media personality dimensions and impact of breakthrough events". In: *Conference on e-Business, e-Services and e-Society*. Springer, pp. 533–544.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

Lawrence, I and Kuei Lin (1989). "A concordance correlation coefficient to evaluate reproducibility". In: *Biometrics*, pp. 255–268.

Ledgerwood, Alison (2014). "Evaluations in their social context: Distance regulates consistency and context dependence". In: *Social and Personality Psychology Compass* 8.8, pp. 436–447.

LeDoux, Joseph (1998). *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster.

Lee, Jiyoung et al. (2019). "Context-aware emotion recognition networks". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10143–10152.

Li, Jamy et al. (2016). "Social robots and virtual agents as lecturers for video instruction". In: *Computers in Human Behavior* 55, pp. 1222–1230.

Li, Jingjing, Jian Zhang, and Zhiguo Yang (2017). "Associations between a leader's work passion and an employee's work passion: a moderated mediation model". In: *Frontiers in Psychology* 8, p. 1447.

Li, Lingzhi et al. (2019). "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping". In: *arXiv preprint arXiv:1912.13457*.

Li, Yuezun and Siwei Lyu (2018). "Exposing deepfake videos by detecting face warping artifacts". In: *arXiv preprint arXiv:1811.00656*.

Li, Yuezun and Siwei Lyu (2019). "Exposing DeepFake Videos By Detecting Face Warping Artifacts". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Li, Yuezun, Xin Yang, et al. (2019). "Celeb-df: A new dataset for deepfake forensics". In: *arXiv preprint arXiv:1909.12962*.

Li, Zhengqi and Noah Snavely (2018). "Megadepth: Learning single-view depth prediction from internet photos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050.

Liao, Sheng-Yang and Tian-Qiang Huang (2013). "Video copy-move forgery detection and localization based on Tamura texture features". In: *2013 6th international congress on image and signal processing (CISP)*. Vol. 2. IEEE, pp. 864–868.

Lin, Ruoyun and Sonja Utz (2015). "The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength". In: *Computers in human behavior* 52, pp. 29–38.

Lin, Tsung-Yi et al. (2014). "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer, pp. 740–755.

Lin, Xiaolin, Mauricio Featherman, and Saonee Sarker (2013). "Information sharing in the context of social media: an application of the theory of reasoned action and social capital theory". In.

Liu, Haomiao et al. (2016). "Deep Supervised Hashing for Fast Image Retrieval". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2064–2072.

Liu, Kuan et al. (2018). "Learn to combine modalities in multimodal deep learning". In: *arXiv preprint arXiv:1805.11730*.

Liu, Yaqi et al. (2021). "Two-Stage Copy-Move Forgery Detection With Self Deep Matching and Proposal SuperGlue". In: *IEEE Transactions on Image Processing* 31, pp. 541–555.

Long, Chengjiang et al. (2019). "A Coarse-to-fine Deep Convolutional Neural Network Framework for Frame Duplication Detection and Localization in Forged Videos." In: *CVPR Workshops*, pp. 1–10.

Lozano, Aurelie C et al. (2009). "Grouped graphical Granger modeling methods for temporal causal modeling". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 577–586.

Lütkepohl, Helmut (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Mahmood, Toqeer et al. (2016). "Copy-move forgery detection technique for forensic analysis in digital images". In: *Mathematical Problems in Engineering* 2016.

Majumder, Navonil et al. (2018). "Multimodal sentiment analysis using hierarchical fusion with context modeling". In: *Knowledge-based systems* 161, pp. 124–133.

Manera, Valeria, Elisa Grandi, and Livia Colle (2013). "Susceptibility to emotional contagion for negative emotions improves detection of smile authenticity". In: *Frontiers in Human Neuroscience* 7, p. 6.

Marchese, Kieron (Mar. 2022). "Fuseproject's ElliQ Robot is Designed to Keep the Elderly Company". In: *DesignWanted*. URL: https://designwanted.com/elliq-robot-fuseproject/.

Markus, Hazel R and Shinobu Kitayama (1991). "Culture and the self: Implications for cognition, emotion, and motivation." In: *Psychological review* 98.2, p. 224.

Martín, Eva García, Niklas Lavesson, and Mina Doroud (2016). "Hashtags and followers". In: *Social Network Analysis and Mining* 6.1, pp. 1–15.

Martinez, Aleix M (2019). "Context may reveal how you feel". In: *Proceedings of the National Academy of Sciences* 116.15, pp. 7169–7171.

Matern, Falko, Christian Riess, and Marc Stamminger (2019). "Exploiting visual artifacts to expose deepfakes and face manipulations". In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, pp. 83–92.

Matsui, Tetsuya and Seiji Yamada (2019). "Designing trustworthy product recommendation virtual agents operating positive emotion and having copious amount of knowledge". In: *Frontiers in Psychology* 10, p. 675.

McDuff, Daniel, Rana El Kaliouby, Jeffrey F Cohn, et al. (2014). "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads". In: *IEEE Transactions on Affective Computing* 6.3, pp. 223–235.

McDuff, Daniel, Rana El Kaliouby, David Demirdjian, et al. (2013). "Predicting online media effectiveness based on smile responses gathered over the internet". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, pp. 1–7.

McKeown, Gary et al. (2010). "The SEMAINE corpus of emotionally coloured character interactions". In: *2010 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1079–1084.

McManus, Ashley (Nov. 2017). "Driver Emotion Recognition and Real-Time Facial Analysis for the Automotive Industry". In: *Affectiva Blog*. URL: https://blog.affectiva.com/driver-emotion-recognition-and-real-time-facial-analysis-for-the-automotive-industry.

McNulty, James K and Frank D Fincham (2012). "Beyond positive psychology? Toward a contextual view of psychological processes and well-being." In: *American Psychologist* 67.2, p. 101.

Media, Common Sense (2020). "Social Media, Social Life: Teens Reveal Their Experiences". In: URL: https://www.commonsensemedia.org/research/social-media-social-life-2020.

Meeren, Hanneke KM, Corné CRJ van Heijnsbergen, and Beatrice de Gelder (2005). "Rapid perceptual integration of facial expression and emotional body language". In: *Proceedings of the National Academy of Sciences* 102.45, pp. 16518–16523.

Mehrabian, Albert (1980). "Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies". In.

— (1996). "Analysis of the big-five personality factors in terms of the PAD temperament model". In: *Australian journal of Psychology* 48.2, pp. 86–92.

Mehrabian, Albert and James A Russell (1974a). *An approach to environmental psychology.* the MIT Press.

— (1974b). *An approach to environmental psychology.* the MIT Press.

Mesquita, Batja and Michael Boiger (2014). "Emotions in context: A sociodynamic model of emotions". In: *Emotion Review* 6.4, pp. 298–302.

Micu, Anca Cristina and Joseph T Plummer (2010). "Measurable emotions: How television ads really work: Patterns of reactions to commercials can demonstrate advertising effectiveness". In: *Journal of Advertising Research* 50.2, pp. 137–153.

Mittal, Trisha, Uttaran Bhattacharya, et al. (2020a). "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues". In: *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832.

— (2020b). "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 02, pp. 1359–1367.

Mittal, Trisha, Pooja Guhan, et al. (2020). "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14222–14231.

Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor (Jan. 2019). "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1, pp. 18–31. ISSN: 2371-9850. DOI: 10.1109/taffc.2017.2740923. URL: http://dx.doi.org/10.1109/taffc.2017.2740923.

Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi (2020). "A BERT-based transfer learning approach for hate speech detection in online social media". In: *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*. Springer, pp. 928–940.

Nahi, Ylenia Camassa et al. (2022). "Recognizing emotions and effects of traumatic brain injury". In: *Cellular, Molecular, Physiological, and Behavioral Aspects of Traumatic Brain Injury*. Elsevier, pp. 515–526.

Narayanan, Venkatraman et al. (2020). "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 8200–8207.

Navarretta, Costanza (2012). "Individuality in communicative bodily behaviours". In: *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Springer, pp. 417–423.

News, AP (Feb. 2020). *False video of Joe Biden claims Democrat candidate greeted wrong US state at a rally | Euronews*. Link. (Accessed on 04/07/2022).

Nguyen, Huy H, Fuming Fang, et al. (2019). "Multi-task learning for detecting and segmenting manipulated facial images and videos". In: *arXiv preprint arXiv:1906.06876*.

Nguyen, Huy H, Junichi Yamagishi, and Isao Echizen (2019). "Capsule-forensics: Using capsule networks to detect forged images and videos". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2307–2311.

Niemeier, Susanne and René Dirven (1997). *The language of emotions: Conceptualization, expression, and theoretical foundation*. John Benjamins Publishing.

Nirkin, Yuval, Yosi Keller, and Tal Hassner (2019). "Fsgan: Subject agnostic face swapping and reenactment". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193.

NPR (July 2008). *Photo Of Iran's Missile Launch Was Manipulated : NPR*. Link. (Accessed on 04/07/2022).

Oliveira, Eva, Pedro Martins, and Teresa Chambel (2011). "Ifelt: accessing movies through our emotions". In: *Proceedings of the 9th European Conference on Interactive TV and Video*, pp. 105–114.

Ong, D. et al. (2019). "Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset". In: *IEEE Transactions on Affective Computing*, pp. 1–1. DOI: 10.1109/TAFFC.2019.2955949.

Ong, Desmond et al. (2019). "Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset". In: *IEEE Transactions on Affective Computing*.

Ong, Desmond C, Jamil Zaki, and Noah D Goodman (2019). "Computational models of emotion inference in theory of mind: A review and roadmap". In: *Topics in cognitive science* 11.2, pp. 338–357.

Panda, Rameswar et al. (2018). "Contemplating visual emotions: Understanding and overcoming dataset bias". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 579–595.

Pantic, Maja et al. (2005). "Affective multimodal human-computer interaction". In: *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 669–676.

Papadamou, Kostantinos et al. (2020). "Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 522–533.

Papadopoulou, Olga et al. (2017). "Web Video Verification Using Contextual Cues". In: *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*. MFSec '17. Bucharest, Romania: Association for Computing Machinery, pp. 6–10. ISBN: 9781450350341. DOI: 10.1145/3078897.3080535. URL: https://doi.org/10.1145/3078897.3080535.

Paszke, Adam et al. (2017). "Automatic Differentiation in PyTorch". In: *NIPS Autodiff Workshop*.

Pediatrics, American Academy of (2021). "The Impact of Social Media on Children, Adolescents, and Families". In: URL: https://pediatrics.aappublications.org/content/147/5/e2020029586.

Peng, Yilang and JOHN B JEMMOTT III (2018). "Feast for the Eyes: Effects of Food Perceptions and Computer Vision Features on Food Photo Popularity." In: *International Journal of Communication (19328036)* 12.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perov, Ivan et al. (2020). "Deepfacelab: A simple, flexible and extensible face swapping framework". In: *arXiv preprint arXiv:2005.05535*.

Peslak, Alan, Wendy Ceccucci, and Patricia Sendall (2012). "An empirical study of social networking behavior using theory of reasoned action". In: *Journal of Information Systems Applied Research* 5.3, p. 12.

Philippot, Pierre (1993). "Inducing and assessing differentiated emotion-feeling states in the laboratory". In: *Cognition and emotion* 7.2, pp. 171–193.

Picard, Rosalind W. (1995). "Affective computing". In: *Proceedings of the 3rd international conference on multimodal interfaces*. ACM, pp. 1–8.

Pilli, Stephen et al. (2020). "Predicting Sentiments in Image Advertisements Using Semantic Relations Among Sentiment Labels". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 408–409.

Prajwal, KR et al. (2020). "A lip sync expert is all you need for speech to lip generation in the wild". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492.

Purohit, Hemant et al. (2015). "Intent classification of short-text on social media". In: *2015 ieee international conference on smart city/socialcom/sustaincom (smartcity)*. IEEE, pp. 222–228.

Qian, Xueming et al. (2013). "Tagging photos using users' vocabularies". In: *Neurocomputing* 111, pp. 144–153.

Quan, Khanh-An C, Vinh-Tiep Nguyen, and Minh-Triet Tran (2018). "Frame-based Evaluation with Deep Features to Predict Emotional Impact of Movies." In: *MediaEval*.

Rainie, Lee, Amanda Lenhart, and Aaron Smith (2012). "The tone of life on social networking sites". In: *Pew Internet Report*.

Recasens, Adriá Recasens Continente (2016). "Where are they looking?" PhD thesis. Massachusetts Institute of Technology.
Resnik, Michael David (1967). "The context principle in Frege's philosophy". In: *Philosophy and Phenomenological Research* 27.3, pp. 356–365.
Ribeiro, Manoel et al. (2018). "Characterizing and detecting hateful users on twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.
Ringeval, Fabien et al. (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *FG*. IEEE, pp. 1–8.
Robitzski, Dan (May 2019). "Amazon Alexa Can Now Analyze Your Emotions". In: *Futurism*. URL: https://futurism.com/the-byte/amazon-alexa-analyzing-emotions.
Rossler, Andreas et al. (2019). "Faceforensics++: Learning to detect manipulated facial images". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–11.
Roth, Daniel et al. (2016). "Avatar realism and social interaction quality in virtual reality". In: *2016 IEEE Virtual Reality (VR)*. IEEE, pp. 277–278.
Russell, James A (2003). "Core affect and the psychological construction of emotion." In: *Psychological review* 110.1, p. 145.
Russell, James A, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols (2003). "Facial and vocal expressions of emotion". In: *Annual review of psychology* 54.1, pp. 329–349.
Ryu, Seung-Jin et al. (2013). "Rotation invariant localization of duplicated image regions based on Zernike moments". In: *IEEE Transactions on Information Forensics and Security* 8.8, pp. 1355–1370.
Sabir, Ekraam et al. (2019). "Recurrent convolutional strategies for face manipulation detection in videos". In: *Interfaces (GUI)* 3, p. 1.
Sahay, Saurav et al. (2018). "Multimodal Relational Tensor Network for Sentiment and Emotion Classification". In: *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pp. 20–27.
Saldias, F Belen and Rosalind W Picard (2019). "Tweet Moodifier: Towards giving emotional awareness to Twitter users". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 1–7.
Sanderson, Conrad (2002). *The vidtimit database*. Tech. rep. IDIAP.
Al-Sanjary, Omar Ismael, Ahmed Abdullah Ahmed, and Ghazali Sulong (2016). "Development of a video tampering dataset for forensic investigation". In: *Forensic science international* 266, pp. 565–572.
Saragih, Jason M, Simon Lucey, and Jeffrey F Cohn (2009). "Face alignment through subspace constrained mean-shifts". In: *2009 IEEE 12th International Conference on Computer Vision*. Ieee, pp. 1034–1041.
Schachter, S and JE Singer (1962). "Psychological review". In: *Psychol. Rev* 69, pp. 379–399.
Schachter, Stanley and Jerome Singer (1962). "Cognitive, social, and physiological determinants of emotional state." In: *Psychological review* 69.5, p. 379.
Schaefer, Alexandre et al. (2010). "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers". In: *Cognition and emotion* 24.7, pp. 1153–1172.
Scherer, Klaus R, Tom Johnstone, and Gundrun Klasmeyer (2003). *Vocal expression of emotion*. Oxford University Press.
Scott, Carol F et al. (2017). "Time spent online: Latent profile analyses of emerging adults' social media use". In: *Computers in Human Behavior* 75, pp. 311–319.

Seo, Minjoon et al. (2018). *Bidirectional Attention Flow for Machine Comprehension*. arXiv: 1611.01603 [cs.CL].

Shan, Caifeng, Shaogang Gong, and Peter W McOwan (2007). "Beyond Facial Expressions: Learning Human Emotion from Body Gestures." In: *BMVC*, pp. 1–10.

Siarohin, Aliaksandr et al. (2019). "First order motion model for image animation". In: *Advances in Neural Information Processing Systems* 32, pp. 7137–7147.

Sikka, Karan et al. (2013). "Multiple kernel learning for emotion recognition in the wild". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 517–524.

Silver, Laura (Aug. 2020). *Misinformation and fears about its impact are pervasive in 11 emerging economies*. URL: https://pewrsr.ch/2ZjbR4o.

Sindhwani, Vikas, Ha Quang Minh, and Aurélie C. Lozano (2013). "Scalable Matrix-Valued Kernel Learning for High-Dimensional Nonlinear Multivariate Regression and Granger Causality". In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. UAI'13. Bellevue, WA: AUAI Press, pp. 586–595.

Sleeper, Manya et al. (2013). "The post that wasn't: exploring self-censorship on facebook". In: *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 793–802.

Soleymani, Mohammad, Sadjad Asghari-Esfeden, et al. (2014). "Continuous emotion detection using EEG signals and facial expressions". In: *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6.

Soleymani, Mohammad, Maja Pantic, and Thierry Pun (2011). "Multimodal emotion recognition in response to videos". In: *IEEE transactions on affective computing* 3.2, pp. 211–223.

Spitzer, Elizabeth G, Eric S Crosby, and Tracy K Witte (2022). "Looking through a filtered lens: Negative social comparison on social media and suicidal ideation among young adults." In: *Psychology of Popular Media*.

Spivak, Russell (2018). "" Deepfakes": The Newest Way to Commit One of the Oldest Crimes". In: *Geo. L. Tech. Rev.* 3, p. 339.

Su, Lichao, Tianqiang Huang, and Jianmei Yang (2015). "A video forgery detection algorithm based on compressive sensing". In: *Multimedia Tools and Applications* 74.17, pp. 6641–6656.

Sun, Jennifer J, Ting Liu, and Gautam Prasad (2019). "Gla in mediaeval 2018 emotional impact of movies task". In: *arXiv preprint arXiv:1911.12361*.

Talevich, Jennifer R. et al. (Feb. 2017). "Toward a comprehensive taxonomy of human motives". In: *PLOS ONE* 12, pp. 1–32. DOI: 10.1371/journal.pone.0172279. URL: https://doi.org/10.1371/journal.pone.0172279.

Tank, Alex et al. (2018). *Neural Granger Causality for Nonlinear Time Series*. arXiv: 1802.05842 [stat.ML].

Tarkiainen, Anssi and Sanna Sundqvist (2005). "Subjective norms, attitudes and intentions of Finnish consumers in buying organic food". In: *British food journal*.

Teixeira, Thales, Rosalind Picard, and Rana El Kaliouby (2014). "Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study". In: *Marketing Science* 33.6, pp. 809–827.

Teräsvirta, T., D. Tjøstheim, and C. Granger (2011). "Modelling Nonlinear Economic Time Series". In.

Thies, Justus, Michael Zollhofer, et al. (2016). "Face2face: Real-time face capture and reenactment of rgb videos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395.

Thies, Justus, Michael Zollhöfer, and Matthias Nießner (2019). "Deferred neural rendering: Image synthesis using neural textures". In: *ACM Transactions on Graphics (TOG)* 38.4, pp. 1–12.
Tiggemann, Marika et al. (2018). "The effect of Instagram "likes" on women's social comparison and body dissatisfaction". In: *Body image* 26, pp. 90–97.
Times, Global (Sept. 2022). "Beijing distributes emotion-sensing equipment to highway and across province bus drivers". In: *Global Times*. URL: https://www.globaltimes.cn/page/202209/1275783.shtml.
Tracy, Jessica L, Richard W Robins, and Roberta A Schriber (2009). "Development of a FACS-verified set of basic and self-conscious emotion expressions." In: *Emotion* 9.4, p. 554.
Trigeorgis, George et al. (2016). "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5200–5204.
Tripathi, Samarth, Sarthak Tripathi, and Homayoon Beigi (2018). "Multi-modal emotion recognition on iemocap dataset using deep learning". In: *arXiv preprint arXiv:1804.05788*.
Tripathy, Soumya, Juho Kannala, and Esa Rahtu (2020). "Icface: Interpretable and controllable face reenactment using gans". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3385–3394.
Tromholt, Morten (2016). "The Facebook experiment: Quitting Facebook leads to higher levels of well-being". In: *Cyberpsychology, behavior, and social networking* 19.11, pp. 661–666.
Valente, Thomas W (2015). "Social networks and health: a relational approach to integrated healthcare". In: *International Journal of Integrated Care* 15.2, e022.
Van Dijck, José and Thomas Poell (2013). "Understanding social media logic". In: *Media and communication* 1.1, pp. 2–14.
Vedula, Nikhita et al. (2017). "Multimodal content analysis for effective advertisements on Youtube". In: *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 1123–1128.
Verdoliva, Luisa and Paolo Bestagini (2019). "Multimedia Forensics". In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: Association for Computing Machinery, pp. 2701–2702. ISBN: 9781450368896. DOI: 10.1145/3343031.3350542. URL: https://doi.org/10.1145/3343031.3350542.
Verduyn, Philippe et al. (2017). "Passive Facebook usage undermines affective well-being: Experimental and longitudinal evidence". In: *Journal of Experimental Psychology: General* 146.2, p. 202.
Vicol, Paul et al. (2018). "MovieGraphs: Towards Understanding Human-Centric Situations from Videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Walther, Joseph B et al. (2008). "The role of friends' appearance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep?" In: *Human communication research* 34.1, pp. 28–49.
Wang, Hee Lin and Loong-Fah Cheong (2006). "Affective understanding in film". In: *IEEE Transactions on circuits and systems for video technology* 16.6, pp. 689–704.
Wang, Junke et al. (2021). "M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection". In: *arXiv preprint arXiv:2104.09770*.
Wang, Kai et al. (2018). "Cascade attention networks for group emotion recognition with face, body and image cues". In: *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 640–645.
Wang, Qi et al. (2014). "Video inter-frame forgery identification based on optical flow consistency". In: *Sensors & Transducers* 166.3, p. 229.

Wang, Yilin and Baoxin Li (2015). "Sentiment analysis for social media images". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, pp. 1584–1591.
Watts, Duncan J, David M Rothschild, and Markus Mobius (2021). "Measuring the news and its impact on democracy". In: *Proceedings of the National Academy of Sciences* 118.15.
Wei, Zijun et al. (2020). "Learning Visual Emotion Representations From Web Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13106–13115.
Welling, Max (2005). "Kernel canonical correlation analysis". In: *Depertment of Computer Science University of Toronto, Canada*.
Whitehill, Jacob et al. (2009). "Toward practical smile detection". In: *IEEE transactions on pattern analysis and machine intelligence* 31.11, pp. 2106–2111.
Wiley, Norbert (2003). "Emotion and film theory". In: *Studies in Symbolic Interaction* 26, pp. 169–190.
Wooldridge, Michael and Nicholas R. Jennings (1995). "Intelligent agents: theory and practice". In: *The Knowledge Engineering Review* 10.2, pp. 115–152. DOI: 10.1017/S0269888900008122.
Wu, Yue, Wael Abd-Almageed, and Prem Natarajan (2018a). "Busternet: Detecting copy-move image forgery with source/target localization". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 168–184.
— (2018b). "long2019coarse". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1907–1915.
Wu, Yuxing et al. (2014). "Exposing video inter-frame forgery based on velocity field consistency". In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 2674–2678.
Xiao, Tete et al. (2018). "Unified Perceptual Parsing for Scene Understanding". In: *European Conference on Computer Vision*. Springer.
Xiong, Caiming, Victor Zhong, and Richard Socher (2018). *Dynamic Coattention Networks For Question Answering*. arXiv: 1611.01604 [cs.CL].
Xu, Bo et al. (2022a). "MET-Meme: A Multimodal Meme Dataset Rich in Metaphors". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2887–2899.
— (2022b). "MET-Meme: A Multimodal Meme Dataset Rich in Metaphors". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, pp. 2887–2899. ISBN: 9781450387323. DOI: 10.1145/3477495.3532019. URL: https://doi.org/10.1145/3477495.3532019.
Yamamoto, Kyoko and Naoto Suzuki (2006). "The effects of social interaction and personal relationships on facial expressions". In: *Journal of Nonverbal Behavior* 30, pp. 167–179.
Yan, Sijie, Yuanjun Xiong, and Dahua Lin (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.
Yang, Songfan et al. (2014). "Zapping index: using smile to measure advertisement zapping likelihood". In: *IEEE Transactions on Affective Computing* 5.4, pp. 432–444.
Yang, Xin, Yuezun Li, and Siwei Lyu (2019). "Exposing deep fakes using inconsistent head poses". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8261–8265.
Yang, Zichao et al. (June 2016). "Hierarchical Attention Networks for Document Classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: https://www.aclweb.org/anthology/N16-1174.

Ye, Keren and Adriana Kovashka (2018). "Advise: Symbolism and external knowledge for decoding advertisements". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 837–855.

Yeh, Hengchin et al. (2008). "Composite agents". In: *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 39–47.

Yen, C (2017). "Exploring user's intention to post photos toward social media". In: *Anais do 28th Research World International Conference, Zurich*, pp. 26–30.

Yi, Ran et al. (2020). "Audio-driven talking face video generation with learning-based personalized head pose". In: *arXiv preprint arXiv:2002.10137*.

Yi, Yun, Hanli Wang, and Qinyu Li (2018). "CNN Features for Emotional Impact of Movies Task." In: *MediaEval*.

Yoon, Seunghyun et al. (2019). "Speech emotion recognition using multi-hop attention mechanism". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2822–2826.

YouTube Video 1 (n.d.). https://www.youtube.com/watch?time_continue=1&v=cQ54GDm1eL0&feature=emb_logo. (Accessed on 02/16/2020).

YouTube Video 2 (n.d.). https://www.youtube.com/watch?time_continue=14&v=aPp5lcqgISk&feature=emb_logo. (Accessed on 02/16/2020).

Zadeh, Amir et al. (2018a). "Memory fusion network for multi-view sequential learning". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

— (2018b). "Memory fusion network for multi-view sequential learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Zadeh, AmirAli Bagher et al. (2018). "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246.

Zajonc, Robert B (1980). "Feeling and thinking: Preferences need no inferences." In: *American psychologist* 35.2, p. 151.

Zhang, Dongyu et al. (Aug. 2021). "MultiMET: A Multimodal Dataset for Metaphor Understanding". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.249. URL: https://aclanthology.org/2021.acl-long.249.

Zhang, Hanlei et al. (2022). "MIntRec: A New Dataset for Multimodal Intent Recognition". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697. DOI: 10.1145/3503161.3547906.

Zhang, Huaizheng et al. (2020). "Look, read and feel: Benchmarking ads understanding with multimodal multitask learning". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 430–438.

Zhang, Jiangning et al. (2020). "Apb2face: audio-guided face reenactment with auxiliary pose and blink signals". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4402–4406.

Zhang, Minghui, Yumeng Liang, and Huadong Ma (2019). "Context-aware affective graph reasoning for emotion recognition". In: *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 151–156.

Zhang, Qi et al. (2018). "Adaptive Co-attention Network for Named Entity Recognition in Tweets". In: *AAAI*.

Zhao, Dong-Ning, Ren-Kui Wang, and Zhe-Ming Lu (2018). "Inter-frame passive-blind forgery detection for video shot based on similarity analysis". In: *Multimedia Tools and Applications* 77.19, pp. 25389–25408.

Zhao, Shanyang, Sherri Grasmuck, and Jason Martin (2008). "Identity construction on Facebook: Digital empowerment in anchored relationships". In: *Computers in human behavior* 24.5, pp. 1816–1836.

Zhao, Yin et al. (2019). "VIDEO AFFECTIVE IMPACT PREDICTION WITH MULTIMODAL FUSION AND LONG-SHORT TEMPORAL CONTEXT". In.

Zheng, Jian et al. (2019). "Image captioning with integrated bottom-up and multi-level residual top-down attention for game scene understanding". In: *arXiv preprint arXiv:1906.06632*.

Zhou, Bolei et al. (2019). "Semantic understanding of scenes through the ade20k dataset". In: *International Journal of Computer Vision* 127.3, pp. 302–321.

Zhou, Peng, Xintong Han, et al. (2017). "Two-stream neural networks for tampered face detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 1831–1839.

Zhou, Peng, Ning Yu, et al. (2021). "Deep video inpainting detection". In: *arXiv preprint arXiv:2101.11080*.

Zi, Bojia et al. (2020). "Wilddeepfake: A challenging real-world dataset for deepfake detection". In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390.