Multimodal Continuous Emotion Analysis

Zhang Su

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2023

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

17-Aug. 2022

Date

Zhang Su

Zhang Su

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

17-Aug. 2022

.....

Date

.....

Prof. Guan Cuntai

Authorship Attribution Statement

Please select one of the following;

*(B) This thesis contains material from 4 paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 and 5 is published as Su Zhang, Chuangao Tang and Cuntai Guan, "Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition," Pattern Recognition (2022): 108833. DOI: 10.1016/j.patcog.2022.108833

The contributions of the co-authors are as follows:

- I designed the methodology, wrote the code, conducted the experiments, and wrote the manuscript.
- Chuangao Tang provided the early input for the general idea and discussed with me on the methodology.
- Prof Guan reviewed and commented the manuscript.

Chapter 4 is published as Su Zhang and Cuntai Guan, "Emotion recognition with refined labels for deep learning," in 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020. DOI: 10.1109/EMBC44109.2020.9176111

The contributions of the co-authors are as follows:

- I detailed the methodology, wrote the code, conducted the experiments, and wrote the manuscript.
- Prof Guan provided the general idea, reviewed and commented the manuscript.

Chapter 6 is published as Su Zhang, Yi Ding, Ziquan Wei and Cuntai Guan, "Continuous emotion recognition with audio-visual leader-follower attentive fusion," in IEEE/CVF Conference on Computer Vision (Workshop), 2021. DOI: 10.1109/ICCVW54120.2021.00397 and Su Zhang, Yi Ding, Ruyi An and Cuntai Guan, "Continuous Emotion Recognition using Visual-audio-linguistic information: A Technical Report for ABAW3," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (Workshop), 2022. DOI: 10.48550/arXiv.2203.13031

The contributions of the co-authors are as follows:

- I designed the methodology, wrote the code, conducted the experiments, and wrote the manuscript.

- Ding Yi proposed the feature fusion module of the proposed deep learning model. He also helped to experiment and reviewed the paper.
- Wei Ziquan and An Ruyi provided extra computational resources and helped to experiment.
- Prof Guan reviewed and commented the manuscript.

17-Aug. 2022

Zhang Su

Date

Zhang Su

Acknowledgements

I am immensely grateful to my supervisor, Prof. Guan Cuntai, for his invaluable support and guidance throughout my research journey. I am especially appreciative of the degree of independence he granted me in pursuing my research, while also upholding the highest standards of research integrity.

I would like to express my gratitude to the colleagues and friends who have supported me throughout my candidature. I am particularly grateful to Ding Yi, my outstanding partner, for his invaluable contributions to our productive collaborations and close friendship. His meticulous guidance, insightful ideas, critical feedback, and unwavering support have been instrumental in enabling me to complete my PhD program successfully. I would also like to extend my heartfelt thanks to Dr. Tang Chuangao, Dr. Chen Haifeng, and Khok Hong Jing for providing me with their code and instructions, which helped me establish my baseline. Their generosity and expertise have been invaluable to my research. I am also deeply appreciative of the assistance I received from Dr. Tushar Chouhan and Dr. Ravikiran Mane during my coursework and TA duties. Their guidance and support were invaluable to my academic progress. Finally, I would like to thank Lu Yuhao for the timely help and convenience provided during my time at NTU.

I express my sincere gratitude to Huang Zichen and my parents for their unwavering emotional support, particularly during these trying times of the Covid era. Their care and encouragement have been a constant source of strength for me. I am also deeply grateful to Feng Yuxi, my lovely girlfriend, for accepting me for who I am and introducing me to her faith. Your presence and support mean the world to me. Thank you for being there when I needed you the most.

Summary

Continuous emotion recognition (CER) involves the sequence-to-sequence regression of various emotion cues, including visual, audio, textual, or physiological. To create a reliable deep learning model for CER, it is essential to achieve temporal modeling and cross-subject generality. To address these challenges, this thesis proposes four methods that utilize the advantages of multi-modality from different perspectives.

Chapter 3 presents the first method, which serves as the foundation for the other three methods. This method feeds unimodal emotion cues and aims to achieve long-range temporal modeling for CER. Chapter 4 describes the second method, which uses the temporal and visual information in continuous labels to enhance the performance of EEG-based emotion classification. In Chapter 5, the third method performs visual-to-EEG knowledge distillation for CER. Finally, Chapter 6 presents the last method, which focuses on multimodal feature fusion for CER. The proposed methods have shown promising experimental results compared to state-of-the-art methods, validating their effectiveness.

Abstract

Emotion recognition is an increasingly popular research topic in various fields, including human-computer interaction and affective computing. Continuous emotion recognition (CER), a sub-task in this area, focuses on performing sequence-to-sequence regression on the provided emotion cues, as opposed to other research topics such as sequence-to-category emotion classification.

To create a trustworthy deep learning model for CER, it is essential to learn the long-range temporal dynamics and preserve the cross-subject generality. The reason is that emotion is a continuous event that depends on past emotional states, making it crucial to consider the dynamics over a longer time frame for a more accurate prediction. Moreover, emotion is susceptible to individual differences because it is linked to personal characteristics such as experience, mood, and personality. To tackle these challenges, we developed four approaches that utilize the advantages of long-range temporal learning and multi-modality in different ways.

Our first method, which serves as the foundation for the other three, focuses on the long-range temporal modeling for CER by utilizing unimodal emotion information. The experiment conducted using the MAHNOB-HCI database shows the superior performance of our method compared to the state-of-the-art method. Additionally, we also explore the contribution of different brain regions and EEG frequency bands towards the emotion process using a saliency map-based visualization method.

The second method proposes using the continuous labels' temporal and visual information to enhance EEG-based emotion classification. The standard configuration assigns a categorical label to each trial, ignoring the temporal variation, which may reduce the classifier's effectiveness. To overcome this limitation, a thresholding scheme is introduced to convert the emotional trace into a discretized label, allowing the training process to occur in an N-to-N manner. By discretizing the trace into three classes, the classifier can fit the features to their corresponding

three-class labels more flexibly. Experimental results show a statistically significant 3% increase in EEG-based emotion classification accuracy.

The third method trains a teacher model on the visual modality and a student model on the EEG modality, where the teacher's temporal embeddings are taken as dark knowledge for the student. By employing L1 loss and concordance correlation coefficient (CCC) loss, the student model learns to fit the teacher's knowledge and predict the continuous labels. Experimental results show that the CKD method outperforms the student model without distillation on root mean square error (RMSE), Pearson correlation coefficient (PCC), and CCC. This approach provides a promising way to leverage the complementarity of different modalities for CER.

The final method proposed in this thesis involves multimodal feature fusion for CER. Utilizing multiple modalities can disambiguate and preserve recognition robustness, improving accuracy in cases such as a crying face with joyful vocal expressions being recognized as happiness instead of sadness. The leader-follower attentive network (LFAN) is introduced to combine the learned encodings of the visual and EEG modalities using a cross-modality co-attention mechanism. The LFAN emphasizes the dominant visual modality, which is believed to have the strongest correlation with the label. Experiments on AVEC2019, MAHNOB-HCI, and AffWild2 databases demonstrate that the proposed LFAN achieves promising results compared to state-of-the-art methods.

Contents

List of Figures

List of Tables

List of Acronyms

BLSTM	Bidirectional Long Short-term Memory
CCC	Concordance Correlation Coefficient
CER	Continuous Emotion Recognition
CKD	Cross-modal Knowledge Distillation
CNN	Convolutional Neural Network
DNN	Deep Neural Network
EEG	Electroencephalogram
FACS	Facial Action Coding System
GRU	Gated Recurrent Unit
KD	Knowledge Distillation
LFAN	Leader-follower Attentive Network
LOSO	Leave-one-subject-out
LSTM	Long Short-term Memory
PCC	Pearson's Correlation Coefficient
PSD	Power Spectral Density
RAM	Random Access Memory
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
TCN	Temporal Convolutional Network
TRS	Trial-wise Random Shuffling
VRAM	Video Random Access Memory

Chapter 1

Introduction

What is emotion? There is no consensus in defining what exactly constitutes an emotion [5]. There are several different perspectives on emotions, each offering a unique approach to understanding this complex phenomenon. The Darwinian perspective views emotions as evolved responses that serve specific survival functions. According to this view, basic behaviors such as facial expressions have been shaped by natural selection as adaptive responses to the environment. These responses help organisms to survive and thrive in their respective environments. The Jamesian perspective, on the other hand, argues that bodily changes are the primary cause of emotions. James believed that automatic responses to events, such as visceral changes or expressive responses, begin immediately upon perception of those events. The resulting feelings are what we commonly think of as emotions. The cognitive perspective proposes that emotions arise not only from bodily changes, but also from cognitive processes such as appraisal. Appraisal is the process of evaluating a situation and labeling it as either positive or negative for the individual. This labeling then elicits the corresponding emotions of happiness or sadness. The social constructivist perspective takes a different approach, viewing emotions as social constructions that are shaped by cultural and societal norms. In this view, emotions can only be described and understood within the context of the society in which they are experienced. While these perspectives differ in their approaches to understanding emotions, there is evidence to suggest that they overlap and have started to converge [6].

Despite the ongoing debate surrounding the nature of emotions, it is widely accepted that emotions play a fundamental role in our daily lives. Emotions are essential to human cognition, including rational decision-making, perception, human interaction, and human intelligence [7]. At the intra-personal level, emotions prepare us for significant events such as birth, battle, and death, and coordinate various bodily systems, including perception, attention, inference, learning, memory, goal choice, motivational priorities, physiological reactions, motor behaviors, and behavioral decision making [8, 9]. At the inter-personal level, emotions convey the psychological state and intentions of the expressor, influencing the behaviors of the perceiver and regulating their interactions. Emotions also play a critical role in maintaining social harmony by coordinating individuals and organizing relationships systematically. While culture provides norms and guidelines for emotions, emotions function as important motivators of behavior and help to regulate our behaviors, reducing social complexity. In summary, despite the ongoing debate surrounding the nature of emotions, their importance cannot be overstated. Emotions play a crucial role in our daily lives, shaping our cognition, perception, and behaviors at the intra-personal, inter-personal, and social and cultural levels.

Automatic emotion recognition is a rapidly growing field of technology that seeks to understand human emotions. Once considered the stuff of science fiction, it has now garnered increasing attention in various fields, such as human-computer interaction, behavioral modeling, opinion mining, psychological health, business intelligence, and entertainment assistant. One of the sub-tasks within automatic emotion recognition is continuous emotion recognition (CER). The term "continuous" has a dual meaning. Firstly, it refers to mapping a subject's recording, such as facial expression, speech, or physiological signals, to a point in a real-valued space defined by certain emotion indicators. Secondly, it implies that such recordings are given sequentially, and the mapping is performed for every time step, producing emotional traces. This stands in stark contrast to other research topics, such as emotion classification, where the entire trial is typically categorized. The focus of this thesis is on the development of CER methods. As emotions are complex and nuanced, continuous emotion recognition techniques are necessary to capture the temporal and spatial aspects of emotional experiences accurately. The development of effective CER methods has the potential to revolutionize many fields, from healthcare to entertainment, by providing a deeper understanding of human emotions and improving the interactions between humans and technology.

1.1 Research Background

Mental disorders are prevalent and the treatment gap is a major obstacle to the well-being of communities. Population-based epidemiological studies have revealed that psychiatric problems are surprisingly common. According to the National Comorbidity Survey Replication, an estimated 9.7% of U.S. adults had a mood disorder in the past year, and an estimated 21.4% of U.S. adults will experience a mood disorder at some point in their lives. Additionally, of adults with any mood disorder in the past year, an estimated 45.0% experienced serious impairment [10]. However, despite the high prevalence of mental disorders, the treatment gap remains a significant concern. The treatment gap is defined as the percentage of people with an illness, disease, or disorder who require treatment but do not receive it [11]. It represents the absolute difference between the true prevalence and the proportion of individuals who are treated. A large multi-country survey supported by the World Health Organization (WHO) revealed that 35.5% of serious cases in developed countries and 76.9% in less-developed countries did not receive treatment in the previous 12 months [12]. The high prevalence of mental disorders and the treatment gap are significant issues that need to be addressed urgently. Failure to provide effective treatment for mental disorders can lead to a range of negative outcomes, including reduced quality of life, impaired functioning, and even suicide. Improving access to mental health services and increasing the availability of evidence-based treatments is critical to improving the well-being of individuals and communities.

Automatic emotion recognition has the potential to provide a promising solution for dealing with mental disorders and the treatment gap. By utilizing objective indicators of human emotions, automatic emotion recognition algorithms can be developed to provide easy accessibility for both patients and doctors. However, to design an effective emotion recognition system, it is crucial to consider three key characteristics: generality, continuity, and multi-modality.

Generality refers to the system's ability to recognize emotions across a wide range of individuals, regardless of personal attributes such as mood, personality, and experience. Emotion, as intrinsic as it may be for every one of us, is highly prone to cross-subject bias. Emotion is triggered by conscious and/or unconscious perception of an event and is usually associated with personal attributes such as mood,

personality, and experience. For example, physically abused children are much quicker than other children to spot the signals of anger [13]. The expression of emotion involves vocal emotional words and non-vocal cues such as facial expressions, voice intonation, and body movement, resulting from the compound effects of physiological arousal, individual feelings/behaviors, and cultural/moral regulation, etc. All of these factors could lead to a large subjective bias and degenerate representation learning. The development of a generality-oriented system requires careful consideration of the diverse nature of emotions and the variation in their expression across different individuals. By doing so, we can create an emotion recognition system that is widely applicable and can be used by various populations.

Continuity refers to the ability of the system to detect emotions continuously as they unfold. Emotions are complex and involve a dynamic interplay of cognitive, physiological, and behavioral processes. Temporally, emotions emerge through dynamic processes involving cognitive appraisal, physiological arousal, and expressive behavior. The processes are ongoing without starting and ending points, such that there is no point during an emotion episode that a certain process is not happening [14]. Emotion episodes contain not only the experience currently occurring [15], but also the attempts of change and regulation [16]. They are a compound of situation, regulation, and person factors [17]. Spatially, instead of being categorized into several basic ones, emotions should be modeled in continuous real value, to describe more subtle changes. The dimensional model [18] represents emotional states in a two-dimensional circular space, where the arousal and valence are the two dimensions. To accurately detect and classify emotions, the system must account for the continuous and evolving nature of emotions. By developing a system that is capable of continuous emotion detection, we can gain a deeper understanding of emotional experiences and provide timely interventions when necessary.

Multi-modality refers to the use of multiple cues, including visual, audio, and bio signals, to recognize emotions. Emotion is multi-modal. A variety of modalities from the subject's internal and external expressions can be used for recognizing emotions. Visual modality refer to everything our eyes can perceive from the subject. It includes facial expressions and bodily postures/gestures. Audio modality refer to the verbal or nonverbal vocal signals, either through explicit linguistic messages or implicit acoustic/prosodic messages. Bio modality are multichannel

recordings from both the central and the autonomic nervous systems [19], which include the EEG, blood perfusion, and galvanic skin responses, etc. Information from different modalities possess a complementary nature. The facial expression, as one of the most common cues, conveys universal and abundant emotional information across humans. It is expressive, but also concealable. Positive correlations of the fundamental frequency, pitch range, and speech rate towards arousal dimension, or certain spoken words such as *good* and *lovely* towards valence dimension are reported for audio modality. Biosignal carries the internal state and is difficult to consciously manipulate. The work in [3] shows that EEG features add information in addition to the facial expression for valence detection. Utilizing multi-modality information benefits CER.

1.2 Research Questions

The ultimate question posed in the thesis is whether we can improve emotion recognition performance by modeling the continuity and multimodal nature of emotions. This question is explored from three different angles in the thesis.

First, the labeling protocol of the traditional emotion classification overlooks the emotion continuity. Typically, the data collected from the subjects, be it facial expressions, audio records, or physiological signals, are labeled by the subjects themselves into basic emotions (nominal) or binary high/low intensity (ordinal). The labeling protocol takes the subjects' emotion state constant for a whole trial, simplifying the sequence prediction problem into an N-to-1 mapping task. The simplicity of the labeling protocol help to develop early emotion recognition methods. However, it contradicts the emotion theories, which we will elaborate in Section 2.3, that emotion is continuous and there is no point during an emotion episode that a certain process is not happening [14]. Therefore, we are particularly interested in exploring:

- How to embed the temporal information of emotion into the traditional emotion classification and therefore improves the accuracy.

Second, given the representative visual and EEG modalities, they have their advantages and disadvantages, as we mentioned in Section 1.1. In addition to which, the

EEG data is relatively hard to collect compared to the visual data. The cross-modal knowledge distillation (CKD) provides a promising way to utilize the complementarity for this circumstance where one modality suffers limited data. CKD has been a hot topic in many fields such as speech recognition and action recognition. However, its study on CER has yet been seen to the best of the author's knowledge. Therefore, we are the first to propose a question that:

- Can the visual modality transfer its knowledge from abundant visual data to the EEG modality and improve the EEG-based CER performance?

Third, in addition to CKD, multimodal feature fusion is another hot technique to utilize multimodal complementarity. Most fusion methods work at either the feature-level or decision-level. However, the feature-level fusion suffers the curse of dimensionality, while the decision-level fusion overlooks the complementarity across different modalities, which is against the evidence [20] in neuroscience, suggesting that multimodal integration occurs at an early stage. Therefore, we propose a question that:

- How to design an intermediate-level feature fusion method, which considers the feature complementarity and modality interactions?

1.3 Contributions of the Thesis

Based on the three questions, there are three major contributions achieved in this thesis. They are listed as follows.

To embed the temporal information of emotion into the traditional emotion classification, a thresholding scheme is designed to categorize the continuous annotation. The ordinal annotation generated from the thresholding scheme refines the categorical label so that the data from one trial are not mapped to one category but a sequence of ordinal labels. The classifier trained using the refined labels has an accuracy gain on EEG modality with statistical significance. In short:

- A thresholding scheme is designed to categorize the continuous annotation and improve the EEG-based emotion classification accuracy with statistical significance.

To transfer the visual knowledge to the EEG modality, unimodal CER models for visual and EEG modalities are designed. The visual and EEG modalities play as the teacher and student, respectively. Through the teacher-student interaction, the EEG model is improved with statistical significance. The code is available at `https://github.com/sucv/Visual_to_EEG_Cross_Modal_KD_for_CER`. In short:

- The visual-to-EEG CKD is achieved which improves the EEG-based CER Performance with statistical significance.

To model the feature complementarity and modality interactions for multimodal feature fusion, the leader-follower attentive fusion network (LFAN) is designed. It combines information from different modalities and emphasizes the leading visual modality in a manner of intermediate-level fusion. We achieved runner-up in the 3rd affective behavior analysis in-the-wild (ABAW3) challenges, and 3rd place in the 5th affective behavior analysis in-the-wild (ABAW5) challenges. The code is available at `https://github.com/sucv/ABAW3`. In short:

- The novel LFAN is designed to achieve the intermediate-level fusion, which works on the visual, audio, linguistic, and physiological modalities.

1.4 Organization of the Thesis

Chapter 2 provides an extensive overview of CER, covering various aspects of the field. It starts by discussing the physiological and neuroscience foundation of emotion and popular theories used to model emotion. This background knowledge is essential for understanding the different modalities used for CER, including facial expressions, audio signals, and physiological signals. The chapter then elaborates on the reasons for utilizing multimodality in CER, such as the complementary nature of different modalities and the ability to capture a more comprehensive representation of emotions. The problem formulation for CER is also defined, providing a clear understanding of the task and its objectives. Furthermore, the chapter provides an introduction to the CER community, including available databases and contests. This information is useful for understanding the progress of the field and the metrics used to evaluate CER performance. Finally, the chapter reviews related

works on unimodal and multimodal CER methods, providing an overview of the existing techniques used to recognize emotions in different modalities. This review of related works serves as a foundation for the subsequent chapters, which explore new approaches to improving CER performance. Overall, this chapter provides a comprehensive overview of CER, setting the stage for the rest of the thesis.

Chapter 3 serves as the baseline for our work, introducing the unimodal CER method. Three subsequent chapters present the novel methods that correspond to the contributions we have achieved.

Chapter 4 presents the proposed label correction method, which uses continuous annotations to improve emotion classification performance. This method addresses the issue of traditional emotion classification labeling protocols that overlook the continuity of emotions. By utilizing continuous annotations, the system can better capture the temporal dynamics of emotions and improve recognition accuracy.

Chapter 5 introduces the visual-to-EEG knowledge distillation method. This method transfers knowledge from visual modalities to EEG modalities, which is useful in situations where one modality suffers from limited data. By utilizing cross-modal knowledge distillation, the system can better capture the complementarity between different modalities and improve recognition accuracy.

Chapter 6 introduces the leader-follower attentive network for multimodal feature fusion. This intermediate-level feature fusion method considers the complementarity of features and modality interactions, improving recognition accuracy by integrating multimodal data at an early stage.

Finally, Chapter 7 discusses the limitations of the thesis and proposes directions for future work, providing insight into potential areas of research that could build upon the contributions of this thesis.

Chapter 2

Review of Continuous Emotion Recognition (CER)

2.1 Introduction

In this chapter, a comprehensive review of the theoretical and technical relevance of CER is presented. From a theoretical perspective, the physiological foundation of emotions is introduced, and the pathway from stimuli to emotional expressions is elaborated upon. Furthermore, several representative theories used to model emotions are introduced, providing insight into the conceptual frameworks that underlie CER research. Additionally, the different modalities of emotion are discussed, as well as the issues associated with using unimodal information for CER. The importance of involving multimodality in CER is argued for, and its potential benefits are highlighted. From a technical perspective, the CER problem is formally defined, and the major databases and contests in the CER community are introduced, providing readers with a good understanding of the progress of CER, including the focus of the topic and the evolution of evaluation metrics. Moreover, the evaluation metrics used in CER research are formally discussed. Finally, related works on both unimodal and multimodal CER methods are reviewed, presenting an overview of the existing techniques used to recognize emotions in different modalities. Overall, this review provides a comprehensive overview of the theoretical and technical aspects of CER, setting the stage for the subsequent chapters, which present novel approaches to improving CER performance.

2.2 Physiological Foundation of Emotion

The human nervous system is comprised of two components, namely the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS, consisting of the brain and spinal cord, and the PNS, comprising the nerves and ganglia outside of the brain and spinal cord, are the two main components. The structure and function of some of the parts associated with this study are first outlined in this section, followed by an elaboration of how facial expressions can result from visual and auditory stimuli.

2.2.1 Cerebrum

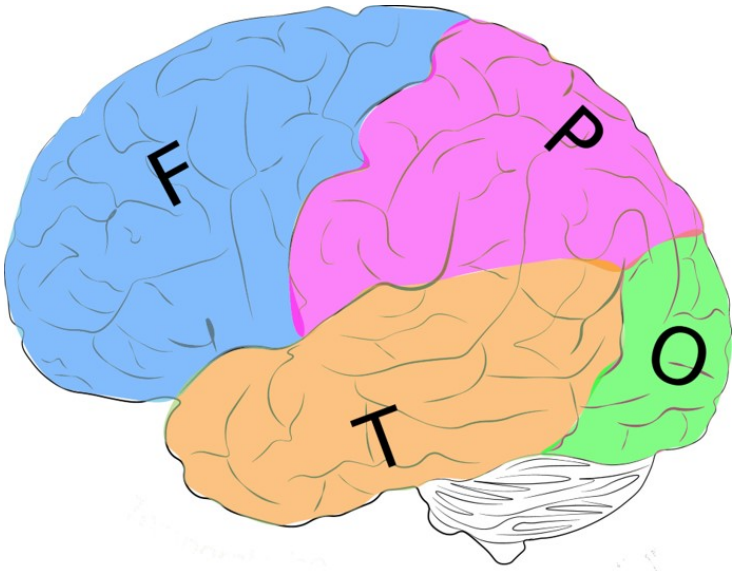

Figure 2.1: The illustration of the cerebrum. It consists of four lobes, i.e., the frontal (F), parietal (P), temporal (T), and occipital (O) lobes.

The cerebrum consists of four lobes [21], which are the frontal lobe, parietal lobe, occipital lobe, and temporal lobe, as shown in Fig. 2.1.

- The frontal lobe is involved in emotions, reasoning, motor control, and language. It contains the motor strip, which plans and coordinates movement; the Broca's area, which is basically for language production.
- The parietal lobe integrates sensory information among various modalities, such as touch, odor, pressure, pain, and temperature, etc.

- The occipital lobe interprets vision, such as color and light. The visual cortex is located in the occipital lobe which receives the electrical nerve signal from the visionary stimulus. The primary visual cortex (PVC) is one component of the whole visual cortex.
- The temporal lobe processes sound, it also interprets languages and speeches. The auditory cortex is located in the temporal lobe. The primary auditory cortex (PAC) is one component of the whole auditory cortex.

2.2.2 The Pathway from Stimuli to Facial Expressions

Vision is generated when the retina of the eye is hit by a stimulus. The photoreceptors in the retina convert the stimulus into an electrical nerve signal, which is then sent to the visual cortex, part of the optic nerve located in the occipital lobe. The signal is transferred along the optic nerves from the eyes to the primary visual cortex. When sound waves in the form of acoustic energy impinge on the tympanic membrane, the information is conducted by the cochlear nerve, part of the vestibulocochlear nerve, to the cochlear nuclei, the superior olivary nucleus, the medial geniculate nucleus, and finally, the auditory radiation to the primary auditory cortex, located in the temporal lobe [22].

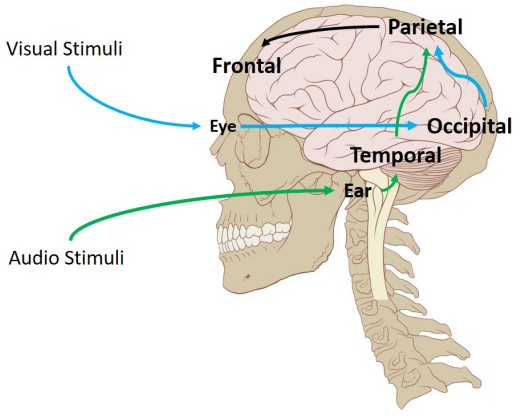

Figure 2.2: The illustration of the two-streams hypothesis. It depicts the pathway of how the visual and audio stimuli could lead to a facial expression through the two-stream hypothesis.

The information processed by the primary visual and auditory cortex is subsequently processed by the two-streams system. The two-streams hypothesis, a model of neural processing of vision and hearing [23–25], posits that when visual information exits the primary visual cortex and auditory information exits the primary

auditory cortex, they follow two pathways: the ventral stream (also known as the "what" pathway) and the dorsal stream (also known as the "where/how" pathway). The former pathway leads to the temporal lobe and is responsible for object identification and recognition, while the latter pathway leads to the parietal lobe and is involved in coordinating objects relative to the viewer, as well as with speech repetition.

The multi-sensory information generated by the two-streams system is associated with the posterior parietal cortex (PPC). The PPC acts as a bridge between the parietal lobe and the frontal lobe, and its role is to transform multi-sensory information into motor planning or motor commands. Subsequently, the primary motor cortex located in the frontal lobe contributes to generating neural impulses that are responsible for controlling facial expressions.

2.2.3 Using EEG to Reveal the Emotion

As previously introduced in Section 2.2.1, 75% of the four lobes are known as associate areas. These areas are involved in higher mental functions such as thinking, speaking, and learning, which are essential for human cognition. These complex functions are not located in precise brain areas but rather result from the synchronized activities of many brain regions [26].

Emotion is a complex function that is processed through a series of pathways. In the context of data acquisition for a CER database, the subject watches short film clips, and the visual and audio stimuli are processed in the occipital and temporal lobes. The sensory information is then integrated in the parietal lobe before being delivered to the frontal lobe, where a variety of judgments and regulations are made. Finally, the frontal lobe directs the subject's facial muscular movement, which is subsequently annotated by an expert. The expert's brain processes the facial expression in a similar pathway, repeating the process of sensory information processing, integration, and motor control.

The electroencephalogram (EEG) signals, which are potential fluctuations produced by the central nervous system, offer promising insights into the decoding of the emotion process. By placing scalp electrodes at specific locations following the 10-20 system of electrode placement, potential fluctuations in the underlying

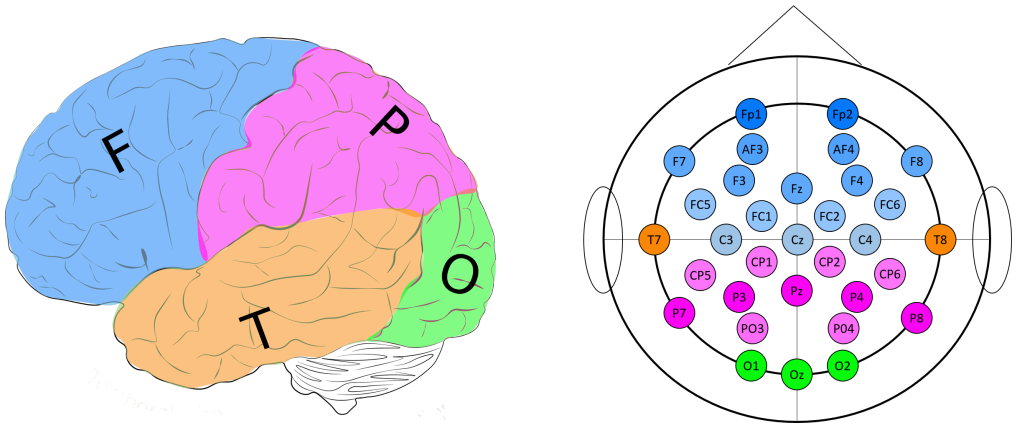

Figure 2.3: The illustration of the four brain lobes (left) and the placement of the 32 EEG electrodes (right). Our brain consists of four lobes, i.e., the frontal (F), parietal (P), temporal (T), and occipital (O) lobes. By placing scalp electrodes on specific locations following the 10-20 system of electrode placement, the potential fluctuations of the underlying cerebral regions can be measured. The correspondences between the four lobes and 32 electrodes are indicated in color.

cerebral regions can be measured, as shown in Fig. 2.3. EEG can be divided into five frequency bands, each corresponding to different mental states, reflecting brain activity. The δ wave $(0.3 - 5Hz)$ is associated with the unconscious mind and appears during anesthesia or dreamless sleep. The θ wave $(5 - 8Hz)$ is associated with the subconscious mind and memory load and appears during sleep and dreaming, during which working memory is encoded into long-term memory. The α wave $(8 - 12Hz)$ is associated with a relaxed yet aware mental state and can be reduced or disappear during external visual or auditory stimuli. The β wave $(12 - 30Hz)$ is associated with an active state of mind, particularly in the frontal lobe, and can be observed during intense focused mental activity. Compared to the "fast idle" β_1 wave $(12 - 18Hz)$, the β_2 wave $(18 - 30Hz)$ is associated with complex thought, integrating new experiences, high anxiety, or excitement. Finally, the γ wave $(> 30Hz)$ is associated with high-level cognitive brain activities or attention-intensive activities, such as perception, transmission, processing, integration, and feedback of information, as well as the processing of multimodal sensory information.

2.3 Theoretical Models of Emotion

Currently, there exist several major theories regarding a general model of emotion mechanisms. They are i) the categorical model, ii) the dimensional model, iii) the componential appraisal model, and iv) the social constructivist model.

2.3.1 The Categorical Model

The categorical model of emotions is established based on a small number of basic emotions, including happiness, sadness, surprise, fear, anger, and disgust [27]. These basic emotions are believed to be intrinsic to the human mind and can be recognized universally [28]. The categorical model was first pioneered by Darwin [29], who insisted that emotions should be understood through their survival value and following the theory of evolution.

Seminal contributions to the categorical model have been made by many researchers. Sylvan Tompkins proposed nine "affects" that are universal, bio-chemical, neurophysiological processes in the body that make or amplify triggering information [30, 31], which can be seen as an interpretation of Darwin's idea. Paul Ekman, Carroll Izard, and their colleagues have also made significant efforts to reveal the universality of certain emotions through human facial expressions. In addition to the six basic emotions proposed by Ekman, ten were identified by Izard, eight by Robert Plutchik [32], and eighteen by Frijda [33]. The number of basic emotions aside, what matters is that they are based on a set of "fundamental" or "primary" emotions, which feature a survival-related pattern of responses towards events in the world that have been selected throughout our evolutionary history [6]. All other emotions are compounds of the basic ones. To date, the categorical model, represented by Ekman's theory on universality, has been the most widely adopted approach for automatic affect recognition.

2.3.2 The Dimensional Model

The dimensional model, which goes beyond discrete emotions and attempts to conceptualize emotions in an emotion space defined by several axes, has been proposed

by some researchers. It argues that a common and interconnected neurophysiological system is responsible for all affective states [34]. According to this model, affective states are not independent components but are systematically related to each other. Emotions can be described in three dimensions. The valence dimension determines the positivity and negativity of the emotion, that is, how pleasant the emotion is. The arousal dimension determines the excitement and apathy of the emotion, that is, how activated the emotion is. Finally, the power dimension determines the sense of control over the emotion.

The circumflex model, introduced by Russell [18], is one representative model that suggests emotions are coordinated in a valence-arousal-axed circular space. In this model, the valence and arousal represent the horizontal and vertical axes of the space, respectively, with the origin representing a neutral state of emotion. Any emotional episode could be located in the continuous space of this two-dimensional plane.

2.3.3 The Componential Appraisal Model

Both the categorical and dimensional models have their advantages and disadvantages. The former cannot handle complex or blended emotions [35], whereas the continuous scale of the latter makes this possible. Despite its advantages, the dimensional model is criticized by many discrete emotion theorists due to the following facts. Describing emotions using only two or three dimensions can result in the loss of information. Moreover, some of the basic emotions, such as fear and anger, are difficult to distinguish in the dimensional space. Additionally, some emotions may not fit into the space at all, such as surprise.

The componential model, which is an extension of the dimensional model, is proposed based on the appraisal theory [36]. It argues that a continuous and recursive subjective evaluation is always ongoing to appraise the individual's internal state and the external environments. Thought and emotion are entangled. When a stimulus is perceived, a primitive urge may first be elicited, which is later "labeled" by our thought, after which we can experience the feeling of happiness or sadness from the primitive urge. According to the componential appraisal model, every

emotion is linked to a particular appraisal pattern, and the latter would then associate many attributes of the person, such as their learning history, temperament, or personality, with the stimuli the person is currently undergoing.

2.3.4 The Social Constructivist Model

In addition to the researchers who view emotions as a matter of primary biology or evolved adaptations, social constructivists argue that emotions are products of culture and learned social rules [6]. According to constructivists, emotions are not just remnants of our physiogenetic past and cannot be explained solely in physiological terms. Instead, emotions are social constructions that can be fully understood only on a social level of analysis [37].

The constructivist perspective emphasizes the social functions of emotions. For example, anger is considered a fundamental and phylogenetically wired emotion. However, according to the constructivist perspective, anger has important social functions interpersonally and socially. It is a product of a sophisticated pattern of appraisals. When person A wrongs person B by violating certain standards, B feels anger based on their knowledge and intentions [38]. For social constructivists, the role of culture is crucial in shaping emotions. Culture provides the content of appraisals and organizes emotions behaviorally, ultimately determining emotions culturally.

2.4 Emotion Modalities and Their Features

Various emotion modalities have been used in CER research. In this section, we will provide a brief overview of the major modalities that will be encountered in this thesis. These include visual, audio, linguistic, and physiological modalities.

2.4.1 Visual Modality

The most abundant emotion features may be contained in the visual modality. The direct visual information provides the first type of visual emotion features. Two representative features in this scope are the facial action units (FAU) [39] and the

facial fiducial landmarks. The FAU is derived from the facial action coding system (FACS), which is a comprehensive system based on anatomy used to describe all visually discernible facial movements. The facial expressions are decomposed into individual components of muscle movement, such as AU1 corresponding to *inner brow raiser* and AU2 corresponding to *outer brow raiser*. The facial fiducial landmarks are marks deliberately placed on the subject's face to function as a point of reference or a measure. The position to place the fiducial points should be stable, robust, easily accessible, and precisely locatable both in the real world and on the images. Eye corners or nose tips are some exemplar positions.

The low-level descriptor of feature engineering is the second type of visual emotion features. These features are derived from various methods, such as optical flow, histogram of gradients (HOG), pyramids of histograms of gradients (PHOG), local binary patterns (LBP), local binary pattern histograms from three orthogonal planes (LBP-TOP), and scale-invariant feature transform (SIFT). The pattern of apparent motion of an object from adjacent frames is known as optical flow, which generates a displacement vector to manifest the movement of points from the starting frame to the ending frame. HOG counts the occurrences of gradient orientation in localized portions of an image. PHOG is an extension of HOG, which divides the image into sub-regions at multiple resolutions and applies the HOG descriptor in each sub-region. LBP uses local spatial patterns and gray-scale contrast to describe a 2-dimensional surface. LBP-TOP is a spatiotemporal extension of LBP in the three-dimensional space. SIFT extracts key points of objects by comparing each pixel to its eight neighboring pixels in each of the neighboring scales of the difference of Gaussian images.

The third type of visual emotion feature is generated through the use of deep learning models, also known as deep features. The advancements in deep learning have led to remarkable progress in various fields, including visual comprehension. In the context of CER, deep features refer to the representations learned by a well-trained deep neural network, usually a ResNet, that has been trained on a large-scale facial recognition or emotion classification database. These deep features are learned from an image-based database, such as VGGFace2 or AffectNet, and are capable of conveying highly abstract information that may not be immediately visible to human observers. When applied to video input, these sequential deep

features can be processed by another deep neural network for temporal learning, resulting in the extraction of spatiotemporal features.

2.4.2 Audio Modality

Low-level descriptors and deep features also play a significant role in audio emotion features. The mel-frequency cepstral coefficients (MFCC), GeMAPS, and eGeMAPS are considered as low-level audio descriptors. A mel-frequency cepstrum (MFC) is formed by collecting coefficients that can efficiently characterize speakers. For instance, it can be used to recognize the speaker's cell phone model details and further the details of the speaker [40]. GeMAPS and eGeMAPS [41] are collections of low-level descriptors that are expert-selected. GeMAPS consists of 62 parameters, such as pitch, jitter, shimmer, loudness, and alpha ratio, while eGeMAPS includes an additional 26 parameters, resulting in 88 parameters in total. They are widely used as benchmark feature sets for emotion recognition studies. Additionally, the spectral centroid, which indicates the mass center of the magnitude spectrum, and the spectral flux, which defines the changing speed of the power spectrum by taking the Euclidean distance between two normalized spectra, are also used in audio emotion features. Deep features, on the other hand, are representations learned by well-trained deep neural networks (DNNs) trained on audio databases such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [**?**] and the EmoReact dataset [**?**]. The sequential deep features extracted from audio signals can also be processed by another DNN for temporal learning, resulting in spatiotemporal features.

Two audio deep features, VGGish[42] and Wav2Vec2[43], can be obtained. VGGish, a pretrained CNN model for audio representation learning, originated from the VGG model [44]. The model consists of convolution and activation layers and is designed to take in patches of log-mel spectrogram with 96×64 bins. These bins can be interpreted as a sequence of images. The number 96 represents the number of windows, and 64 denotes the number of mel-spaced frequency bins. The audio information of 960ms is contained in 96 windows, where the hop length of the window is 10ms, making this the basic unit for a representation point. The final feature produced by VGGish is a 128-dimensional vector [42].

In contrast to VGGish, which is trained in a fully-supervised manner, Wav2Vec2 makes use of self-supervised learning from a BERT transformer [45] to achieve generality across multiple languages, dialects, and domains. The speech audio is a continuous signal that lacks clear segmentation into words, thus, the authors define the basic unit as 25ms long for high-level contextualized representation learning. The self-supervision is achieved by training the model to predict masked speech units. In contrast to traditional supervised models that are trained with annotated datasets limited to certain languages, the self-supervised Wav2Vec2 model can recognize more languages.

2.4.3 Linguistic Modality

The low-level linguistic descriptors, including N-gram, syntactic N-gram [46], bag of words (BoW), and term frequency-inverse document frequency (TF-IDF) [47], among others, are included in the linguistic emotion features. N-gram is a method of tokenizing the sentence into a set of N-words by resampling the word sequence of the text using a window length of N and a hop size of 1. The syntactic N-gram considers the syntactic dependency or constituent trees instead of the straightforward tokenization of the text. BoW represents the text by counting the frequency of each word, which ignores the sentence structure, grammar, and semantic relationship of the words. TF-IDF reduces the effect of implicitly common words in the text by combining the term frequency with the inverse document frequency.

Word embeddings, which are another type of linguistic emotion features, include Word2Vec [48, 49], GloVe [50], FastText [51], and BERT [45]. Word2Vec employs shallow neural networks and two different models, namely continuous bag-of-words and skip-gram, to extract high-dimensional vectors for words, thereby discovering relationships and similarities of words in the text corpus. GloVe represents each word as a high-dimensional vector and is trained based on surrounding words over a massive corpus. Pretrained models are available in 50, 100, 200, and 300 dimensions of representation using different corpora. FastText, proposed by Facebook AI research, represents each word as a bag of character n-grams, thus capturing the morphology of words, with pretrained models available for 294 languages. BERT is a transformer-based language processing model developed by Google [52], which learns contextual embeddings of words through pretraining tasks, such as language

modeling and next sentence prediction, and can be fine-tuned on smaller datasets for specific application scenarios.

2.4.4 Physiological Modality

Physiological modalities that are representative include electrocardiogram (ECG), electromyography (EMG), blood volume pulse (BVP), galvanic skin response (GSR), electrodermal activity (EDA), and electroencephalogram (EEG), among others.

The activity of the heart is measured by the ECG signal. A low HRV can indicate a relaxed state, while a high HRV can indicate stress and frustration. EMG measures the frequency of muscular activity or tension and is correlated with negative valence emotions. BVP indicates blood flow, and an increase in negative emotions, such as anxiety or fear, can increase it. GSR measures skin conductance, and higher GSR is associated with high arousal or stress. The EEG modality records neuronal potentials that reflect the functional and physiological changes of the central nervous system, and it contains valuable psychophysiological information. Individuals with disabilities can use their EEG signals to control wheelchairs or robotic arms.

One of the key features of EEG signals is that signals from different frequency bands correspond to different mental states. Delta wave (1-4 Hz), theta wave (4-8 Hz), alpha wave (8-13 Hz), Beta wave (13-30 Hz), and gamma wave (>30 Hz) are the five frequency bands for EEG processing. The delta wave with the lowest frequency corresponds to deep dreamless sleep, whereas the gamma wave with the highest frequency usually corresponds to high-level functions that require intense attention.

The features typically extracted from EEG signals include short-time Fourier Transform (STFT), discrete Fourier transform (DFT), power spectral density (PSD), wavelet transform (WT), approximate entropy (AE), differential entropy, sample entropy, wavelet entropy, and common spatial patterns (CSP), among others. STFT, DFT, and WT are standard operations that convert the signals from the time domain to the frequency domain. PSD involves the computation of average band power, which calculates a single scalar that summarizes the contribution of a given frequency band to the overall power of the signal. The entropy-based features are used to describe the complexity of the time series.

2.4.5 Problems of Uni-modality

Carrying out CER using unimodal input has several problems. In this subsection, we first discuss the modality-specific problems, we then discuss the general problem.

Visual modality could be one of the most effective and efficient ones for CER. The foundation of the basic emotion view [28] hypothesizes that the emotion features of anger, sadness, happiness, and some other basic emotion categories own universally identifiable facial expressions across different countries and cultures. According to this view, it is feasible to infer subjects' emotions based on their facial expressions with high confidence. The majority of published studies have been guided by this hypothesis.

However, recent studies [53] show that the visual modality could be problematic in certain cases. It is due to the intra-category variability and inter-category similarity. Specifically, suppose that we have six basic emotion categories. For each category, we have ten instances of expression recorded from subjects. The instances of the same category may have a moderate to a large amount of variability. The instances of different categories could also have a non-trivial similarity. Both could deteriorate the accurate prediction of emotional states.

For the linguistic modality, it suffers from the following issues. First, currently, a speech recognition model can not truly understand the contextual relation of words and sentences, which leads to misinterpretations of what the speaker meant to say or achieve. Second, the voice inputs can divert too much from the average, in terms of the accents and local differences. For example, American English and Singapore English have a non-trivial difference in voice pattern. A Singapore speaker may confuse a speech recognition model trained by American speakers' data. Third, when the speech is mixed with loud background noise, for example, collected from a speaker in the urban outdoors or in large public spaces, it could be challenging for the speech recognition model to separate the foreground and produce effective recognition results. Moreover, the requirement of privacy and data security hinders data availability and diversity.

For the physiological modality, e.g., the EEG, it usually suffers from a low signal-to-noise ratio, the non-stationarity over time, and a limited amount of training data. EEG signals are collected using an EEG cap placed on the subject's skull. The

electrodes, which are in a limited number, such as 32, 64, or 128, record the EEG signals from specific positions following the 10-20 placement system. During the recording of EEG signals, the muscular movement, including the eyes, eyebrows, face, and head, can lead to artifacts and further deteriorate or inflate the analysis results. The limited number of electrodes also results in a small spatial resolution. The non-stationarity over time causes even the same-user EEG signals to vary between or within runs, from which the trained model may not be general enough for practical scenarios.

2.5 The Multimodal Nature of Emotion and Its Involvement in CER

2.5.1 Our Brains Are Multimodal

The world surrounding us is multimodal, we can perceive this world because our brains are also multimodal. Take speech perception as an example. It is usually regarded as a unimodal process of purely auditory. Instead, vision can influence the speech perception. In [54], an experiment was conducted to demonstrate this phenomenon. The researchers showed a short video of a young woman's talking head, in which the utterances of the syllable *ba* had been dubbed onto lip movements for *ga*. A majority reported hearing *da*. However, an accurate perception was reported by the majority if only the soundtrack or the visual-only video was provided.

Another fact is that many actions can be recognized both by their sound and by their vision, thanks to the audiovisual mirror neurons [55]. An experiment showed that the so-called mirror neurons can discharge both when monkeys perform a specific action and when they see or hear the same action performed by another individual. Intuitively, when we see or hear someone knocking on our door, or when we do it ourselves, or other people do it, our brains can always extract a single meaning, i.e., *knocking*, from all these modalities.

The above mirror effect is called the McGurk effect [54]. It shows that in daily conversation, not only speech and auditory perception, but also vision, such as gestures, eye contact, and facial expressions, all play an important role to facilitate

communication. Such effect is related to language development according to the study based on preschool and school children who manifested a weaker McGurk effect. The ability to match audio and visual speech events has been reported in infants who are under 5 months old before they start to perceive language. Further experiments [56] suggest that the audio-visual interaction begins at a very early level, i.e., the visual input could improve speech detection in a noisy environment.

2.5.2 Utilizing Multi-modality for CER

Emotions are biological states associated with the nervous system brought on by neurophysiological changes variously associated with thoughts, feelings, behavioral responses, etc [57]. Multiple modalities are greatly relevant to emotions. Though emotion recognition using a single modality has made impressive progress so far, we can still say that considerable ground is to be covered before it can be fully integrated into everyday interfaces and devices. For example, due to the limitation of recording devices, a camera used for collecting facial expression video is usually 60 Hz, and a headband used for EEG data collection usually has 32 or fewer electrodes. As a result, despite the possibility that an individual may deliberately pose an expression, facial video capture at 60 Hz may not be sensitive enough to identify microexpression. Also, the EEG data collected by only 32 electrodes can be extremely limited considering the fact that brain waves actually result from the interaction of billions of brain cells. Therefore, an AI system should also be able to interpret and reason about multimodal inputs, in order to achieve an improvement over the unimodal counterpart for CER.

According to the review [58], the multimodal settings include the following perspectives. We first enumerate them, followed by discussing the relevance in our context.

- **Representation** refers to summarizing the data by exploiting the complementarity and redundancy of different modalities.
- **Translation** refers to mapping the data from one modality to another modality. An exemplar case is to describe an image by words.

- **Alignment** refers to identifying the direct relations between different modalities. An exemplar case is to align the ingredients of a receipt to the time steps of a video cooking tutorial.
- **Fusion** refers to joining information from multiple modalities for prediction. An exemplar case is to recognize the spoken words through lip motion and speech signal.
- **Co-learning** refers to the transfer of knowledge between modalities. An exemplar case is to train a model in a modality with abundant data and then transfer the knowledge to another model in the modality with scarce data.

In these thesis, *alignment*, *fusion/representation*, and *co-learning* are explored. We start the discussion with alignment. In the conventional settings of EEG-based emotion recognition, the emotion is usually elicited by watching a visual stimulus. The stimuli may be several seconds to several minutes long. After which, the subject is required to self-report the emotional state following the well-known self-assessment manikins (SAM) [59]. The self-assessment is based on categorical labels, i.e., given the N time steps of the recordings, they are all mapped to 1 label. Such an experimental protocol has been widely accepted and adopted due to its simplicity [2, 60], however, we would like to pose one question: Does the protocol overlook the temporal continuity of the emotional states?

Fusion essentially increases the amount of available data for a better representation of learning. It takes advantage of the merit of every single modality, which brings three benefits. First, more data can result in more robust predictions [61]. Second, data from different modalities allow us to exploit the complementarity within. For example, EEG data is usually collected with a time resolution of 1024 Hz, which is far more detailed than that from a camera, and the emotion measured by EEG is also hard to be deliberately hidden. On the other hand, the facial expression is less noisy and more universal than the EEG signal, given individuals of different genders, ages, regions, and races, etc., and it has favorable spatial and depth information, though it can be consciously concealed. Fusion provides an ideal way to complement their merits while attenuating the inferiors. Third, when the data from one modality is not informative, the system can still operate. For example, in the case when a person is not speaking, an audiovisual emotion

recognition system should also be able to produce predictions though it is fed by non-spoken audio signals.

Co-learning aids the representation learning of a modality by exploiting knowledge from another modality. It is especially desirable when the data from one modality are limited in terms of low signal-to-noise ratio or annotation sufficiency. One particular characteristic of co-learning is that most often the teacher modality participates only in the training, and during the test, the student modality is the only input. Generally, the visual and audio modalities have relatively more abundant data than their physiological counterpart. And, data from some modalities are hard to acquire simultaneously due to interference. For example, collecting both the facial expression and EEG signal can be challenging, as the electromyography (EMG) caused by the facial muscular movement or the electrooculogram (EOG) caused by the eyeball movement can degrade the quality of EEG. EMG and EOG may have a non-trivial correlation with the stimuli and contribute to the prediction. Knowledge distillation provides a promising solution to circumvent this issue by separating the training and testing modalities.

2.6 Multimodal Databases and Contests of CER

The CER databases are relatively scarce compared to other fields such as video understanding and object detection. Here the representative CER databases are introduced. An overview is listed in Table 2.1.

2.6.1 HUMAINE

The HUMAINE database [62] is the first database for CER that contains continuous labels. It is created by choosing recordings from a large collection of multiple databases following some criteria, and then continuously labeling the chosen data. Range of content, multi-modality, and labeling are three priorities during the development of HUMAINE. For the range of content, the developers of HUMAINE decide to move from emotion in monologue, to emotion in sedentary interaction, to emotion in action. The data are from diverse situations such as old acquaintances sedentarily discussing TV chat shows and religious programs, subjects conducting

Table 2.1: The overview of the CER databases. For Column Modal, V denotes visual, A denotes audio, P denotes physiological, and L denotes linguistic. For Column Duration, it represents the duration of the whole database as hh:mm. For Column # Anno, it denotes how many annotators are employed for annotation. For Column Emo Dim, it denotes the labeled emotion dimension, in which V denotes valence, A denotes arousal, I denotes intensity, P denotes power, E denotes expectation, L denotes liking, and T denotes trustworthiness.

Dataset	Modal	Duration	# Anno	Emo Dim
HUMAINE [62]	VA	4:11	6	VA, I
SEMAINE [63]	VA	6:30	6	VA, I, P, E
RECOLA [64]	VA	3:50	6	VA
MAHNOB-HCI [2, 3]	VAP	5:23	5	V
SEWA [65]	VA	4:39	5	VAL
AffWild2 [66]	VA	43:00	6	VA
Muse-CaR [67]	VAL	40:12	5	VA, T

human-computer conversations, subjects taking part in outdoor activities, and subjects playing driving simulators, etc. For the multi-modality, most of the records are audiovisual data, with a small subset being physiological data. For the labeling, it discards the old labeling scheme which was at a very basic level, i.e., labeling a whole trial categorically. Rather, the label is not only at a global level, but also at the frame level, changing over time.

Overall, after the data selection, HUMAINE contains 50 clips of naturalistic and induced data, which is up to 4 hours and 11 minutes long. Both audio and visual modalities are available. 6 annotators are employed to continuously label the data.

2.6.2 SEMAINE

The SEMAINE database [63] is a large audiovisual database containing the interactions between a subject and a Sensitive Artificial Listener (SAL) agent. The SAL scenario is intended to provide fluent, sustained, and emotionally colored conversations. The interaction includes two ends. At one end are recordings of paired people engaged in emotionally colored conversations. At the other end are recordings of individuals interacting with a cartoon character that is animated by one of the parties in the first end. The first end involves a human user and human operator, and the second end involves a human user and a machine operator (the cartoon character). The operator always tries to engage or even provoke the user

to vent their emotions. Note that the data are recorded in a tightly controlled laboratory environment.

Overall, SEMAINE contains recordings as long as 6 hours and 30 minutes, in audio and visual modalities. The recorded videos are of 580×780 resolution and 50 fps. 6 annotators are employed to continuously label the data.

2.6.3 RECOLA

The RECOLA database [64] is a multimodal corpus of spontaneous collaborative and affective interactions in French. The participants are paired to have a video conference, during which they are required to complete a task requiring collaboration. In addition to the audio and visual modalities, the ECG and EDA modalities are also recorded continuously and synchronously. Note that the data are recorded in a tightly controlled laboratory environment.

Overall, RECOLA contains recordings as long as 3 hours and 50 minutes. The data are available in audio, visual, ECG, and EDA modalities. The recorded videos are of 1080×720 resolution and 25 fps. 6 annotators are employed to continuously label the data. The annotations are sampled at 25 Hz.

2.6.4 MAHNOB-HCI

MAHNOB-HCI is a multimodal database [2]recorded in response to affective stimuli with the goal of emotion recognition and implicit tagging research. It provides the synchronized recording of facial videos, audio signals, eye gaze data, EEG signals, and other physiological signals from 30 subjects. The subjects are asked to watch 20 emotional video clips, resulting in 440 trials. The video clips are between 35 and 117 seconds long. The EEG signals are acquired from 32 electrodes on the 10-20 international system. The sampling frequency is 256 Hz. The facial videos are captured at 60 fps and 780×580 resolution. For each trial, four integers ranging from 1 to 9 and self-reported by the subjects are used to label the valence, arousal, dominance, and emotional keywords, respectively.

A subset [3] of the original MAHNOB-HCI database is chosen to be continuously labeled. It contains 239 trials from 24 subjects with obvious facial expressions.

The trial number for each subject is not even. Five experts are employed for the annotation using FEELTRACE and a joystick. Only the valence is continuously labeled. The reason is that the subjects are quiet and passively watching videos, which makes the annotation of arousal, power, or expectation unavailable [3]. The continuous valence label is determined by the average of the five experts' labels and is sampled at a frequency of 4 Hz.

2.6.5 SEWA

SEWA [65] is an audiovisual, multilingual database of richly annotated facial, vocal, and verbal behavior recordings made in-the-wild. It contains audiovisual recordings of spontaneous behaviors and features an unconstrained environment with commercial webcams and microphones for data recordings. Subjects from different cultural backgrounds (British, German, Hungarian, Greek, Serbian, and Chinese) and age groups (20+, 30+, 40+, 50+, 60+) participate in the data recording. The participants are required to watch 4 advertisements, which are chosen for eliciting particular mental states. In order to reach a consistent understanding of the advertisement, the chosen ones have no dialogues and are featured with simple visuals and music. A questionnaire is required to be filled in by the participants after watching the advert, which records the participants' emotional state and sentiment. After which, paired participants are asked to discuss the adverts through a video conference. The discussion is intended to elicit certain emotional reactions to the adverts. Finally, another questionnaire is required to be filled in to record participants' emotion states and sentiments. The data are annotated in multiple ways, including facial action units, facial landmarks, vocal and verbal features, and continuous valence and arousal traces.

Overall, SEWA contains recordings as long as 4 hours and 39 minutes. The data are available in audio and visual modalities. The recorded videos are in a resolution ranging from 320×240 to 640×360, and the frame rate is between 20 to 30 fps. 5 annotators are employed to continuously label the data. The labels are sampled at 10 Hz.

2.6.6 AffWild2

The AffWild2 database [66] contains 564 Youtube videos of spontaneous facial behaviors of daily life in arbitrary conditions, unlike other databases whose scenarios are only limited to the speaking participants sitting in front of a camera. Some scenarios include subjects giving an interesting speech in ceremonies, participating in interviews, reacting to something that brings them happiness, etc. Overall, the subjects' age, ethnicity, profession, head pose, illumination conditions, and occlusions are in a wide range. Four experts annotate the videos in valence and arousal. The final annotations are determined by firstly performing the median filtering for the annotation of a video from one annotator, and then averaging the 4 median-filtered annotations. The videos are in various resolutions, and the frame rate is about 30 fps.

Overall, AffWild2 contains recordings more than 43 hours long. The data are available in audio and visual modalities. The videos have various resolutions, and the frame rate is approximately 30 Hz. 6 annotators are employed to continuously label the data. The annotations are sampled at the same frequency as the corresponding videos.

2.6.7 MuSe-CaR

The MuSe-CaR database [67] is a large, extensively annotated multimodal dataset featuring YouTube videos on car reviews. It has 40 hours of user-generated videos with more than 350 reviews and 70 host speakers. The data are extensively annotated in multiple annotation tiers in addition to dimensional emotion traces. The speakers include professional, semi-professional, and casual reviewers, with their ages ranging from their mid-20s to late-50s. For the 303 videos that are annotated, the average duration is 8 minutes. Three types of roles are employed for data annotation, which are the annotator, the auditor, and the administrator. While the annotators are in charge of labeling the data according to the subsequent instruction, the auditor is responsible to inspect the labeling quality, and the administrator assigns duties during the entire annotation process.

Overall, Muse-CaR contains recordings of 40.2 hours long. The data are available in audio, visual, and textual modalities. 5 annotators are employed to continuously label the data. The annotations are sampled at 0.25 frequency.

2.6.8 CER Contest

CER community also has several representative contest series which greatly help to boost the progress of this area. In this subsection, a brief introduction to the contest series is provided. From which we can witness some interesting adaptions in terms of data acquisition protocol and winning criterion.

For data acquisition protocol, the early CER contests, focus on the data from a controlled environment. The scene would contain a single subject mostly sitting in front of the camera, with few variations on head poses, illumination conditions, and backgrounds. The focus later shifts to in-the-wild scenarios, where the scene could be about any activities, with or without one or more subjects.

For the winning criterion, the early CER contests employ root mean square error (RMSE) or Pearson's correlation coefficient (PCC), and later it shifts to the concordance correlation coefficient (CCC). A discussion regarding the criteria property would be provided in Sec. 2.7.

2.6.8.1 AVEC series

One of the most influential contests for the CER community is the audiovisual emotion challenge (AVEC) contest. It started in 2011 as a yearly event and ended in 2019. The goal of the AVEC contest series is to provide common benchmark test sets, to bring together the audio and video emotion recognition communities, and to establish to what extent the fusion of multi-modality information is possible and beneficial.

AVEC2011 [68] used categorical labels from SEMAINE [63] database. Back then the dimensional affect recognition problem was still less explored and was usually reduced to a binary classification problem. For example, the valence or arousal would have two classes, named positive/negative, for a 2-class classification.

AVEC2012 [69] was the first contest in the AVEC series that used continuous labels and posed a regression problem of CER. The SEMAINE [63] database was used. There were two sub-challenges in AVEC2012. The fully continuous sub-challenge is for a standard CER scenario, where the level of emotion has to be predicted for every moment of the recording in the dimensional space. The word-level sub-challenge only required the participants to predict the emotion when the subject is speaking. The criterion for AVEC2012 was PCC.

AVEC2013 [70] followed its predecessor for CER using the SEMAINE database of natural dyadic interactions. In addition to which, it extended the scope to a more complex scenario called depression recognition. Both the emotion and depression recognition problems were cast as regression problems. Thus, there were two sub-challenges, namely, the affect recognition sub-challenge as a standard CER scenario using the SEMAINE [63] database, and the depression recognition sub-challenge using the audiovisual depressive language database. Note that for the latter, there was only one continuous value for one multimedia file. The criteria for AVEC 2012 are twofold. The affect recognition sub-challenge used the PCC, while the depression recognition sub-challenge used the RMSE. AVEC2014 [71] also contained one CER sub-challenge using the SEMAINE database and one depression recognition sub-challenge using the depressive language database. And the PCC and RMSE were used as the criteria.

Compared to AVEC 2014 [71], there were the following changes in AVEC 2015 [72]. First, there was only one scenario, i.e., the standard CER one as it was in AVEC2014. Second, the database employed was the RECOLA [64] database. Third, in addition to audiovisual modalities, the physiological signal, including the ECG and EDA signals, were included. In AVEC 2016 [73], the depression recognition sub-challenge was included again. And the criterion for it was the F1 score. In AVEC 2017 [74], the CER sub-challenge employed the SEWA database and also the CCC as the criterion.

In AVEC2018 [75], there were three sub-challenges. The cross-cultural emotion sub-challenge includes the German and Hungarian subjects and the collected audiovisual recordings which were an extension of the SEWA database. The participants had to train their model using the German data and test their model using the Hungarian data. The prediction was required for valence, arousal, and liking dimensions in a temporally continuous manner. And the CCC was the criterion.

The gold-standard emotion sub-challenge is a new task focusing on the generation of dimensional emotion labels. There were time-continuous labels from several annotators, the participants were required to first merge the labels as one and then trained their models using the merged labels. The RECOLA database with the audiovisual and physiological data was employed. The CCC was the criterion. The bipolar disorder sub-challenge was a new focus rather than the depression analysis from previous AVEC contests. The bipolar disorder database was used. The audiovisual data was available for a 3-class classification, i.e., mania, hypomania, and remission. The unweighted average recall, which was the average of the recall in percentage obtained in each of the three classes, was the criterion.

In AVEC2019 [76], three sub-challenges were included. In addition to the cross-cultural emotion sub-challenge and depression recognition sub-challenge, the state-of-mind sub-challenge was available. The goal of this sub-challenge is to predict the human state-of-mind in a 10-point scale from the USoM corpus [77]. Moreover, for the cross-cultural emotion sub-challenge, the database was extended with Chinese subjects, so that there were German, Hungarian, and Chinese subjects in total, where the latter were taken as the test set.

2.6.8.2 ABAW series

As we introduced, the data from AVEC series are mostly recorded in the laboratory or controlled environments. And the number of subjects, head pose variations, present occlusion, background, and illumination are all limited in diversity. Moreover, the duration of the videos is up to about 4 hours, which is rather short.

In an attempt to tackle the aforementioned limitations, the affective behavior analysis in-the-wild (ABAW) contest has been organized since 2017. By the date when writing this thesis, three contests of the series, i.e., ABAW1, ABAW2, and ABAW3, were held.

ABAW1 was held in 2020. It is aimed at the automatic analysis of valence-arousal estimation, basic expression classification, and action unit detection. It had three sub-challenges, each one addressing a respective task out of the three. It employed the AffWild2 [66] database. For the valence-arousal estimation sub-challenge, the CCC was the criterion. As for the other two sub-challenges, the F1 score and

accuracy were the criterion, as they were classification problems. Following the same configuration, ABAW2 was held in 2021, followed by ABAW3 in 2022.

2.6.8.3 Other CER Contests

Multimodal Sentiment Analysis in Real-life Media (MUSE) series is another CER contest series. It aims to provide a benchmark for the fusion of audiovisual and language modalities. In addition to the audiovisual emotion recognition, it also included the sentiment analysis perspective and contributed to bringing the two communities together.

MuSe2020 [78] had three sub-challenges, the multimodal sentiment in-the-wild (MuSe-Wild) sub-challenge, the multimodal emotion target engagement (Muse-Topic) sub-challenge, the multimodal trustworthiness (MuSe-Trust) sub-challenge. The MuSe-Wild sub-challenge was in a standard CER configuration which aimed to predict the level of valence-arousal in a time-continuous manner from audio-visual recordings, and employed the CCC as the criterion. The MuSe-Topic was labeled categorical so that the criterion was the weighted sum of unweighted average recall and F1 score. The MuSe-Trust sub-challenge used the same setting as in Muse-Wild. The MuSe-Car database was employed.

MuSe2021 [67] had four sub-challenges, the multimodal continuous emotions in-the-wild (MuSe-Wilder) sub-challenge, the multimodal sentiment in-the-wild classification (MuSe-Sent) sub-challenge, the multimodal emotional stress (MuSe-Stress) sub-challenge, and the multimodal physiological-arousal (MuSe-Physio) sub-challenge. Except for the MuSe-Sent sub-challenge which was a classification task and employed the F1 score as the criterion, the other three sub-challenges all followed a standard CER configuration, with the CCC being the criterion. MuSe-Wilder and Muse-Sent employed the MuSe-Car database, while MuSe-Stress and MuSe-Physio employed the Ulm-TSST DATABASE, which required the subjects to give a free speech task under a highly stress-induced environment, following the TSST protocol [79].

By the time when writing this thesis, MuSe2022 is hosting for participation. It has three sub-challenge, the MuSe-Humor sub-challenge, the MuSe-Reaction sub-challenge, and the MuSe-Stress sub-challenge. MuSe-Humor utilizes the Passau-SFCH database. It contains an audiovisual recording of 10 football coaches who

would express humor during press conferences. MuSe-Humor employs a two-dimensional humor model [80] in the humor style questionnaire, and the label was binary, indicating if the subjects' communication is humorous or not frame by frame. The area under curve is taken as the criterion. MuSe-Reaction utilizes the Hume-Reaction database, which contains more than 70 hours of audiovisual data regarding human emotional reactions. The database is labeled categorically into seven basic emotions, and the PCC is employed as the criterion. And MuSe-Stress uses the same configuration as in MuSe2021.

2.7 Evaluation Metrics for CER

CER is basically a sequence prediction task. The agreement between the prediction sequences from a deep learning model and the gold standard sequence, or, the ground truth, is needed to be measured. Generally, the root mean square error (RMSE), Pearson's correlation coefficient (PCC), and concordance correlation coefficient (CCC) are employed. In this section a formal definition is provided for the three metrics, followed by a discussion regarding their properties.

Given the prediction $\hat{\mathbf{Y}} = \{\hat{y}_i\}_{i=1}^N$ from a trained deep learning model, and the annotation $\mathbf{Y} = \{y_i\}_{i=1}^N$ from another independent observer, our goal is to quantify the agreement, or the consistency in-between.

The root mean square error (RMSE), due to its simplicity, is one of the most popular distance metrics. It is formulated as:

$$\epsilon(\hat{\mathbf{Y}}, \mathbf{Y}) = \left\| \frac{\hat{\mathbf{Y}} - \mathbf{Y}}{N} \right\|^2 = \sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N}}. \tag{2.1}$$

The Pearson's correlation coefficient (PCC) is formulated as:

$$\rho_p(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sigma_{\hat{Y}Y}}{\sigma_{\hat{Y}}\sigma_Y} = \frac{\sum_{i=1}^N (\hat{y}_i - \mu_{\hat{Y}})(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \mu_{\hat{Y}})^2}\sqrt{\sum_{i=1}^N (y_i - \mu_Y)}}, \tag{2.2}$$

where σ_{XY} is the covariance, σ_X and σ_Y are the variances, and μ_X and μ_Y are the means of $\mathbf{X}$ and $\mathbf{Y}$.

The concordance correlation coefficient (CCC) is formulated as:

$$\rho_c(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{2\sigma_{\hat{Y}Y}}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 - (\mu_{\hat{Y}} - \mu_Y)^2}. \tag{2.3}$$

RMSE has no bound so unless a further magnitude comparison is conducted can any meaningful interpretation be made. It also fails to measure the correlation. Suppose that the two sequences to measure are placed close enough, their RMSE could be small though they may have very different trends. PCC can measure the correlation of two vectors. However, it cannot distinguish the identity relationship, or, the accuracy, because the PCCs of one vector against itself with and without an offset are equal. CCC, on the other hand, penalizes any deviation from the identity relationship while capturing the linear relationship thanks to the terms to account for the mean difference and covariance, respectively. In short, CCC owns the best credit out of the three metrics.

The machine learning community has grown an increasing interest in utilizing CCC as the performance measure of choice. For example, the AVEC contest series which we have introduced can show the trend. In the early series, the AVEC utilizes RMSE and PCC as the winning criteria, whereas in the late series it is the CCC that plays the role. In this thesis, the CCC is the main criterion for evaluating the CER methods.

Table 2.2: The relevant papers for CER. The following acronyms are from the table columns. Cita.: citation. Pub.: publication title. Mod.: modality, containing V (visual), A (audio), L (linguistic), T (thermal), and P (physiological). TDN: time-delay networks. 1DCNN: 1-dimensional convolutional neural networks. AE: auto-encoder. GRU: gated recurrent unit. LSTM: long short-term memory recurrent neural networks. CvLSTM: convolutional LSTM. DBN: deep belief networks. Att: attention. Transf: transformer. Non-DL: non-deep learning methods, i.e., the conventional methods. SNN: spiking neural networks. SSL: semi-supervised learning. KD: knowledge distillation. ADV: adversarial learning. DA: domain adaption. MT: multitask learning. USL: unsupervised learning. LC: label correction. The following acronyms are from the table columns. A: audio. T: thermal. P: physiological. V: visual. AVEC: audiovisual emotion challenge. ABAW: affective behavior analysis in-the-wild workshop. CVPR: IEEE international conference on computer vision and pattern recognition. MM: ACM conference on multimedia. TAC: IEEE transactions on affective computing. TIP: IEEE transactions on image processing. TC: IEEE transactions on cybernetics. TPAMI: IEEE transactions on pattern analysis and machine intelligence. TMM: IEEE transactions on multimedia. IF: information fusion. IVC: image and vision computing. STSP: IEEE selected topic on signal processing. ICASSP: IEEE international conference on acoustics, speech, and signal processing.

Cita.	Pub.	Year	Mod.	Model													Training Setting						
				TDN	1C	2C	3C	AE	GRU	LSTM	CvLSTM	DBN	Att	Transf.	Non-DL	SNN	SSL	KD	ADV	DA	MT	USL	LC
[81]	TAC	2011	AL												✓								
[82]	TAC	2011	VA							✓					✓								
[83]	AVEC	2014	VA												✓								
[84]	AVEC	2014	VAL												✓								
[85]	TAC	2014	V													✓							
[86]	TPAMI	2014	VA												✓								
[87]	AVEC	2015	VAP							✓													
[88]	AVEC	2015	VAP							✓													
[89]	TAC	2015	VA																				✓
[90]	TAC	2015	A									✓											
[91]	TC	2015	VAL												✓								
[92]	AVEC	2016	VAL			✓				✓													
[93]	MM	2016	VAL							✓													
[94]	TPAMI	2016	V																				
[95]	ICAASP	2016	VA												✓								
[96]	TC	2016	V	✓																			
[97]	MM	2017	VA							✓											✓		
[98]	IVC	2017	VA							✓					✓			✓					
[99]	ICASSP	2017	A							✓					✓			✓					
[100]	AVEC	2017	VAL			✓				✓											✓		
[101]	STSP	2017	VA			✓				✓													
[102]	AVEC	2017	VAL							✓													
[103]	ABAW	2017	V			✓															✓		
[104]	TAC	2017	A												✓								
[105]	ICAASP	2017	A												✓								
[106]	AVEC	2018	VAL																				
[107]	AVEC	2018	VAL							✓										✓			
[108]	AVEC	2018	VAL							✓													
[109]	ICAASP	2018	-																				✓
[110]	ICAASP	2018	A			✓				✓													
[111]	ICAASP	2018	V				✓				✓												

Table 2.2 Continued.

[112]	TAC	2018	A											✓								
[113]	AVEC	2019	V		✓	✓																
[114]	FG	2019	V			✓																
[115]	IF	2019	VL			✓							✓							✓		
[116]	AVEC	2019	VAL															✓				
[117]	AVEC	2019	VAL		✓	✓				✓												
[118]	TMM	2019	A								✓											
[119]	TAC	2019	V			✓		✓													✓	
[120]	CVPR	2020	V				✓															
[121]	ABAW	2020	VA			✓	✓													✓		
[122]	ABAW	2020	V			✓				✓										✓		
[123]	ICASSP	2020	VA							✓										✓		
[124]	ICASSP	2020	VA											✓								
[125]	TAC	2020	L												✓							
[126]	TIP	2020	VT			✓				✓			✓									
[127]	ABAW	2021	VA		✓	✓																
[128]	ABAW	2021	VA			✓										✓				✓		
[129]	ABAW	2021	V			✓											✓			✓		
[130]	ABAW	2021	V			✓														✓		
[131]	MUSE	2021	VAL			✓				✓										✓		
[132]	MUSE	2021	VAL			✓		✓														
[133]	ABAW	2021	V			✓			✓													
[134]	IF	2021	VAL			✓				✓			✓									
[135]	CVPR	2021	V			✓																
[136]	CVPR	2021	V							✓			✓									
[137]	TAC	2021	A			✓															✓	
[138]	TMM	2021	VA		✓	✓								✓								
[139]	TMM	2021	V							✓												
[140]	TIP	2021	V			✓				✓												
[141]	MUSE	2021	AL							✓												✓
[142]	ABAW	2022	V			✓			✓													
[143]	ABAW	2022	VAL		✓	✓																
[144]	TPAMI	2022	VA						✓													
[145]	ABAW	2022	VA			✓				✓												
[146]	ABAW	2022	VA			✓			✓								✓			✓		
[147]	IF	2022	VP				✓															

2.8 Related Works

In this section, we conduct a literature review of CER. The relevant papers are collected from several major academic online platforms including the SemanticScholar, IEEE Xplore, ScienceDirect, and ACM Library. The papers are published either in top journals (including TPAMI, TC, IF, TMM, TAC, TIP, and IVC), top conferences (including CVPR, ICASSP, MM), or the contests (including AVEC, ABAW, and MUSE) we introduced in Sec 2.6.8. Accordingly, 67 papers are collected and summarized in Table 2.2.

Overall, we see that in terms of modalities, 55.2% of methods are multimodal. 79.1% methods involve the visual modality, 62.7% methods involve the audio modality, and 26.9% methods involve the linguistic modality. In terms of computational models, 43.3% methods use 2-dimensional CNN, 37.3% methods use LSTM, and 16.4% methods use the combined CNN-LSTM models. In terms of training settings, a majority of the methods are trained in supervised settings for valence-arousal prediction.

2.8.1 Unimodal CER methods

2.8.1.1 Visual-based CER methods

[96] develops a two-stage automatic system for expression-based CER. The time-delay neural network is employed to model the consecutive frames and exploit the slow-changing dynamics between emotional states. [85] proposes a downsampling method with a dynamic sampling rate for facial expression-based CER. In which, an individual with idle and active expressions would allocate fewer and more frames for data analysis. [94] develops a vector machine, called the doubly sparse relevance vector machine, for facial-based CER. The proposed machine enforces double sparsity by jointly selecting the most relevant training examples and the most important kernels corresponding with relevant facial parts. [108] employs the 3-dimensional convolutional neural network with LSTM, i.e., the ConvLSTM, to merge feature extraction and regression into a unified system. [113] takes head pose and eye gaze in addition to facial expression for CER. The first two features guide the facial features through an attention mechanism for CER. [114] employs tucker

tensor regression to model the continuous emotion in naturalistic settings. The tucker tensor regression model allows to capture of multimodal data structure and reduces the total parameter number. [120] unifies low-rank tensor decompositions onto MobileNet for efficient multidimensional convolutions. The proposed is able to be generalized from 2D images to videos thanks to the higher-order transduction. [122] applies data balancing techniques and multitask learning for CER. It first trains a teacher model to perform the basic emotion classification, action unit classification, and valence-arousal prediction simultaneously, later the predictions produced by the teacher are taken as the soft labels for student training. [129] trains a unified CNN-GRU cascade network to perform facial expression classification and valence-arousal prediction. The knowledge distillation technique is employed, which uses both the ground truth labels and soft labels from the trained teacher to train the student model. [130] proposes a multitask streaming network based on the assumption that the basic emotions, action units, and the valence-arousal are intrinsically associated with each other. An advanced facial expression embedding is used as the prior knowledge to aid the learning. [136] presents a time-series emotion prediction method for multimedia content. It models the temporal causality using attention-based methods and Granger causality. Facial features, scene understanding, visual aesthetics, action description, and movie script are employed to obtain an affect-rich representation for CER. [137] introduces a CNN-based network to simultaneously align and predict labels in an end-to-end manner. The proposed network is a stack of convolutional layers followed by an aligner network that aligns the speech signal and emotion labels. [112] formulates speaker variability in terms of probability distributions in both feature and model spaces. Two compensation techniques based on partial least square dimensional reduction and feature mapping are proposed. [142] develops a facial expression-based method by combining the GRU and transformer for valence-arousal prediction. [103] employs the action unite to estimate the valence-arousal intensity, based on the assumption that both of them are entangled in the same emotional space. [119] develops a novel algorithm for 2D face frontalization, a novel frequency-domain convex optimization algorithm for unsupervised training, and an extended Kalman filtering-based algorithm for affect estimation. [139] proposes to learn rich affective state dynamics by defining the affective state in terms of valence, arousal, and their higher-order derivatives. It combines an RNN and a Bayesian filter and is able to handle high-dimensional observations and efficient optimization in an end-to-end fashion. [140] proposes a

framework for extracting 3D facial spatiotemporal features from monocular image sequences using an extended 3D morphable model. An LSTM is used to evaluate the extracted features on multiple tasks.

2.8.1.2 Audio-based CER methods

[99] proposes a prediction-based learning framework for speech-based CER. The support vector regression (SVR) and bidirectional long short-term memory recurrent neural network (BDLSTM-RNN) are used and concatenated forming a united cascade framework. [104] introduces a CER system based on ensembles of single speaker regression models. The emotion is estimated by combining a subset of the initial pool of single speaker regression models, which are the most concordant ones. [90] applies multi-task learning to leverage valence and arousal information for audio-based CER using the deep belief network. The classification of the categorical emotion and the prediction of valence and arousal are the two tasks involved in the combined loss function. [105] uses the phone log-likelihood ratio features to index valence and arousal in a pairwise low/high framework for speech-based CER. A multi-stage staircase regression framework that enables fusion at three different stages is also designed. [110] employs CNN and LSTM to model the contextual information of the speech data for CER. [125] investigates the phonetic features on several widely used corpora, including RECOLA [64] and SEMAINE [63] to explore the acoustic space partitioning information and phonetic content. [118] proposes a new bidirectional convolutional recurrent sparse network for music-based CER. The sequential-information-included affect-salient features are learned from the 2D music spectrogram. [133] employs a CNN-GRU architecture for valence-arousal prediction. [135] proposes a CER framework with three key novel components, including a global latent variable model, the temporal context modeling via task-specific predictions in addition to features, and the smart temporal context selection.

2.8.2 Multimodal CER Methods

[81] proposes linguistic analysis approaches for CER. A feature combination of character n-gram and audio low-level descriptors is explored and fused. [82] fuses facial expression, shoulder gestures, and audio features for CER. The output-associative

fusion framework incorporates correlations and covariances between the emotion dimensions. The BLSTM and SVR are employed and compared. [84] uses multimodality to account for context during the frame-wise inference and linearly fuse the outcomes from the audio-visual input. [86] proposes a novel generative model, which discovers temporal dependencies on the shared and individual spaces. It also introduces a latent warping process to counter temporal lags. [87] extracts multiple low-level descriptors in visual, audio, and physiological modalities for feature selection. The BLSTM is used for valence-arousal prediction. [88] investigates the ϵ-insensitive loss and temporal pooling for CER. The first one is claimed to be more robust for the label noises and the second one enables temporal modeling in the input features. [89] proposes a label correction method by modeling the reaction lag of evaluators. [91] develops the temporal Bayesian fusion for CER combining the visual, audio, and lexical modalities. It integrates the temporal prediction model prior to the Bayesian fusion and uncertainties about the unimodal predictions. [93] utilizes the LSTM network to capture the temporal dependencies. The attention mechanism is employed so that the model can dynamically pay attention to salient modalities at each time step. [95] proposes a mixture of experts-based fusion model to dynamically combines audiovisual information. The expectation-maximization algorithm is used to model the emotional dynamics. [97] proposes a soft prediction framework to provide a human-like emotion prediction. The uncertainty is modeled to indicate the perception and inter-rater disagreement. [98] proposes the strength modeling, which includes two models in a hierarchical manner. The strength information of the first model is joined with the original features, thus expanding the feature space for the successive model to learn from. [100] explores and fuses multiple low-level features and deep features from the visual, audio, and linguistic modalities. And the LSTM and SVR are employed and compared. [101] utilizes CNN to extract visual and speech features. An LSTM is then used to combine the features for the valence-arousal prediction. [115] proposes the multimodal data fusion method to combine the visual and audio data. The temporal attention filters are proposed to align the visual and audio data, the aligned data are then combined in the common space. [106] proposes an interesting method to exploit the difficulties in learning to boost the machine learning process. The reconstruction error is taken as the difficulty, which is then used with the original features to update the model. [107, 116] investigate different interaction methods among modalities to imitate the real interaction patterns in daily lives. [117] proposed the

CNN-TCN architecture to extract features from different modalities. The features are then fed to BLSTM for CER. [121] proposed a two-stream CER model, where the video frames and audio spectrum are taken as the input. A TCN is employed for temporal learning. [126] present a novel method fed by color, depth, and thermal recordings. The model learns spatiotemporal attention volumes to robustly recognize the valence-arousal according to the attention-boosted feature volumes. [127, 143] propose a leader-follower co-attention network for multimodal fusion from the visual, audio, and linguistic input. [128] designs a multitask learning method to learn three types of facial-based representations simultaneously. After which, a teacher model is used to produce soft labels from the unlabeled data for student training. [134] utilizes the transformer-based network and the high-resolution network to process the visual modality. Together with the audio modality, the embeddings are combined via an LSTM network. [138] developed a transformer encoder with multimodal multi-head attention for CER, which progressively refine the inter- and intra-modality temporal dependency. [141] proposes a label correction method called the rater aligned annotation weighting to align the annotations in a translation-invariant way. [144] proposes a novel fusion framework, which first learns latent distributions over audiovisual space, and then constrains the variance vectors of each modality in order to force them to represent the amount of information with respect to emotion recognition. [146] uses deep ensembles to capture uncertainty for basic emotions, action units, and valence-arousal scores. The iterative self-distillation technique is employed to improve the model performance alternative teacher and student role-playing. [147] proposes a multimodal approach to combine the EEG signals, eye modality, and face modality for CER.

Chapter 3

Unimodal CER with Temporal Modeling

This chapter [1] aims to establish the whole pipeline for unimodal CER, which is the foundation of this thesis. The downstream expansions upon the unimodal CER are to be presented later in this thesis, including alignment, co-learning, and fusion, as we introduced in Section 2.5.2.

3.1 Introduction

CER is an N-to-N sequence prediction problem. A Unimodal CER system employs sequential emotion cues from only one modality and produces sequential prediction in the continuous space axed by valence and arousal. The pipeline usually contains data preprocessing, model training, and postprocessing. Two modalities are taken into consideration, i.e., the visual and EEG modalities. Other modalities, such as the audio and linguistic modalities, will be involved in a peripheral role in the multimodal fusion.

A reliable deep learning model for the unimodal CER should be able to model the temporal dynamics. CER aims to understand the unfolding emotion, an ongoing causal event with temporal dependencies. Specifically, the facial expressions over the time span $T = \{0, 1, \ldots, t\}$ are a composition of t facial video frames, each

[1] The work in this chapter has been published in [4].

having its own valence and arousal values. The facial frame at time step t is not only the direct successor of that at $t-1$, but also the effect of the emotional roller coaster over all the predecessors. A more reliable prediction could be made by considering the emotional dynamics over a large time window.

To tackle the issue of temporal dynamic learning, we resort to large window resampling. The resampling window determines the length of context the model draws to make a prediction for each time step. A large window containing more time steps has increased global expressiveness and is favored by long-term dependency learning. Similar to mini-batch learning with a large batch size, taking longer sequences as inputs improve the generality as the model parameters only update once per sequence.

With this in mind, the architecture of our unimodal CER model is designed as a cascade CNN-TCN network. The reason for the choice is explained as follows. Generally, there are two fundamental frameworks of neural networks for visual-based emotion recognition: (i) the cascade spatiotemporal architecture and (ii) the standalone architecture. Type (i) usually contains a CNN to extract spatial information, from which the temporal information is obtained by using temporal models such as Time-delay, recurrent neural networks (RNN), long short-term memory networks (LSTM), or TCN. Type (ii) combines the two separated steps into one and extracts the spatiotemporal feature using a unified model like the 3D-CNNs.

We choose Type (i) due to the following facts. First, a 3D-CNN model[148] usually has considerably more parameters than 2D-CNNs due to the extra kernel dimension, and therefore requires more data and a longer time to train. However, 3D-based emotion recognition databases [149] are typically based on posed behavior with a few subjects, little diversity, and limited continuous labels. By contrast, there are a large amount of 2D-based facial image or emotion databases, such as MS-CELEB-1M [150], VGGFace2 [151], which are more diverse and determinant. Though there are abundant 3D video understanding databases that might be available for self-supervised or semi-supervised pretraining of a potential 3D-CNN-based emotion recognition model, the techniques involved are still a hot research topic. Second, 3D-CNNs alone may not be suitable to capture long-range temporal dependencies. As we mentioned before, CER requires mapping a composition of complex one-actions with varied intensity and order to a sequence

of continuous labels. However, most 3D-CNN-based networks are designed for at most 128 time steps [152], whereas an exclusive temporal model can easily exceed this limit.

This chapter is structured as follows. First, the methodology is formulated mathematically, followed by the demonstration of the model architecture. Next, the implementation details, including the database, feature extraction, and data partitioning, are introduced. Next, the settings of parameters, hyperparameters, and training are elaborated. After which, the results against the state-of-the-art CER method are reported. Finally, we visualize the skull saliency map for each band and subject, based on the trained student model and the peak response mapping [153] (PRM), as we are also particularly interested in revealing the contribution of each brain lobe and the band of EEG towards the emotion process (as we introduced in Section 2.2.3).

3.2 Definition of CER and Problem Formulation

Based on the aforementioned theories of emotion, we can therefore define the continuous emotion recognition (CER) termed in this thesis. Spatially, CER follows the dimensional model, or to be more specific, the circumflex model [18], to describe emotions in the continuous space axed by valence and arousal. Temporally, CER follows the componential appraisal model, to repetitively carry out the appraisal process at each time step of the given data.

Now let us formally formulate the CER problem. Suppose that we are given synchronous sequences of emotional features $\mathbf{X} = \{\mathbf{X}^{(m)}\}_{m=1}^{M}$ from M modalities, where $\mathbf{X}^m \in \mathbb{R}^{T \times d_m}$, and the corresponding ground truth $\mathbf{Y} = \{y_1, y_2, \ldots, y_T\}$, where $\mathbf{Y} \in \mathbb{R}^T$. In each modality, the sequence $\mathbf{X}^k = \{x_1^{(m)}, x_2^{(m)}, \ldots, x_T^{(m)}\}$ contains T samples, where T denote the length of the resampling window. Our goal is to find a function $f : \mathbf{X} \to \mathbf{Y}$ so that its output $\hat{\mathbf{Y}} = f(\mathbf{X}) \in \mathbb{R}^T$ minimizes some loss $L(\mathbf{Y}, \hat{\mathbf{Y}})$ against the ground truth $\mathbf{Y}$.

3.3 Methodology

Suppose that we are given one sequence of emotion cue $\mathbf{X} = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{T \times d}$, and the corresponding ground truth $\mathbf{Y} = \{y_1, y_2, \ldots, y_T\} \in \mathbb{R}^T$. The sequence $\mathbf{X}$ contains T samples, each of which is d-dimensional. Note that T also denotes the length of the resampling window. Our goal is to find a function $f : \mathbf{X} \rightarrow \mathbf{Y}$ so that its output $\hat{\mathbf{Y}} = f(\mathbf{X}) \in \mathbb{R}^T$ minimizes some loss $L(\mathbf{Y}, \hat{\mathbf{Y}})$ against the ground truth $\mathbf{Y}$. Note that the causal constrain is applied, which requires that y_t depends only on $x_1, \ldots, x_t$ and not on any future inputs $x_{t+1}, \ldots, x_T$.

3.3.1 Temporal Modelling

To learn the temporal dependency of the given sequences, some feature extractors b are firstly used upon the raw input data, updating $x_t \leftarrow b(x_t)$ for each sample, so that their representability is increased. Depending on the data modalities, the actual extractors b are varied. For example, a convolution neural network (CNN) backbone trained on a large-scale 2D face database could be used to extract the deep facial feature, and the average band power can be used as the low-level EEG feature, etc.

Several deep neural networks are capable of temporal modelings, such as the long-short term memory (LSTM) and the TCN. Systematical comparison [154] demonstrated that TCNs convincingly outperform recurrent architectures across a broad range of sequence modeling tasks. With the dilated and casual convolutional kernel and stacked residual blocks, the TCN is capable of looking very far into the past to make a prediction. Therefore, we employ the TCN as our temporal model.

A typical TCN consists of a stack of 1D dilated convolutions, and residual connections, which are formulated as:

$$TCN(\mathbf{X}) = Activation(\mathbf{X} + \sum_{i=0}^{k-1} f(i) \cdot \mathbf{X}_{s-d \cdot i}), \tag{3.1}$$

where k, s, and d denote the kernel size, stride, and dilation, respectively. $s - d \cdot i$ stands for the direction of the past. The receptive field is therefore determined by

$$1 + N_{pair} \cdot (k-1) \cdot N_{stack} \cdot \sum_i d_i, \tag{3.2}$$

where N_{pair} denotes the number of the paired convolutions and residual connections in one TCN block, and N_{stack} denotes how many TCN blocks are available for one TCN layer. Throughout this paper we employed $N_{pair} = 2$ and $N_{stack} = 1$. Given a TCN with 4 layers, according to Eq. 3.2 and our settings, its receptive field would be 121 time steps, if $k = 5$ and $d_i = 2^i$ for $i = 0, 1, 2, 3$.

After updating $\mathbf{X} \leftarrow TCN(\mathbf{X})$, the inference is produced using a regressor, which is a fully connected layer:

$$\hat{\mathbf{Y}} = FC(\mathbf{X}). \tag{3.3}$$

3.3.2 Loss Function

Given the inference $\hat{\mathbf{Y}}$ produced by our unimodal CER model and the continuous label $\mathbf{Y}$, the CCC loss:

$$\ell(\hat{\mathbf{Y}}, \mathbf{Y}) = 1 - \rho_c(\mathbf{Y}, \mathbf{Y}) \tag{3.4}$$

is employed for optimization, where ρ_c is defined in Eq. 2.3.

3.3.3 Model Architecture

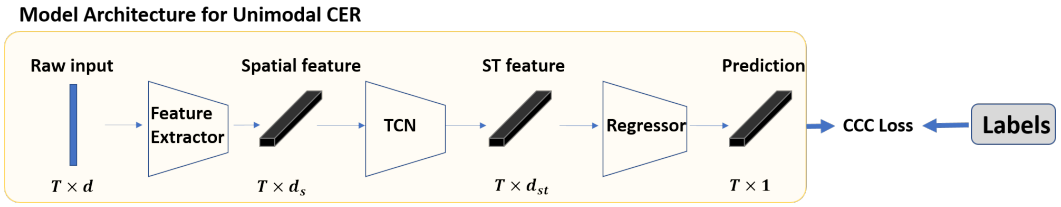

Figure 3.1: The illustration of the unimodal model. The figure shows the architecture of the unimodal model. The model takes T samples sized at d as the input. The feature extractor yields the per-sample features. The latter is then fed to TCN producing the spatiotemporal features. And finally, the regressor maps each feature point onto the 1-D space. ST: spatiotemporal.

Our unimodal CER model is illustrated in Fig. 3.1. It consists of a feature extractor, a TCN, and a regressor (i.e., a fully connected layer). Fed by T consecutive samples, the feature extractor produces T d_s-D spatial features. The latter is then

fed to the TCN producing T d_{st}-D spatiotemporal features. Finally, the regressor maps the features onto the 1-D.

3.4 Implementation Details

3.4.1 Database

MAHNOB-HCI is a multimodal database recorded in response to affective stimuli with the goal of emotion recognition and implicit tagging research [2]. It provides the synchronized recording of facial videos, audio signals, eye gaze data, EEG signals, and other physiological signals from 30 subjects. The subjects are asked to watch 20 emotional video clips, resulting in 440 trials. The video clips are between 35 and 117 seconds long. The EEG signals are acquired from 32 electrodes on the 10-20 international system. The sampling frequency is 256 Hz. The facial videos are captured at 60 fps and 780×580 resolution. For each trial, four integers ranging from 1 to 9 and self-reported by the subjects are used to label the valence, arousal, dominance, and emotional keywords, respectively.

A subset [3] of the original MAHNOB-HCI database is chosen to be continuously labeled. It contains 239 trials from 24 subjects with obvious facial expressions. The trial number for each subject is not even. Five experts are employed for the annotation using FEELTRACE and a joystick. Only the valence is continuously labeled. The reason is that the subjects are quiet and passively watching videos, which makes the annotation of arousal, power, or expectation unavailable [3]. The continuous valence label is determined by the average of the five experts' labels. Our work is based on this subset.

3.4.2 Data Preprocessing

3.4.2.1 Facial Video

Given the facial video of a trial, it contains the facial expression of the subject during the stimuli watching and self-reporting. The latter is excluded by trimming the facial video according to the time stamp information. The video is then changed

to 64 fps for more convenient synchronization with the continuous valence label which is at 4 fps, i.e., every 16 consecutive frame corresponds to 1 valence label point. Finally, the video frames are resized to $48 \times 48 \times 3$.

3.4.2.2 EEG Signal

Given the EEG signal of a trial, the first and last 30s of the recording which do not correspond to stimuli watching are excluded according to the database manual [2]. The signals from the 32 electrodes are then re-referenced to the average reference to enhance the signal-to-noise ratio. The default API *set_eeg_reference* from MNE toolkit [3] is used for the average reference. After which, the average band power on the six bands is calculated.

The average band power computes a single scalar that summarizes the contribution of a given frequency band to the overall power of the signal. Given a windowed EEG discrete signal $X(n)$ with N samples from one EEG electrode in the time domain, the fast Fourier transform (FFT) returns N complex number whose real and imaginary parts represent the amplitude and phase of the signal in the frequency domain. The magnitude-squared of the FFT can be used to obtain an estimate of the power spectral density

$$S_X(f) = \frac{1}{N}\left|\sum_{n=1}^{N} X(n)e^{-i2\pi f n\Delta t}\Delta t\right|^2 \tag{3.5}$$

at f, based upon which the average band power in the frequency band $[f_1, f_2]$ is defined as

$$P_{[f_1,f_2]} = \int_{f_1}^{f_2} S_X(f)df. \tag{3.6}$$

In our work, $S_X(f)$ is obtained using Welch's method from the *scipy* library.

The physiological motivation for the six-band division is elaborate in Section [?]. The window size and hop size for band power calculation are 2s and 0.25s, respectively. The resulted $6 \times 32 = 192$-D band power features at the frequency of $4Hz$ are therefore synchronized with the continuous valence labels. Note that the EEG preprocessing was carried out following the baseline method [3] which employed

[2] `https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf`
[3] `https://mne.tools/stable/index.html`.

only the average reference and band-pass filtering. We did not employ other techniques to deal with the artifacts caused by motion and respiratory. Theoretically, the delta band $(0.3 - 5Hz)$ could contain such artifacts.

3.4.3 Data Partitioning

Two data partitioning schemes are used: (i) trial-level random shuffling (TRS, 10-fold) [3] and (ii) leave-one-subject-out (LOSO, 24-fold). TRS focuses on the trial-level and overlooks from which subject the trial comes. It first randomly shuffles the 239 trials and then splits the 239 trials into 129, 86, and 24 trials for training, validation, and test, so that the test set contains 10% of the data and the training and validation sets contain the 60% and 40% of the remaining data. LOSO focuses on the subject-level. For the i-th fold, trials from the i-th subject are taken as the test set. All the trials from the remaining 23 subjects are randomly shuffled, with 80% and 20% being the training and validation sets, respectively.

TRS may lead to data leakage. The random shuffling would split the data from the same subject to training, validation, and test sets. Compared to the data from different subjects, the data from the same subject has greater consistency. The model trained in this manner has actually seen the test data to some extent and would inflate the test performance. TRS is widely used in fields like computer vision and natural language processing, where the data usually are vastly greater in diversity and therefore invulnerable to the overfitting problem. However, the negative influence becomes nontrivial for fields with limited training data. In AI-based emotion recognition, both the TRS and LOSO are widely used. In our experiment, we choose to employ both schemes and objectively report the results.

3.4.4 Model Training

The teacher model is trained as follows. A Resnet50 is used as the visual backbone. It is pre-trained on the MS-CELEB-1M dataset[4] [150] as a facial recognition task, it is then fine-tuned on the FER+ [155] dataset as a facial expression recognition task.

[4] `https://github.com/TreB1eN/InsightFace_Pytorch#2-pretrained-models--performance`.

Table 3.1: The training settings for the teacher model. The Adam optimizer and ReduceLROnPlateau are from the PyTorch library.

Optimizer		Scheduler	
Adam with the CCC loss		ReduceLROnPlateau	
Learning rate	$1e-5$	Patience	5
Weight decay	$1e-4$	Factor	0.5
Others			
Maximum epoch	30	batch size	2
Early stopping counter	20	Window length (s)	24, equal to 96 data points
Minimum learning rate	$1e-6$	Hop length (s)	8, equal to 32 data points
Random flip (0.5) + random crop (40) for training			
Only center crop (40) for validation			
Normalization of video frames: mean = std = 0.5			

The training settings of our visual model are summarized in Table 3.1. To fine-tune the teacher model on the MAHNOB-HCI dataset, two groups (i.e., the output layer and the whole layer4, according to the Pytorch official implementation) of the Resnet50 backbone are selected. The backbone is initially frozen. When the minimum learning rate is reached, unfreeze one group (starting from the output layer) and reset the scheduler. At the end of each epoch, the best model parameters are loaded. The training would stop if i) there is no remaining backbone layer group, ii) the early stopping counter reaches 20, or iii) the epoch reaches 30.

The training of our EEG model is much the same as the teacher training except for the following. First, since no images are involved, data augmentation and normalization are not employed. Second, the maximal epoch number and early stopping counter are both set to 15 to prevent gradient explosion. Finally, since the student model does not contain a Resnet backbone, it is no need to reset the scheduler.

3.5 Results and Analysis

The experiment examines that the unimodal CER model can produce results no worse than the baseline method on the valence regression task. The best model from Soleymani et al. [3], i.e., a two layers LSTM network is adopted as the baseline for the valence regression task. To make a fair comparison, the baseline model is implemented and incorporated into our pipeline. The experimental results for the valence regression and CKD experiments are obtained in two steps. First, the model outputs for all the trials of a partition are concatenated along the temporal dimension. Recall that our resampling windows have 66.7% overlap. A direct

concatenation is not welcomed as it would produce an over-length prediction vector and further inflates the metrics. Instead, the concatenation is done by placing each output segment according to its windowing indexes. The obtained prediction vector is therefore temporally restored to the original form, which is N-to-N corresponded to the labels. The mean values are taken for the overlapped steps. Second, the RMSE, PCC, and CCC are calculated based on the concatenated prediction vectors and the continuous labels. The results are averaged over the N-fold. Specifically, for the TRS and LOSO partitioning, we have 10 and 24 groups of evaluation results, and the final results are the average across the groups, respectively. This evaluation protocol has been widely used in many CER contests [67, 73–76, 78, 156–158].

The result of our visual model against the baseline using the TRS and LOSO data partitioning is reported in Table 3.2. For the TRS partitioning, it can be observed that results from the test set are more consistent with those from the validation set. Only a slight drop in PCC and CCC when it goes from validation to test on both of the two methods. As we mentioned before, the reasons are two-fold: i) the visual modality is highly determined, and ii) data from one subject are already seen by the model during the training stage. For the LOSO partitioning, the gap between the validation and test results is relatively larger. Up to 13.38% and 10.57% drop on CCC can be observed from ours and the baseline, respectively. Overall, our method produces superior results to that of the baseline methods on RMSE, PCC, and CCC. Given a scenario, e.g., the test on LOSO partitioning, the 24 CCC pairs of CCC from the two methods are evaluated using the paired t-test. The p-values obtained from the t-tests are all smaller than 0.01. The statistical significance is obtained for TRS and LOSO partitioning over the visual and EEG modalities.

Table 3.2: The result of our visual model against the baseline using the TRS and LOSO data partitioning. The mean and standard deviation are reported. Given a scenario, e.g., the test on LOSO partitioning, the 24 CCC pairs of CCC from the two methods are evaluated using the paired t-test. The p-values obtained from the t-tests are all smaller than 0.01 between two corresponding scenarios. TRS: trial-wise random shuffling. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

Visual modality	**Ours with TRS**		Soleymani et al. with TRS		**Ours with LOSO**		Soleymani et al. with LOSO	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
RMSE ↓	0.054±0.006	0.054±0.006	0.060±0.006	0.058±0.009	0.053±0.005	0.049±0.015	0.060±0.007	0.055±0.020
PCC ↑	0.697±0.054	0.686±0.079	0.623±0.079	0.611±0.150	0.699±0.068	0.684±0.203	0.611±0.105	0.602±0.264
CCC ↑	0.690±0.057	0.674±0.085	0.606±0.086	0.589±0.163	0.695±0.068	0.602±0.228	0.596±0.107	0.533±0.251

Table 3.3: The result of our EEG model against the baseline using the TRS and LOSO data partitioning. The mean and standard deviation are reported. Given a scenario, e.g., the test on LOSO partitioning, the 24 CCC pairs of CCC from the two methods are evaluated using the paired t-test. The p-values obtained from the t-tests are all smaller than 0.01 between two corresponding scenarios. TRS: trial-wise random shuffling. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. Bold fonts indicate the best results.

EEG modality	**Ours with TRS**		Soleymani et al. with TRS		**Ours with LOSO**		Soleymani et al. with LOSO	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
RMSE ↓	0.067±0.005	0.066±0.009	0.083±0.006	0.080±0.012	0.068±0.007	0.066±0.025	0.087±0.020	0.081±0.034
PCC ↑	0.463±0.103	0.435±0.205	0.347±0.091	0.353±0.228	0.467±0.116	0.474±0.267	0.360±0.147	0.427±0.267
CCC ↑	0.444±0.109	0.415±0.201	0.331±0.092	0.333±0.214	0.445±0.118	0.377±0.250	0.348±0.129	0.306±0.257

The result of our EEG model against the baseline using the TRS and LOSO data partitioning are reported in Table 3.3. For the TRS partitioning, consistent results between the validation and test are also observed for the two methods. For the LOSO partitioning, up to 15.28% and 12.07% drop on CCC are observed from ours and the baseline, respectively, which is larger than their visual counterpart. In this modality, our EEG model also produces better results against the baseline in terms of RMSE, PCC, and CCC.

3.5.1 Interpretation

In our work, we are also particularly interested in revealing the contribution of each brain lobe and the band of EEG towards the emotion process. To this end, we visualize the skull saliency map for each band and subject, based on the trained student model and the peak response mapping [153] (PRM). The PRM is based on an observation that the backward propagation of the peak logit usually leads to informative regions of an image corresponding to the class.

Given the EEG band power calculated in the 6 bands ($0.3-5Hz$, $5-8Hz$, $8-12Hz$, $12-18Hz$, $18-30Hz$ and $30-45Hz$) from 32 electrodes (note that the β band is split to two sub-bands β_1 and β_2), the PRM is adopted as follows, with Fig. 3.2 illustrating the pipeline. The EEG band power from the LOSO partitioning is fed to the trained student model that is corresponding to the $i-th$ subject. For each trial of the test set from the $i-th$ subject, the peak scalar from the prediction is used to carry out the backward propagation, producing T_j 32×6 gradient vectors, where T_j denotes the time steps of the j-th trial. The temporally averaged gradient vector ($N \times 32 \times 6$ where N denotes the trial number) of this trial together with those from other trials are averaged again to obtain the gradient vector (32×6) for the $i-th$ subject. Normalization of the gradient vectors over the 6 bands is employed so that the inter-band information is preserved. Note that a per-band normalization rescaling each band independently is inappropriate since the visualization would all look similar in terms of color intensity. After which, for the k-th out of the 6 bands, 32 scalars corresponding to the 32 electrodes of the 10-20 system are used to generate the 6 skull saliency maps for the i-th subject. The same process is conducted on all the 24 subjects obtaining 24×6 topographic saliency maps. Finally, the results from all the subjects are averaged and visualized

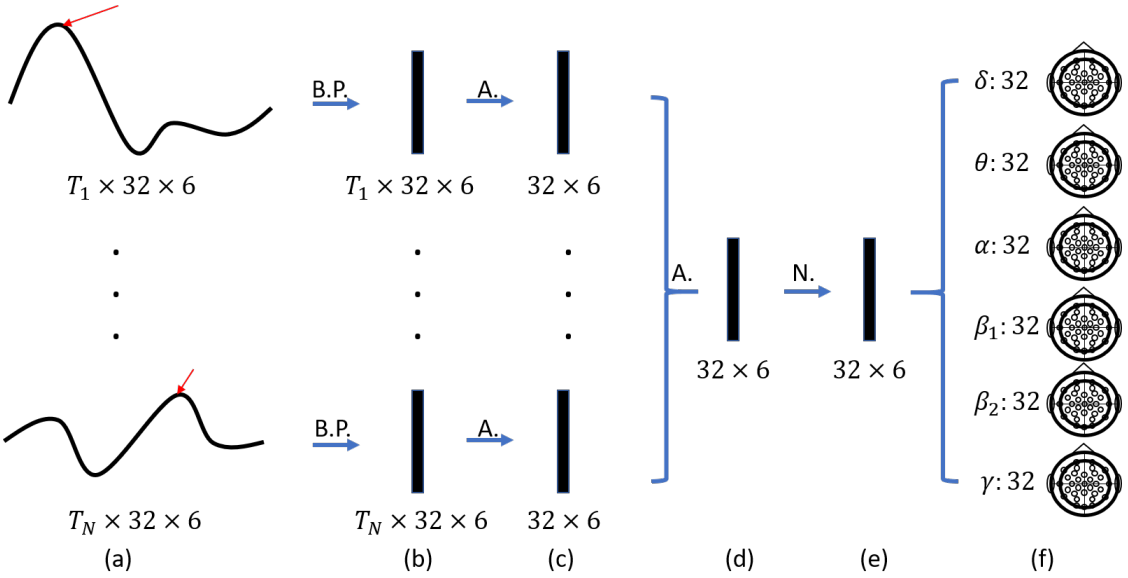

Figure 3.2: Illustration of generating the peak response mapping for interpretability [1] investigation. The figure shows the procedure of obtaining the saliancy map. Given the trained EEG model for the i-th subject, (a) N $T_j \times 32 \times 6$ valence predictions for N trials are obtained. By selectively back-propagate the peaks, (b) N $T_j \times 32 \times 6$ gradient vectors for the N trials are obtained. By averaging on the temporal dimension, (c) N 32×6 gradient vectors are obtained. After which, the average over the trial dimension is conducted producing (d) the 32×6 gradient vector of the i-th subject. (e) The normalized version of the latter is finally used to plot (f) the heatmap on the six bands using the MNE toolkit. B.P.: backward propagation. A.: average. N.: normalization. The red arrow points to the peak value for the backpropagation.

as well. Theoretically, we expect that the saliency map of the β and γ bands should manifest a warmer color compared to those from other bands.

The per-subject topographic saliency maps over the six bands are illustrated in Fig. 3.3. Note that by referencing the saliency maps to the brain division and 32-electrode placement shown in Fig. 2.3, we can locate the active and less active brain regions for the state of high valence. We see that the last two columns, corresponding to the β_2 and γ bands, are apparently warmer than the rest four columns. Specifically, in the β_2 band, active frontal lobes are observed on all subjects, while in the γ band, the occipital and parietal lobes take the place. We could explain that during the experiment, subjects should be focused on watching the movie clips, with their occipital and temporal lobes perceiving the visual and audio stimuli, the parietal lobe integrating the perceived, and the frontal lobe making the decision and directing the facial expression.

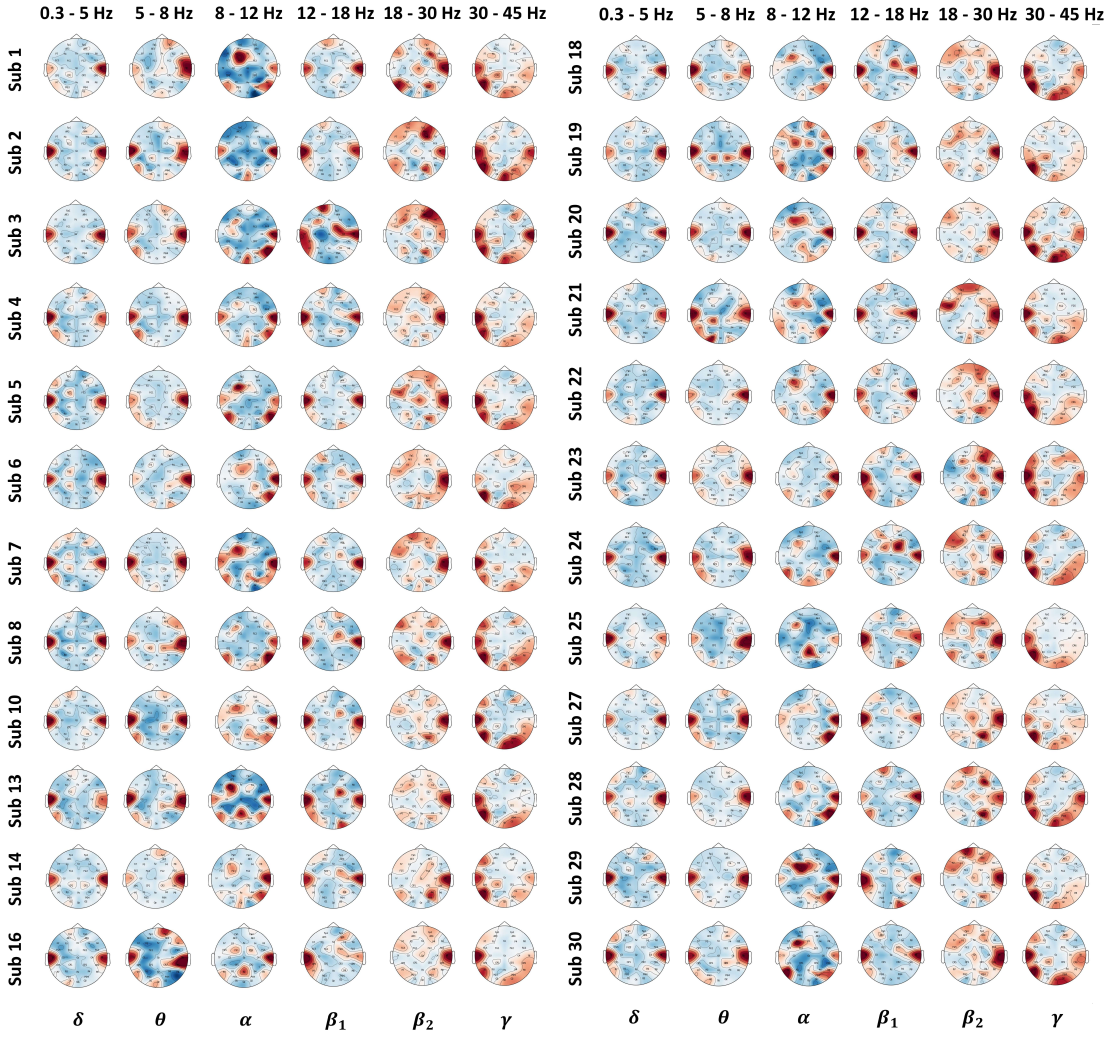

Figure 3.3: Topographic saliency maps. The figure shows the topographic saliency maps for the 24 subjects. The gradients of the EEG band power over the 32 electrodes are calculated following the procedure shown in Fig. 3.2. The warmer color means a higher gradient. A region having warmer color implies that it contributes more to the valence prediction. Therefore, the red regions tend to be more informative for the neural networks to infer the valence compared to the blue counterparts. Sub: subject. The subject numbering is determined by the MAHNOB-HCI database [2]. The missing subjects are not included in the subset [3] since they are not continuously labeled in valence.

Moreover, active temporal lobes are observed over all the subjects and six bands. In the δ band $(0.3-5Hz)$, no lobe is comparably active against the temporal lobe. In the θ band $(5-8Hz)$, mild activation of the frontal lobe can be observed from Subject 1, 2, 3, 10, 14, 16, 18, 20, 23, and 27. And so are the active parietal lobe observed from Subject 8, 10, 13, 14, 19, and 21. And the active occipital lobes are observed from Subject 1, 5, 16, 19, 20, 21, 25, 27, and 30. In the α band $(8-12Hz)$, more active frontal lobes can be observed from Subject 1, 5, 6, 7, 13, 19, 20, 21, 22, 29, and 30. Highly active parietal lobes are observed for all the subjects except for Subject 2. And, Subject 2, 3, 4, and 24 have active occipital lobes.

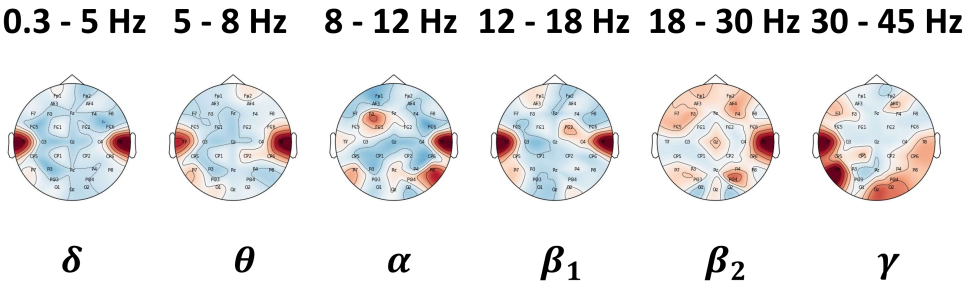

Figure 3.4: Overall topographic saliency maps. The figure shows the saliency maps averaged over all the 24 subjects. The warmer color means a higher gradient, and further implies more contribution to the valence prediction.

To summarize, we focus on the overall saliency map shown in Fig. 3.4. In terms of brain lobes, all the four lobes can be active on the six bands, which conforms to the fact that complex mental functions do not reside in any one place [13], instead of locating complex functions in precise brain areas [159–161]. In terms of bands, the β_2 and γ, with the frequency of $18-30Hz$ and $30-45Hz$, contribute the most to the human emotion process compared to other bands. The observation complies with the knowledge we discussed before, that i) the β band corresponds to a focused state of mind, and is more obvious in the frontal lobe, and ii) the γ band corresponds to the high-level cognition process such as the perception, transmission, and integration of the visual and audio stimuli from the occipital, temporal, and parietal lobes.

3.6 Discussion and Conclusion

This chapter introduces the unimodal CER model, which addresses the challenge of temporal modeling by incorporating a TCN with a large resampling window.

The model is designed to process either a video frame or EEG signal as input. When using video, a Resnet50 serves as the visual backbone for feature extraction, while the EEG signal is processed using the mean band power as the feature. By leveraging these inputs, the TCN effectively learns the spatiotemporal features required for the final inference.

The proposed unimodal model obtained 0.60+ and 0.30+ in CCC for visual and EEG modality, where a large performance gap can be observed. One possible reason for the difference in CCC is the quality and quantity of data. The facial landmark modality may have a larger dataset available, which can lead to improved accuracy due to having more examples to train the model. In contrast, EEG signals may be more difficult and expensive to collect, which can result in a smaller and less diverse dataset, leading to lower accuracy. Another possible reason is the complexity of the modality. Facial landmark is a relatively simple modality that captures visual cues related to facial expressions. EEG signals, on the other hand, are more complex and require specialized knowledge to preprocess and extract relevant features. If the feature extraction or modeling process for the EEG data is not optimized, it can lead to lower accuracy. The modality itself may also play a role in the difference in accuracy. Facial expressions are overt cues that are more directly related to emotions, while EEG signals are covert cues that capture neural activity associated with emotions. The relationship between neural activity and emotions is complex, and the EEG signal may be affected by a variety of factors, such as individual differences in brain function, noise, or artifact.

which modality is the best for CER? Can we simply conclude that the visual modality is much stronger than the EEG modality for CER? The answer to this question is the motivation and starting point for the development based on the unimodal model. Visual modalities, such as facial landmarks and video frames, can provide important visual cues that may be indicative of a person's emotional state. For example, changes in facial expressions such as frowning or smiling can provide valuable information for emotion recognition algorithms. However, visual modalities can be affected by factors such as lighting, angle, and occlusions, which can make it difficult to accurately detect and interpret facial expressions. Audio modalities, such as sound waves and spectrograms, can capture important auditory cues related to emotions, such as changes in tone, pitch, and volume. These cues can be particularly useful in detecting emotional states such as anger or sadness.

However, audio modalities can also be affected by background noise, accent, and the quality of the recording equipment, which can impact the accuracy of emotion recognition algorithms. Physiological modalities, such as the EEG signal, can provide information about changes in brain activity that may be related to emotional states. This can be particularly useful for detecting subtle changes in emotional states that may not be visible through visual or auditory cues alone. However, physiological modalities require specialized equipment and can be invasive, which can limit their practicality in certain settings. Overall, the best modality for emotion recognition depends on the specific context and application. In some cases, visual cues may be the most informative, while in other cases, audio or physiological cues may be more useful. Additionally, many emotion recognition algorithms use a combination of modalities to improve accuracy and robustness.

Furthermore, we investigate the interpretability of the unimodal model from a physiological perspective by leveraging the PRM visualization method [153]. This approach allows us to identify the contribution of different brain regions and EEG frequency bands to the emotion process. Our findings demonstrate that all four brain lobes work synergistically in this process, highlighting the complex nature of brain function. Moreover, we observe that the β_2 and γ frequency bands, covering the range of $18 - 30Hz$ and $30 - 45Hz$ respectively, are the primary contributors to the emotion process, surpassing the influence of other frequency bands.

Chapter 4

Multi-modal Emotion Recognition with Refined Labels for Deep Learning

This chapter [1] aims to present the label correction method we proposed for CER. The traditional emotion classification framework usually fits all the feature segments of the same trial to a fixed annotation. Considering the fact that emotion is a continuous reaction to stimuli that lasts for varied periods, we argue that the indiscriminate annotation is equivalent to taking the emotional state as fixed within the whole trial, leading to a decrease in the classification accuracy. In this study, we attempt to alleviate this issue by developing a thresholding scheme, converting the continuous emotional trace into a three-class sequential annotation. The features within a trial are therefore assigned to varied emotional states, resulting in an improvement in the accuracy. A long short-term memory (LSTM) networks-based emotion classification framework is implemented, to which the proposed thresholding scheme is applied. A subset of the MAHNOB-HCI dataset with continuous emotional annotation is used. The EEG signal and frontal facial video are used for feature extraction. The experiment results demonstrate that the proposed scheme provides statistically significant improvement to the three-class classification accuracy of the EEG feature-based LSTM network (p-value = 0.0329).

[1] The work in this chapter has been published in [19].

4.1 Introduction

Emotion recognition is the process of distinguishing the emotional state, where the emotional state is a term temporally referring to the current "feeling" of a person elicited by stimuli. Emotion recognition plays an important role in various fields such as affective computing, human-computer interaction, education, and gaming. There are mainly two directions in this research, depending on the output form. Emotion classification is an N-to-1 problem, which feeds N samples of emotion cues, and produces only 1 categorical inference. Whereas CER is an N-to-N problem, which feeds N samples of emotion cues while producing N continuous inference. In this study, we attempt to utilize the continuous annotation from the latter to improve the classification accuracy of the former.

The traditional training process of emotion classification usually contains the following steps. First, given the emotion cues from a trial, be it video frames, audio signals, or EEG signals, a sliding window is employed with a certain overlap ratio for re-sampling. So that a certain number of cue segments are obtained. Second, for each segment, it is then assigned the same categorical annotation. Finally, the classifier will be trained to fit each segment to its associated annotation. In this study, we argue that the classification accuracy of the traditional training process may be hindered by the N-to-1 mapping. The test protocol of the typical datasets [2, 60] usually uses short video clips with a length of around 30 to 90 seconds for emotion elicitation, and simultaneously collects the participant's visual, auditory, and physiological data. The participant's self-rating in the form of a scalar will play as the annotation. However, the fixed scalar is problematic considering the fact that emotion is a reaction to stimuli that lasts for seconds or minutes. For instance, a participant may label a video clip as high valence, though he/she may not always be in a state of high valence during the whole trial. In this case, the data that corresponds to other emotional states are therefore falsely labeled. The indiscriminate label can lead to an increase in noisiness and confusion of the classifier.

In this study, we are particularly interested in investigating two questions. (i) Can the aforementioned issue be alleviated by training the classifier using the discretized sequential label? (ii) Whether the idea can work on different modalities? To this end, a thresholding scheme is developed to yield the discretized label from the

emotional trace. Specifically, for each trial, the trace is discretized to three classes, and the classifier is trained to fit the features to its corresponding three-class labels flexibly, instead of its indiscriminate counterpart. To validate our thresholding scheme, a traditional emotion classification framework [2] based on the EEG feature and facial expression feature is implemented, to which the thresholding scheme is then applied. A subset [3] of MAHNOB-HCI dataset [2] with continuous valence trace is used. The subjects' EEG signals and frontal facial video are used as the emotion cues. A two-layer long short-term memory (LSTM) network is used as the classifier. The experiment results show that, without the thresholding scheme, the facial expression-based classifier is superior to its EEG-based counterpart for about 2% accuracy. And with the thresholding, the latter obtains a gain of 3% accuracy and therefore claims superiority over the former.

4.2 Methodology

Suppose that we are given one sequence of emotion cue $\mathbf{X} = \{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{N\times d}$, the corresponding ground truth $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^N$, and the categorical ground truth $\mathbf{C} = \{c_1, c_2, \ldots, c_N\} \in \{0, 1, 2\}^N$, where $c_1 = c_2 = \ldots = c_N$. All of which contains N sample points.

Our goal is to define a thresholding scheme D, discretizing the continuous label $\mathbf{Y}$ into a set of three-class categorical labels $\mathbf{Z} = D(\mathbf{Y}) = \{D(y_1), D(y_2), \ldots, D(y_N)\}$, where $D(y_i) \in \{0, 1, 2\}$. After which, we find a function $f : \mathbf{X} \to \mathbf{Z}$ so that its output $\hat{\mathbf{Z}} = f(\mathbf{X}) \in \{0, 1, 2\}^{N\times d}$ minimizes some loss $L(\mathbf{Z}, \hat{\mathbf{Z}})$. Hopefully, during the test, the model inference $\hat{\mathbf{Z}}$ can have a higher accuracy towards $\mathbf{C}$than that from the traditional way by fitting $f : \mathbf{X} \to \mathbf{C}$. Note that the causal constrain is applied, which requires that y_t depends only on $x_1, \ldots, x_t$ and not on any future inputs $x_{t+1}, \ldots, x_T$.

4.2.1 Thresholding Scheme

The experiment paradigm of MAHNOB-HCI database [2] employs the self-assessment manikins (SAM) [59] to categorically label the data. The subjects first watch short video clips for each trial. At the end of the trial, the subjects are asked to report

their emotion ratings and tags. The emotion ratings are integers ranging from 1 to 9 for the valence and arousal dimensions. The emotion tags are nine basic emotions including neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear. The nine categories are then grouped into three classes, which are (i) pleasant including amusement and joy, (ii) neutral including neutral and surprise, and (iii) unpleasant including anxiety, sadness, disgust, anger, and fear. The authors also request five annotators to continuously label the subjects' valence traces according to their facial expression, obtaining an averaged valence trace in the interval of $[-1.0, 1.0]$ [3]. In sum, from the 239 trials that are continuously labeled, we have 239 three-class emotional categorical labels, 239 integer valence ratings, and 239 valence traces involved in our study.

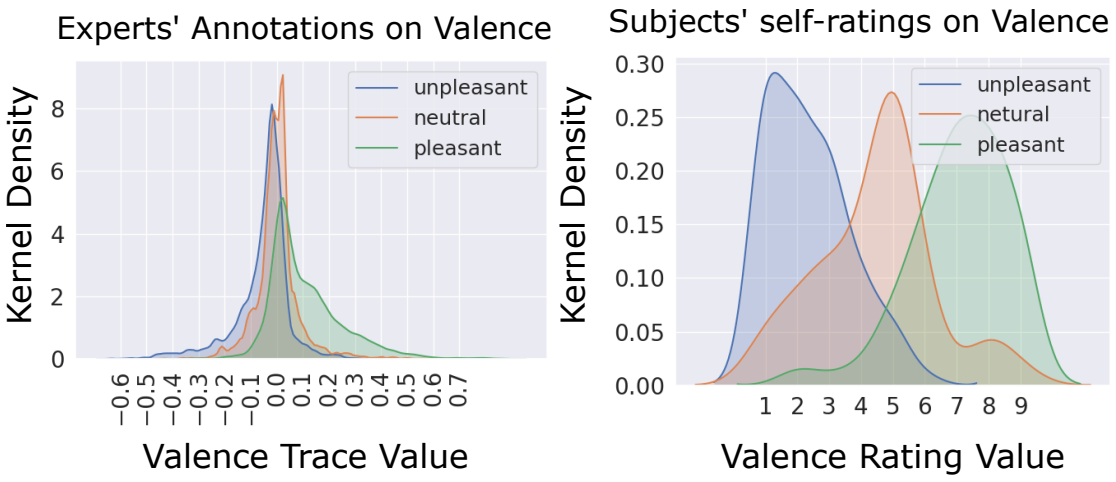

Figure 4.1: Kernel density estimate (KDE) plot. The figure shows the KDF plot of the experts' valence traces (left) and subjects' valence ratings (right) against the 3-class subjects' emotion tag, respectively. The subjects' 3-class emotion tag have a more consistent distribution with the subjects' valence ratings compared to that with the experts' valence traces. Best viewed in color.

An exploratory data analysis using the kernel density estimate plot of the Seaborn library is conducted. The results are shown in Fig. 4.1. (i) The values of valence traces and (ii) the subjects' nine-class valence self-ratings of the 239 trials are grouped with respect to their three-class labels, as shown in the left and right subfigure of Fig. 4.1respectively. We see that the distribution of the valence trace mostly overlaps with each other. In contrast, the distribution of the valence ratings has a much lower overlap ratio, and they are also consistent with the tendency as the unpleasant/pleasant class tends to have relatively low/high valence ratings, respectively. Such a discrepancy in the distributions implies the gap between the annotator's observation of the subject's overt emotional state (i.e., the subject's

expression), and the subject's covert emotional state (i.e., the subject's experience). Also, a trial rated as pleasant does not necessarily mean that the subject feels constant pleasant over the whole trial. Therefore, we would like to differentiate the overt state-based valence trace by considering the information from the covert state-based valence rating, followed by obtaining the specific annotation for a small window from the trace.

Based on the exploratory data analysis, the thresholding scheme is designed as follows. Given one point y of the trace, let D be the discretion operation, the goal is to find a threshold τ, discretizing the valence trace into three classes:

$$D(y) = \begin{cases} 0, & \text{if } y < -\tau \\ 1, & \text{if } -\tau \leq y < \tau \\ 2, & \text{otherwise} \end{cases}, \tag{4.1}$$

where 0, 1 and 2 correspond to unpleasant (low valence), neutral (mild valence), and pleasant (high valence). However, considering the high overlap ratio among the valence traces of different classes, it is awkward to directly apply the discretion. We assume that the distribution of the unpleasant and pleasant traces should be symmetric about the y-axis, according to the observation of the distribution of valence ratings. Therefore a translation parameter $s \in \mathbb{R}$ is employed to globally shift the traces aside, obtaining:

$$\begin{cases} y^{(0)} \leftarrow y^{(0)} - s \\ y^{(1)} = y^{(1)} \\ y^{(2)} \leftarrow y^{(2)} + s \end{cases}, \tag{4.2}$$

where the superscripts (0), (1) and (2) correspond to the unpleasant, neutral and pleasant classes. The shifted traces from Eq. 4.2 are then processed by Eq. 4.1. We resort to experiment to empirically find the appropriate τ and s.

4.2.2 Model Architecture

An LSTM network is implemented using Pytorch API with GPU acceleration. The network contains two LSTM layers and two linear layers, as shown in Fig. 4.2. The number of LSTM cells contained in the two LSTM layers equals half and a quarter

of the number of the input features. The first linear layer has the same neuron number as the hidden unit number of the adjacent LSTM layer, and the second linear has 3 neurons with Softmax to carry out the classification. The two hidden layers and the first linear layer all have a 0.4 dropout rate. The network contains around $1E+5$ and $2E+4$ parameters for EEG and facial data experiments.

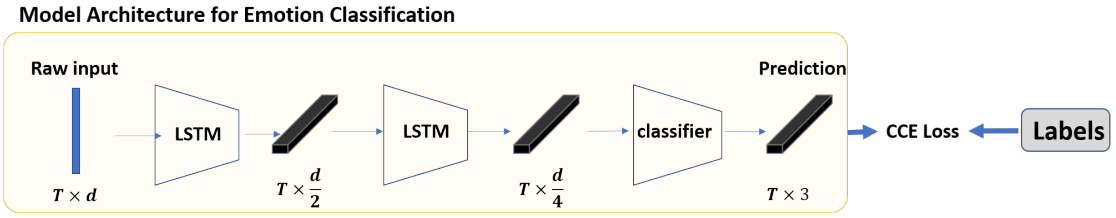

Figure 4.2: The architecture of the LSTM network. The figure shows the architecture of the LSTM network employed for the experiment. CCE Loss: categorical cross-entropy loss.

4.3 Implementation Details

In order to enable the mini-batch training on the sequential input with varied time-step, the sizes of all the inputs are zero-padded according to the trial with the largest time-step, followed by the pack and unpack operation before and after feeding to the LSTM cells. The loss function is also masked to ensure that the gradient from the padded entries will not affect the optimization.

4.3.1 Database and Data Partitioning

The database used in this chapter is the same as in the last chapter (please see Section 3.4.1), i.e., it is the subset [3] of the MAHNOB-HCI database [2]. The leave-one-subject-out (LOSO) is employed to partition the data into the training, validation, and test sets.

4.3.2 Data Preprocessing

For the EEG signal, the same technique as in Section 3.4.2.2 is used to calculate the average band power from five bands, i.e., $4-8$ Hz, $8-12$ Hz, $12-18$ Hz,

$18 - 30$ Hz and > 30 Hz, yielding $32 \times 5 = 160$ features for each time step at a frequency of 4 Hz.

For the visual cues, facial landmarks are employed. For every 25 frames of a video, 68 facial fiducial points are extracted using the Dlib library [162] at the first frame, and the fiducial points are obtained using the Lucas-Kanade (LK) algorithm [163] for the rest 24 frames. By which the fiducial points will not collapse due to the accumulation of discontinuity among frames. To reduce the variation of the point sets caused by face shape, head pose, etc., the same technique as in [3] is employed. Specifically, a mean face is calculated across all the subjects by averaging the fiducial point sets of the first frame. For each subject, an affine transformation is established by registering his/her point set from the first frame to the mean face. The transformation is then applied to all the other point sets of the same subject. After which the points with indexes $\{28, 29, 30, 31, 40, 43\}$ are selected to calculate the reference point. For each frame, the reference point is obtained by averaging the selected points, and the distances from each point to the reference point are served as the features, yielding 68 feature for each frame. Finally, the features are down-sampled to 4 Hz.

4.4 Results and Analysis

Recall that we would like to explore two questions, i.e., (i) can the aforementioned issue be alleviated by training the classifier using the discretized sequential label? (ii) whether the idea can work on different modalities? The experiment is designed as follows. First, we carry out the standard classification for the visual and EEG modalities, which trains the model solely using the three-class categorical label **C**. The accuracy obtained will then serve as the baseline for examining the validity of the thresholding scheme. Then, we train the model again with the three-class discretized label **Z** from our thresholding scheme.

The experiments are performed on a PC with a 3.2 GHz AMD Core CPU, Nvidia RTX 2070 GPU, and 16 GB memory using Python code. The masked categorical cross-entropy loss function with Adam optimizer [164] is employed. The learning rate is set to $2E - 5$ with a momentum of 0.9. The maximum epoch in training

is set to 100. If there is no improvement in the cross-entropy on the validation set after 10 epochs, the training is stopped with the early stopping strategy.

Table 4.1: The accuracy of the baseline for each subject, as well as the overall mean and standard deviation. The table reports the performance without using the thresholding scheme, which could serve as the baseline. S: subject. The round bracket indicates the number of trials involved in that subject.

	S01(19)	S02(5)	S03(13)	S04(9)	S05(13)	S06(9)	S07(13)	S08(13)	S10(5)	S13(6)	S14(13)	S16(2)
EEG	0.6034	0.5015	0.6113	0.5754	0.5444	0.5577	0.2254	0.5794	0.4797	0.2801	0.5097	0.5415
Face	0.5235	0.7169	0.5837	0.6057	0.5121	0.6572	0.4174	0.5932	0.4360	0.4217	0.6083	0.5131
	S18(1)	S19(11)	S20(17)	S21(10)	S22(14)	S23(15)	S24(12)	S25(16)	S27(6)	S28(4)	S29(10)	S30(3)
EEG	0.5963	0.6156	0.5214	0.6150	0.4893	0.5058	0.6293	0.5809	0.5677	0.5118	0.2922	0.5233
Face	0.6226	0.5755	0.5233	0.5211	0.5711	0.5266	0.6685	0.5299	0.5424	0.4880	0.2296	0.5203
Overall	EEG: 0.5191 ± 0.1075							Face: 0.5378 ± 0.0992				

4.4.1 Result on the Baseline

In this series of experiments, the implemented emotion classification framework is validated without the thresholding scheme, i.e., the annotation is fixed to the subjects' emotional categorical labels for all the training, validation, and test sets. The baseline accuracy for each subject, as well as the mean and standard deviation, are shown in Table 4.1. The subjects' indexes follow the manual of the MAHNOB-HCI dataset.

4.4.2 Result on Our Thresholding Scheme

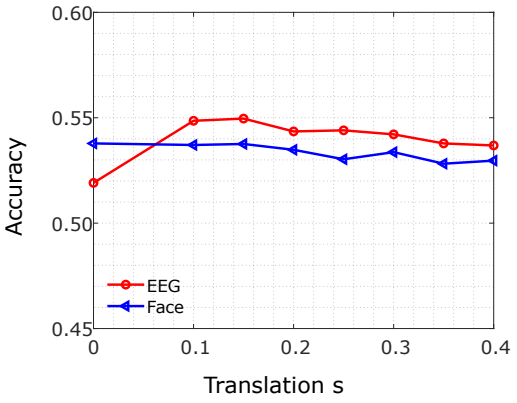

Figure 4.3: Accuracies obtained using different translation s on the visual and EEG modalities. The line graph shows an increase of accuracy on EEG modality when $s = 0.10$ and $s = 0.15$. And there is no improvement on the visual modality. Best viewed in color.

In this series of experiments, seven degrees of translation, i.e., $s \in \{0.10, 0.15, \ldots, 0.40\}$ are selected for experimenting. For each degree, the thresholding interval is set to $\tau \in [0.01, 0.9s]$, with a step-size of 0.01. To obtain the result for a translation s, the following procedure is carried out. Suppose that $s = 0.20$, then according to the definition of threshold τ, the latter should be set to $[0.01, 0.18]$. Under the step-size 0.01 there are 18 thresholds to experiment. Each threshold will be pared with $s = 0.20$ to perform the label refinement (following Eq. 4.2 and Eq. 4.1) and an accuracy will be obtained from the experiment. The final result for translation $s = 0.20$ is obtained by averaging the results from the 18 pairs. The same applied to all other translation s and all the 24 folds from the LOSO partitioning. The result is shown in Fig. 4.3. It is worth noting that the thresholding scheme is only applied to the training and validation sets, while the originally fixed annotations

Table 4.2: The comparison of the accuracy of the EEG feature-based accuracy against its baseline accuracy. STD: standard deviation. P-value: the p-value of the paired one-tailed t-test. The smaller the p-value is, the more it supports the hypothesis that our thresholding scheme can increase the accuracy. Normality: the p-value of the Shapiro Wilk test. The larger the normality is, the more it supports that the samples are from a normally distributed population. The bond fonts indicate the best result.

Translation	Mean±STD	P-value	Normality
$s = 0.0$	0.5191 ± 0.1075	-	-
$s = 0.10$	0.5486 ± 0.1132	0.0303	0.6819
$s = 0.15$	$\mathbf{0.5496 \pm 0.1141}$	0.0329	0.8726
$s = 0.20$	0.5435 ± 0.1098	0.0692	0.7566
$s = 0.25$	0.5440 ± 0.1125	0.0703	0.8731
$s = 0.30$	0.5421 ± 0.1120	0.0775	0.8542
$s = 0.35$	0.5378 ± 0.1105	0.1177	0.9369
$s = 0.40$	0.5369 ± 0.1099	0.1325	0.7830

are preserved for accuracy calculation using the test set. The red and blue lines are the results of EEG and facial features, respectively. The dashed line presents the baseline accuracy obtained from the first experiment. We can see that the thresholding scheme can increase the accuracy of EEG feature-based classification, which peaks when $s = 0.15$, while it has no positive effect on the facial feature. A paired one-tailed t-test is conducted to investigate the statistical significance of the EEG feature-based result against its baseline accuracy. The null hypothesis H_0 and alternative hypothesis H_1 are defined as $\mu \leq \mu_0$ and $\mu > \mu_0$, respectively, where μ and μ_0 denote the subject-wise averaged accuracy from our thresholding scheme and its counterpart from the traditional manner. The p-values are listed in Table 4.2.

Our thresholding scheme cannot benefit the facial modality. The effect discrepancy of our thresholding scheme on the two modalities might result from the human reaction time (RT). In the field of mental chronometry, RT is defined as the interval between the onset of a target stimulus and the initiation of the subject's immediate motor response to the target appearance [165]. In our study, there exist two temporally adjacent processes of this perceptual-motor chain, i.e., the subject's perception of visual/auditory stimuli and facial expression elicitation, followed by the annotator's perception of the subject's expression and the annotation operation, where the latter directly links the valence trace to the facial modality. Therefore, the subject's EEG signal and the annotator's annotation have a greater latency

than that between the subject's expression and the annotation. Our thresholding scheme may compensate for this gap, leveraging the performance of the EEG modality, yet still bounded by the second process.

4.5 Discussion and Conclusion

In this study, we attempt to answer two questions. (i) Whether the training using the discretized emotional trace can improve the performance of the classifier? (ii) Whether this idea can function in different modalities? The EEG signal and facial videos are used for feature extraction in our study, from which the band power and distance of 68 facial fiducial points to the reference point are extracted as the features, respectively. The accuracy of our LSTM network using the EEG feature is around 51.9%, while 57.0% is reported from the MAHNOB-HCI dataset [2]. The reason can be explained by the balance and sufficiency difference. Only 239 trials with meaningful facial expressions are continuously labeled and involved in our study, and the trials are not evenly distributed across the 24 subjects, as shown in Table 4.3. In contrast, 540 trials with even distribution across 27 subjects are used in [2] to achieve higher performance. On the other hand, before applying the thresholding scheme, the facial feature-based classification has around 2% superiority in accuracy compared to the EEG feature-based classification, which can be explained by the fact that the facial expression is more discriminant than the EEG signal.

Indeed, the difference between overt and covert modalities can impact how they contribute to emotion recognition. Overt modalities, such as facial expressions, are visible to others and can be observed in an objective manner. This makes them relatively easy to detect and interpret using computer vision techniques, such as facial landmark detection or facial expression analysis. Overt modalities can be particularly useful in social contexts, where facial expressions are an important part of human communication. Covert modalities, such as EEG signals, are not visible to others and require specialized equipment to detect. However, they can provide valuable information about the underlying neural activity that is associated with emotional states. EEG signals can detect changes in brain activity that are not visible through overt modalities, such as changes in attention or cognitive processing. Covert modalities can also be particularly useful in situations where

overt cues may be unreliable or difficult to interpret, such as in individuals with facial paralysis or those who are attempting to conceal their emotions.

The accuracy obtained above is then played as the baseline for the exploration of the two aforementioned questions. For EEG feature-based classification, our threshold scheme can increase the accuracy by around 3%. When $s = 0.10$ and $s = 0.15$, the statistical significance between the performances from the baseline and the proposed thresholding scheme exists when $\alpha < 0.05$ according to the p-value of the t-test. The thresholding scheme leverages the accuracy of EEG feature-based classification to 54.96%, which is higher than 53.78% obtained from the facial feature-based baseline. However, there is no improvement when applying the thresholding scheme on the facial feature-based baseline. The difference in the improvement obtained when combining EEG with visual-based human annotation versus facial landmark with visual-based human annotation may be due to the complementary nature of the modalities.

EEG signals capture covert cues related to neural activity associated with emotions, while facial landmarks capture overt cues related to facial expressions. When these two modalities are combined, they may provide a more complete picture of the emotional state than either modality alone. The visual-based human annotation can also provide additional information about the emotional state that may not be captured by either the EEG or facial landmark modalities alone. In contrast, when combining facial landmark with visual-based human annotation, there may be less complementary information provided by the modalities. Both facial landmark and visual-based human annotation capture visual cues related to facial expressions, and they may be measuring similar aspects of the emotional state. In this case, combining the two modalities may not provide much additional information beyond what is already captured by facial landmark or visual-based human annotation alone.

Chapter 5

Multimodal CER through Visual-to-EEG Cross-modal Knowledge Distillation

This chapter [1] presents the visual-to-EEG cross-modal knowledge distillation. Visual modality is one of the most dominant modalities for current continuous emotion recognition methods. Compared to which the EEG modality is relatively less sound due to its intrinsic limitation such as subject bias and low spatial resolution. This work attempts to improve the continuous prediction of the EEG modality by using dark knowledge from the visual modality. The teacher model is built by a cascade convolutional neural network - temporal convolutional network (CNN-TCN) architecture, and the student model is built by TCNs. They are fed by video frames and EEG average band power features, respectively. Two data partitioning schemes are employed, i.e., the trial-level random shuffling (TRS) and the leave-one-subject-out (LOSO). The employment of the visual-to-EEG cross-modal KD further improves the prediction with statistical significance, i.e., p-value < 0.01 for TRS and p-value < 0.05 for LOSO partitioning. The code is available at https://github.com/sucv/Visual_to_EEG_Cross_Modal_KD_for_CER.

[1] The work in this chapter has been published in [4].

5.1 Introduction

Continuous emotion recognition (CER) is the process of identifying human emotion in a temporally continuous manner. The emotional state, once understood, can be used in various areas including entertainment, e-healthcare, recommender system, and e-learning. To describe the human state of feeling, psychologists have developed categorical and dimensional models. The categorical model aims to obtain a discrete estimate of emotional category. It features simplicity and universality and has been extensively exploited in affective computing. The dimensional model, on the other hand, aims to obtain a continuous estimate in a dimensional space. It can describe more complex and subtle emotions. This paper focuses on developing a CER method based on the dimensional model.

CER can utilize information from various modalities. The visual modality, usually featured by facial expressions [166, 167], is one of the most dominant modalities for emotion recognition. By utilizing either a finely hand-crafted descriptor, e.g., facial action coding system (FACS) [168], or a powerful convolutional neural network, e.g., the Resnet for feature extraction, an emotion recognition method can achieve promising results. In recent years, Electroencephalography (EEG) has drawn considerable attention from researchers [169], due to its simple, cheap, portable, and easy-to-use solution for identifying emotions [170]. In addition to the visual and EEG information, the audio/speech, text, and some other physiological signals (e.g., heart rate, blood pressure, and eye gaze) are also widely used.

Two general differences between the visual and EEG modalities are of the most relevance to our interest. First, facial expressions and gestures are overt and determined, whereas the EEG signal is covert and highly subject-dependent. As a result, it is feasible to directly label the emotion based on the visual modality from an annotator, yet for EEG modality it is done either by predefined experiment protocol or by subjects themselves. Second, the visual modality usually has high and low resolutions on spatial and temporal dimensions (e.g., $40 \times 40 \times 3$ and 30 fps, respectively, whereas the EEG modality is high on temporal resolutions (e.g., 256 Hz) yet low on spatial resolutions (e.g., 32 electrodes). The greater the resolution is, the more detailed structural or phase changes in response to emotional stimuli can be studied. Based on the differences in modalities and the assumption that incorporating multimodal data will produce results that are superior to unimodal

data, it is natural to utilize the multimodal data which can essentially increase the amount of available data and hopefully attenuate the defects of each modality.

Knowledge distillation (KD) is one of the promising solutions to combining multimodal data. In deep learning, KD is an effective technique that has been widely used to transfer information from one network to another network whilst training constructively [171]. Many cross-modal KD methods have been proposed to leverage the synchronization of visual and audio information in the video data. A joint embedding can be learned by distilling the knowledge between RGB/depth, face/voice, and CT/MRI images. However, to the best of the authors' knowledge, there is no prior work relevant to visual-to-EEG cross-modal KD on CER.

We, therefore, pose a question: Can the CER performance of the EEG modality be improved if we transfer the knowledge from the visual modality? Given a dataset containing synchronous facial videos and EEG signals of different subjects, the facial video modality tends to have stronger relevance with respect to the expert-labeled continuous trace. The reasons are two-fold. First, the experts conduct the labeling according to the subject's facial expression. Second, the EEG signal has a low information-to-noise ratio, large bias can be existing among signals recorded at a different time or from different subjects. According to the results in Table 3.2 and 3.3, we can see a performance lead of about 0.30+ in CCC when comparing the visual model against the EEG model. It suggests that on the MAHNOB-HCI dataset, the visual model is potentially more powerful for the CER task. It inspires us to teach the EEG modality using the visual knowledge.

In this work, we explore to what extent can the EEG modality gain from the visual modality using the cross-modal knowledge distillation (CKD) for CER. A teacher model is firstly trained in the visual modality using the facial video. Its intermediate features, a.k.a. dark knowledge, are then used to supervise the student model training in an offline manner. Specifically, the teacher and student models comprise a cascade spatiotemporal and a single temporal network, respectively. The inputs to the teacher and student are facial video frames and the synchronous EEG average band power. The temporal embeddings from the trained teacher's temporal component are taken as dark knowledge. During the training of the student, its temporal embeddings are guided by the dark knowledge using L1 loss. Together with the concordance correlation coefficient (CCC) loss, which punishes

the inconsistency between the prediction and label sequences by scaling the correlation coefficient with their mean square difference, the student is able to learn from the visual and EEG modalities simultaneously. (A formal definition of CCC is provided by Eq. 2.3.) During the test of the student, it infers based on the EEG modality and the learned visual knowledge. Results from experiments manifest statistical significance (p-value < 0.01 or 0.05 depending on the data partitioning scheme) on root mean square error (RMSE), Pearson correlation coefficient (PCC), and CCC, compared to its counterpart without KD.

This chapter is structured as follows. First, the methodology is formulated mathematically, followed by the demonstration of the model architecture. Next, the implementation details, including the database, feature extraction, and data partitioning, are introduced. Next, the settings of parameters, hyperparameters, and training are elaborated. After which, the results against the state-of-the-art CER method are reported. Finally, we visualize the skull saliency map for each band and subject, based on the trained student model and the peak response mapping [153] (PRM), as we are also particularly interested in revealing the contribution of each brain lobe and the band of EEG towards the emotion process (as we introduced in Section 2.2.3).

5.2 Related Works

5.2.1 Cross-modal Knowledge Distillation (CKD)

Cross-modal knowledge distillation (CKD) uses the teacher's representation as a supervision signal to train the student to learn another task [171]. It is helpful especially when the data or labels for the target modalities are hard to get. Based on the hypothesis that the emotional content of speech correlates with the facial muscle movement and facial expression of the speaker, Afouras et al. [172] transfer voice knowledge to train lip reading-based visual speech recognition models, while Nagrani et al. [173] transfer the visual knowledge to learn voice feature-based speech classification, both of which are without access to any form of human-labeled ground truth. Hoffman et al. [174] utilize the RGB information to teach a depth network, and fuse the information across modalities. Gupta et al. [175] learn a student model on unlabeled depth images and optical flow by transferring

the knowledge of a teacher model trained on well-annotated RGB images. Zhao et al. [176] use radio data to guide human pose estimation on occluded images. Thoker and Gall [177] employ paired RGB videos and skeleton sequences for CKD. The knowledge learned on RGB videos is transferred to the student model for skeleton-based human action recognition. Garcia et al. [178] use additional depth images to generate a hallucination stream for RGB image modality and thereby improve the action recognition performance. Tian et al. [179] employ a contrastive loss to transfer relation-based knowledge across modalities. Roheda et al. [180] use generative adversarial networks (GAN) for distillation among the missing and available modalities. We see that most of the CKD methods are for target detection and action recognition. It is rarely explored in the area of CER.

5.2.2 Multimodal CER Methods

The term "continuous" possesses two characteristics in our context. Spatially, it aims to place the emotional state as a continuous-valued point in the multi-dimensional space of the dimensional theory, instead of choosing categorical labels. Temporally, it continuously predicts the emotional state for a fixed time interval, constituting the emotional trace of the subject over a specified time span.

The CER has always been challenging due to the following causes. First, the emotion itself is highly subjective and subject-dependent. For example, the perception of emotion is influenced by individual experiences. Physically abused children are much quicker than other children to spot the signals of anger [13]. As a result, the data from the subjects and the ground truth from the annotators are prone to personal bias. Multimodality and Transfer learning among visual, audio, and physiological data are two promising techniques to alleviate this issue and develop reliable CER models. Second, by taking the facial muscular movement as actions, the complex emotion cues over a large time span are a composition of complex one-actions [181]. Typical one-actions can be defined by FACS [182] that codes the movements of individual facial muscles. However, as atomic as the FACS may be, human emotion, no matter from which modality it is observed, usually exhibits large variations in terms of intensity and order in their duration and takes longer to unfold. Models which can learn the long-range temporal dependencies are in need to counter this issue.

In Section 2.8, we have provided a completed review of a number of CER methods. We see that most CER methods are based on visual and audio modalities. To the best of our knowledge, out of all the publicly available CER databases, there is only a subset [3] of the MAHNOB-HCI database [2] where the facial video, EEG signal, and continuous valence label are available. Our work is based on this subset of the MAHNOB-HCI database.

Here we highlight the two methods that are the most relevant to ours, i.e., Soleymani et al. [3] and Chen et al. [117]. Soleymani et al. [3] propose a multimodal method for continuous valence prediction based on facial landmark sequence and EEG signal. A long short-term memory (LSTM) network is used for feature learning. The features from the two modalities are fused using feature-level and decision-level fusion schemes before feeding into the fully-connected layers. Chen et al. [117] combine a pretrained 2D-CNN and a TCN to learn deep spatiotemporal features from video frames and audio spectrograms, and use a spatiotemporal graph convolutional network to encode facial landmarks graph. Finally, a bidirectional LSTM network is employed for unimodal and multimodal predictions.

The differences are explained as follows. First, concerning the motivation, our work intends to investigate the CKD on visual and EEG modalities, while the two papers are for multimodal feature fusion. In the case where the visual information is not available, our model can still work and infer based on the EEG signal and the learned visual knowledge. Second, our visual model comprises a cascade 2DCNN-TCN architecture. The produced spatiotemporal features are directly fed to a linear layer to infer. Whereas in [117], the 2DCNN-TCN is first trained as a feature extractor. An independent bidirectional LSTM network is then trained on top of the extracted features to infer. Third, our EEG model use TCN to learn the temporal encoding of the EEG band power, while in [3] an LSTM network is used for the same purpose.

5.3 Methodology

Suppose that we are given one sequence of emotion cue $\mathbf{X} = \{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{N \times d}$, and the corresponding ground truth $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^N$, both of which contain N sample points. We are also given a train teacher model T from

the visual modality, and a fresh student model S from the EEG modality. Our goal is to find a function $f : \mathbf{X} \to \mathbf{Y}$ so that its output $\hat{\mathbf{Y}} = f(\mathbf{X}) \in \mathbb{R}^T$ minimizes some loss $L(\mathbf{Y}, \hat{\mathbf{Y}})$ against the ground truth $\mathbf{Y}$. The function is found by training the student model S. Note that the causal constrain is applied, which requires that y_t depends only on $x_1, \ldots, x_t$ and not on any future inputs $x_{t+1}, \ldots, x_T$.

5.3.1 Cross-modal Knowledge Distillation

The knowledge distillation requires the teacher and student to interact through interchanging their knowledge. The teacher's knowledge $\mathbf{U} \in \mathbb{R}^{T \times F}$ and the student's knowledge $\mathbf{V} \in \mathbb{R}^{T \times F}$ refer to intermediate features produced by some certain layers of the model that are in the same dimension.

The interaction is done by aligning the student's knowledge $\mathbf{V} \in \mathbb{R}^{N \times F}$ onto the fixed teacher's knowledge $\mathbf{U} \in \mathbb{R}^{N \times F}$. Inspired by Romero et al. [183] which distills the knowledge by enforcing the proximity of intermediate feature maps using the L2 loss, we further use the sparser L1 loss as the KD loss, in order to produce a more reasonable magnitude relevant to the CCC loss and makes the training more controllable. The L1 loss is formulated as follows.

$$L_1(\mathbf{U}, \mathbf{V}) = \frac{1}{TF} \sum_{i=1}^{T} |u_i - v_i|, \tag{5.1}$$

where $u_i \in \mathbb{R}^F$ and $v_i \in \mathbb{R}^F$ are the feature points in each time step.

The final loss function is designed as a weighted sum of CCC and L1 loss as follows:

$$\ell(\mathbf{X}, \mathbf{Y}, \mathbf{U}, \mathbf{V}) = 1 - \rho_c(\mathbf{X}, \mathbf{Y}) + w \cdot L_1(\mathbf{V}_t, \mathbf{V}_s) \tag{5.2}$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the predictions and the labels, and $\mathbf{U}$ and $\mathbf{V}$ denote the spatiotemporal features of the teacher and student model, respectively, with the constant w being the trade-off. The CCC loss ρ_c is formulated at Eq. 3.4. The grid searching is employed to find the optimal w.

5.3.2 Model Architecture

The goal of the visual-to-EEG KD is to use the visual knowledge (i.e., the spatiotemporal visual features produced by the TCN of the visual model) along with the labels to train an improved EEG model. The interaction between the teacher and student is illustrated in Fig. 5.1.

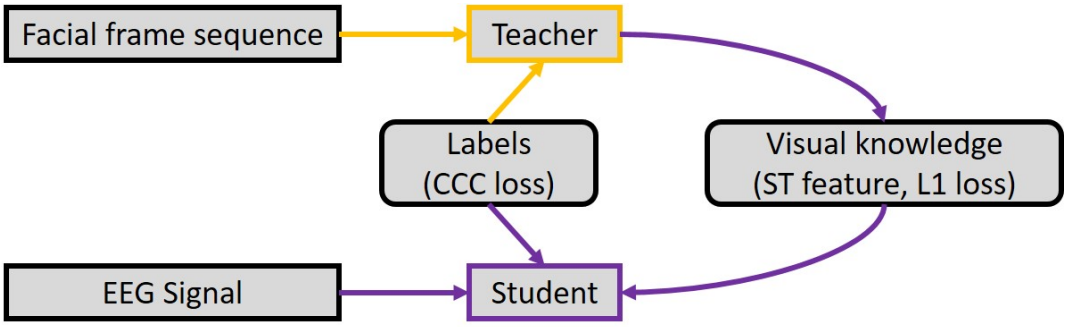

Figure 5.1: The illustration of the teacher-student interaction. The figure shows the 2-stage teacher-student interaction for the cross-modal CKD. ST feature denotes the spatiotemporal features. The training of the teacher and student models are colored in yellow and purple, respectively.

Two stages are involved in the teacher-student interaction. In the first stage, the teacher model is trained by minimizing the CCC loss function between the frame sequences and the corresponding labels. In the second stage, the trained teacher model is used to extract the spatiotemporal features of the visual modality, and the student model is then trained using the EEG Signal, the corresponding labels, and visual spatiotemporal features.

5.4 Implementation Details

The implementation details are carried out following the same procedure introduced in Section 3.4. On top of this, the feature synchronization is the only new step that requires detailing.

Recall that after the data preprocessing, we have facial frames, EEG band power features, and continuous valence labels at the frequency of 64Hz, 4Hz, and 4Hz, respectively. In order to synchronize the facial frames with the other two features, downsampling is employed. A consecutive 16 frames are taken as one group, corresponding to one valence label point. During the teacher training, the n-th frame

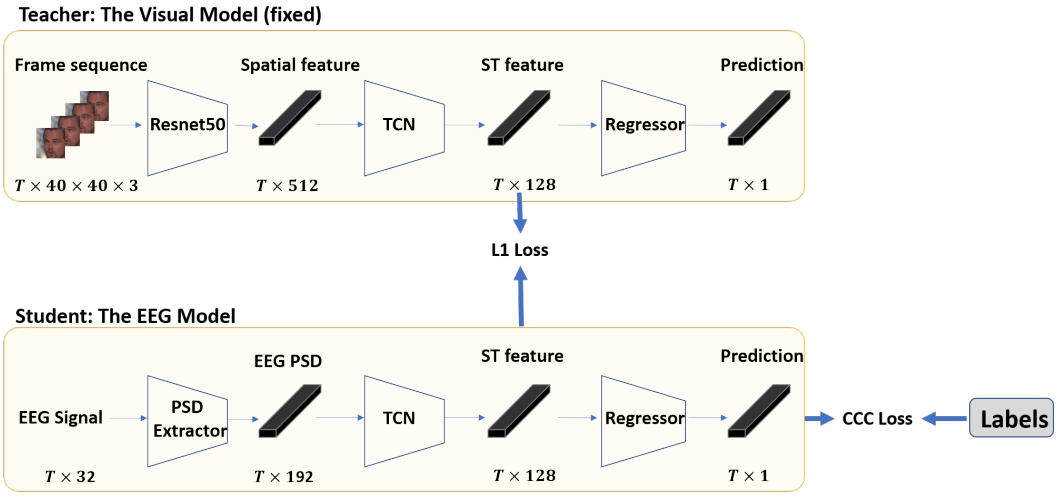

Figure 5.2: The illustration of the cross-modal knowledge distillation model. The figure shows the architecture of the teacher and student models.

for each group is loaded in sequential and fed to the teacher model. The integer n is randomly chosen from 0 to 15 for each epoch. For the inference, only the 0-th frames of each group are loaded.

The visual knowledge is generated by feeding the facial frames without downsampling to the trained teacher model. The generated visual knowledge is at the frequency of 64Hz. During the student training, the same downsampling scheme used on the video frames is applied to the visual knowledge for the synchronization with the EEG band power features and continuous valence labels.

Table 5.1: The result of our EEG model taught by visual knowledge against the standalone counterpart using the TRS partitioning. The mean, standard deviation, and p-value are reported. The p-value is obtained using the one-tailed paired t-test over the 10-fold TRS partitioning. TRS: trial-wise random shuffling. ↑: the higher the better. ↓: the lower the better. ⋆: 0.01 <p-value≤ 0.05. ⋆⋆: 0.001 <p-value≤ 0.01. ⋆⋆⋆: p-value≤ 0.001. Bold fonts indicate the best results.

TRS		Without KD	$w = 0.2$		$w = 0.4$	
		mean±std	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.067±0.005	0.066±0.005	0.001 (⋆⋆⋆)	0.066±0.005	0.004 (⋆⋆)
	PCC ↑	0.463±0.103	0.469±0.103	0.019 (⋆)	0.469±0.105	0.003 (⋆⋆)
	CCC ↑	0.444±0.109	0.450±0.110	0.014 (⋆)	0.449±0.111	0.030 (⋆)
Test	RMSE ↓	0.066±0.009	0.066±0.010	0.013 (⋆)	0.065±0.010	0.010 (⋆⋆)
	PCC ↑	0.435±0.205	0.442±0.205	0.013 (⋆)	0.442±0.204	0.011 (⋆)
	CCC ↑	0.415±0.201	0.422±0.202	0.022 (⋆)	0.422±0.201	0.015 (⋆)

TRS		$w = 0.6$		$w = 0.8$		$w = 1.0$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.065±0.005	0.004 (⋆⋆)	0.065±0.005	0.001 (⋆⋆⋆)	**0.065 ± 0.005**	**< 0.001** (⋆⋆⋆)
	PCC ↑	0.473±0.107	0.001 (⋆⋆⋆)	0.477±0.108	0.001 (⋆⋆⋆)	**0.478 ± 0.107**	**< 0.001** (⋆⋆⋆)
	CCC ↑	0.453±0.114	0.011 (⋆)	0.457±0.115	0.005 (⋆⋆)	**0.458 ± 0.113**	**0.002** (⋆⋆)
Test	RMSE ↓	0.065±0.010	0.002 (⋆⋆)	0.065±0.010	0.003 (⋆⋆)	**0.065 ± 0.009**	**< 0.001** (⋆⋆⋆)
	PCC ↑	0.450±0.202	0.001 (⋆⋆⋆)	0.457±0.204	0.008 (⋆⋆)	**0.454 ± 0.207**	**0.010** (⋆⋆)
	CCC ↑	0.428±0.199	0.002 (⋆⋆)	0.436±0.202	0.011 (⋆)	**0.433 ± 0.205**	**0.010** (⋆⋆)

TRS		$w = 1.2$		$w = 1.4$		$w = 1.6$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.065±0.005	< 0.001 (⋆⋆⋆)	0.064±0.005	< 0.001 (⋆⋆⋆)	0.064±0.005	< 0.001 (⋆⋆⋆)
	PCC ↑	0.479±0.108	< 0.001 (⋆⋆⋆)	0.480±0.109	0.002 (⋆⋆)	0.481±0.108	< 0.001 (⋆⋆⋆)
	CCC ↑	0.459±0.114	0.004 (⋆⋆)	0.461±0.116	0.009 (⋆⋆)	0.461±0.115	0.006 (⋆⋆)
Test	RMSE ↓	0.064±0.009	< 0.001 (⋆⋆⋆)	0.064±0.009	< 0.001 (⋆⋆⋆)	0.064±0.009	< 0.001 (⋆⋆⋆)
	PCC ↑	0.456±0.206	0.001 (⋆⋆⋆)	0.459±0.205	0.004 (⋆⋆)	0.460±0.205	0.002 (⋆⋆)
	CCC ↑	0.436±0.204	0.005 (⋆⋆)	0.437±0.204	0.007 (⋆⋆)	0.438±0.204	0.006 (⋆⋆)

TRS		$w = 1.8$		$w = 2.0$			
		mean±std	p-value	mean±std	p-value		
Validation	RMSE ↓	0.064±0.005	0.003 (⋆⋆)	0.064±0.005	0.005 (⋆⋆)		
	PCC ↑	0.476±0.103	0.084	0.475±0.105	0.140		
	CCC ↑	0.453±0.108	0.317	0.452±0.110	0.415		
Test	RMSE ↓	0.064±0.009	0.028(⋆)	0.064±0.009	0.024 (⋆)		
	PCC ↑	0.450±0.207	0.252	0.450±0.207	0.249		
	CCC ↑	0.427±0.206	0.369	0.427±0.206	0.377		

Table 5.2: The result of our EEG model taught by visual knowledge against the standalone counterpart using the LOSO partitioning. The mean, standard deviation, and p-value are reported. The p-value is obtained by using the one-tailed paired t-test over the 24-fold LOSO partitioning. LOSO: leave-one-subject-out. ↑: the higher the better. ↓: the lower the better. ⋆: 0.01 <p-value≤ 0.05. ⋆⋆: 0.001 <p-value≤ 0.01. ⋆⋆⋆: p-value≤ 0.001. Bold fonts indicate the best results.

LOSO		Without KD	$w = 0.2$		$w = 0.4$	
		mean±std	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.068±0.007	0.067±0.006	0.096	0.067±0.006	0.002 (⋆⋆)
	PCC ↑	0.467±0.116	0.475±0.110	0.225	0.480±0.111	0.039 (⋆)
	CCC ↑	0.445±0.118	0.451±0.115	0.259	0.454±0.115	0.050 (⋆)
Test	RMSE ↓	0.066±0.025	0.065±0.025	0.059	**0.063 ± 0.025**	**0.001** (⋆⋆⋆)
	PCC ↑	0.474±0.267	0.480±0.269	0.034 (⋆)	**0.482 ± 0.269**	**0.014** (⋆)
	CCC ↑	0.377±0.250	0.382±0.250	0.319	**0.387 ± 0.253**	**0.033** (⋆)

LOSO		$w = 0.6$		$w = 0.8$		$w = 1.0$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.067±0.006	0.004 (⋆⋆)	0.067±0.006	0.004 (⋆⋆)	0.066±0.006	0.004 (⋆⋆)
	PCC ↑	0.477±0.111	0.107	0.477±0.111	0.107 (⋆)	0.479±0.112	0.067
	CCC ↑	0.452±0.115	0.149	0.452±0.115	0.149	0.454±0.115	0.084
Test	RMSE ↓	0.064±0.025	0.003 (⋆⋆)	0.064±0.025	0.003 (⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)
	PCC ↑	0.482±0.268	0.022 (⋆)	0.482±0.268	0.022 (⋆)	0.481±0.270	0.030 (⋆⋆)
	CCC ↑	0.383±0.251	0.177	0.383±0.251	0.177	0.385±0.254	0.084

LOSO		$w = 1.2$		$w = 1.4$		$w = 1.6$	
		mean±std	p-value	mean±std	p-value	mean±std	p-value
Validation	RMSE ↓	0.066±0.006	< 0.001 (⋆⋆⋆)	0.066±0.006	< 0.001 (⋆⋆⋆)	0.066±0.006	< 0.001 (⋆⋆⋆)
	PCC ↑	0.484±0.111	0.015 (⋆)	0.484±0.115	0.016 (⋆)	0.486±0.114	0.011 (⋆)
	CCC ↑	0.458±0.116	0.012 (⋆)	0.458±0.118	0.015 (⋆)	0.459±0.118	0.012 (⋆)
Test	RMSE ↓	0.064±0.024	0.003 (⋆⋆)	0.064±0.024	0.001 (⋆⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)
	PCC ↑	0.481±0.271	0.037 (⋆)	0.480±0.270	0.200	0.481±0.270	0.150
	CCC ↑	0.381±0.253	0.324	0.378±0.250	0.798	0.380±0.251	0.504

LOSO		$w = 1.8$		$w = 2.0$			
		mean±std	p-value	mean±std	p-value		
Validation	RMSE ↓	**0.065 ± 0.006**	**< 0.001** (⋆⋆⋆)	0.065±0.006	< 0.001 (⋆⋆⋆)		
	PCC ↑	**0.487 ± 0.116**	**0.009** (⋆⋆)	0.485±0.116	0.017		
	CCC ↑	**0.460 ± 0.120**	**0.011** (⋆)	0.457±0.118	0.035		
Test	RMSE ↓	0.063±0.024	< 0.001 (⋆⋆⋆)	0.063±0.024	< 0.001 (⋆⋆⋆)		
	PCC ↑	0.481±0.270	0.140	0.479±0.271	0.220		
	CCC ↑	0.381±0.251	0.483	0.382±0.253	0.437		

5.5 Results and Analysis

Based on Eq. 5.2, the grid search is employed for w ranging from 0.2 to 2.0, with a step of 0.2. All the other settings remain the same.

The results of the student with and without CKD using TRS and LOSO partitioning are reported in Table 5.1 and Table 5.2, respectively. In the interval of $0.2 \leq w \leq 1.0$, the best validation results are found when $w = 1.0$ and $w = 0.4$ for TRS and LOSO, respectively. They also lead to the best test results with statistical significance (p-value ≤ 0.01 and p-value < 0.05 on the three metrics for TRS and LOSO partitioning, respectively). When $1.0 < w \leq 2.0$, though better validation results are yielded for LOSO, the corresponding test results are without statistical significance.

Comparing the results from TRS partitioning against LOSO partitioning, we can see that the former tend to have more stars (i.e., smaller p-values) and more consistent metrics between the validation and test set. LOSO is prone to over-fitting when $w > 1.0$. We can therefore infer that the CKD using TRS partitioning is more effective. Indeed, it can be explained that both the teacher and student have seen examples similar to the test examples, and therefore produce joint embeddings that are of greater representability during the testing.

5.6 Discussion and Conclusion

The goal of CER is to continuously predict the emotional trace in the multi-dimensional space over a specified time span. However, the recognition of emotion, if driven by data, suffers from the subject bias that exhibits in multiple stages of the emotion process. The issue is escalated for physiological signals, compared to the more objective and determinant cues from visual or audio modalities. Also, emotion cues over a large time span are a composition of complex one-actions, manifesting large variations in intensity and order in their duration. A model that is capable of capturing long-range dependencies is crucial in this area.

In this chapter, on top of the visual and EEG unimodal CER model, we explore the idea of teaching an EEG-based CER model using the visual knowledge from a

visual-based CER model. The teacher model features a cascade CNN-TCN architecture and is fed by video frames. A subset [3] of the MAHNOB-HCI database [2] that includes facial videos, EEG signals, and continuous valence labels of 24 subjects is employed for the experiment. Two data partitioning schemes, i.e., the TRS and LOSO are employed. The results reported in Table 3.2 and Table 3.3 validated the performance of the standalone teacher and student models in visual and EEG modalities and obtained promising results compared to the baseline. After which, the spatiotemporal feature of the trained teacher is taken as the dark knowledge. The latter, together with the continuous label, is used to teach the student model. The experiment using the TRS and LOSO partitioning schemes both show an increase with statistical significance, i.e., p-value < 0.01 for TSR and p-value < 0.05 for LOSO partitioning on RMSE, PCC, and CCC.

The improvement is obtained thanks to the complementary nature of multi-modality, and the transfer of knowledge between the modalities can help to leverage this complementary information. The visual-based model may have learned to detect and interpret visual cues related to facial expressions, which can provide important information about the emotional state. The EEG-based model, on the other hand, may have learned to detect and interpret neural activity associated with emotions, which can provide additional complementary information about the emotional state. By transferring knowledge from the visual-based model to the EEG-based model, the latter may be able to leverage this complementary information and improve its performance. The visual-based model may provide a guide for the EEG-based model to learn how to interpret and extract relevant features from EEG signals that are related to emotional states. Additionally, the transfer of knowledge may help to regularize the EEG-based model and prevent overfitting. The visual-based model may provide a regularization effect by encouraging the EEG-based model to learn a more general representation of the emotional state that is not specific to the EEG modality. Overall, the cross-modal knowledge distillation approach for continuous emotion recognition can work by leveraging complementary information between modalities and providing regularization to the target modality, which can result in improved performance with statistical significance.

Chapter 6

Multimodal CER through Leader-follower Attentive Fusion

This chapter [1] presents the multimodal feature fusion for CER. Continuous emotion recognition (CER) aims to sequentially map a subject's emotional recordings to a real-valued space axed by valence and arousal. Temporal dynamic learning and cross-subject generality are two crucial cornerstones for developing a reliable deep learning model for CER, based upon which we propose the leader-follower attentive network (LFAN). It extends our previous work for the affective behavior analysis in-the-wild (ABAW) contest. Our LFAN aims to learn the per-modality long-term dependencies first and then combine the learned encodings using the cross-modal co-attention mechanism. Specifically, the temporal convolutional network (TCN) is employed for the per-modality long-term dependency learning. A window length of 300 is employed to resample the sequential input feeding the TCN, resulting in abundant historical information for our LFAN to reveal the emotional roller coaster. The learned per-modality encodings are fused using the cross-modal co-attention block, which is lightweight and capable of weighing across the modalities for each time step. We believe that the visual modality has the strongest correlation with the label. To emphasize such dominance, the visual encoding outputted by the visual TCN is taken as the leader and is concatenated to the fused encoding. Experiments on AVEC2019, MAHNOB, and AffWild2 databases are carried out,

[1] The work in this chapter has been published in [127, 143], and has extended to a full paper and is currently under review in IEEE Transactions on Affective Computing.

where our LFAN achieves promising results compared to its variants and state-of-the-art methods.

6.1 Introduction

Emotion recognition is the process of understanding human emotion. It plays a vitally crucial role in many fields such as human-computer interaction, behavioral modeling, opinion mining, psychological health, business intelligence, and entertainment assistance. Continuous emotion recognition (CER) is one of the sub-tasks in this area that aims for the sequential regression of the given emotion cues. It contrasts sharply with other research topics like emotion classification where the whole trial is usually annotated categorically.

A reliable deep learning model for CER should be able to model the temporal dynamics and preserve the cross-subject generality. In Chapter 3 we have discussed the former, that is, the emotion cues over the time span $T = \{0, 1, \ldots, t\}$ are a composition of t cues, each has its own valence and arousal values. And the cue at the time step t is not only the direct successor of that at $t-1$, but also the effect of the emotional roller coaster over all the predecessors. A more reliable prediction could be made by considering the emotional dynamics over a large time window. In this chapter, we involve cross-subject generality. Emotion, as intrinsic as it may be for every one of us, is highly prone to cross-subject bias. Emotion is triggered by conscious and/or unconscious perception of an event and is usually associated with personal attributes such as mood, personality, and experience. For example, physically abused children are much quicker than other children to spot the signals of anger [13]. The expression of emotion involves vocal emotional words and non-vocal cues such as facial expressions, voice intonation, and body movement, resulting from the compound effects of physiological arousal, individual feelings/behaviors, and cultural/moral regulation. These factors could lead to a large subjective bias and degenerate representation learning.

To tackle the issue of temporal dynamic learning and cross-subject generality, we employ large window resampling and multimodal fusion. For the former, as we discussed in Chapter 3, the resampling window determines the length of context

the model draws to make a prediction for each time step. As for multimodal fusion, it is widely accepted that incorporating multimodal data would yield superior predictions than that from unimodal data, thanks to their complementary nature. Indeed, on the one hand, multimodal data help to disambiguate. A crying face with joyful vocal expressions would be recognized as happiness instead of sadness, and a neutral face with a harsh intonation might be a sign of anger or contempt. On the other hand, multimodal data helps to preserve recognition robustness. In natural settings where the environment is uncontrolled, subject diversity, various illumination conditions, spontaneous behaviors, background noise, and unclear speech are ubiquitous. In a scene where the actor's face is not in the camera, their voice or body gesture could serve as an emotional cue and support the ongoing learning.

We present the leader-follower attentive network (LFAN) to achieve temporal dynamic learning and cross-subject generality. Our LFAN aims to learn the per-modality long-term dependencies first and then combine the learned encodings using the cross-modality co-attention mechanism. Specifically, the temporal convolutional network (TCN) [154] is employed for the per-modality long-term dependency learning. With the dilated convolutional kernel and stacked residual blocks, the TCN is capable of looking very far into the past to make a prediction [154]. Ablation studies show that the TCNs alone are capable of long-term dependency learning without employing the temporal self-attention, which is of T^2 complexity. The learned per-modality encodings are fused using the cross-modal co-attention block. The latter is lightweight and is capable of weighing across the modalities for each time step. To emphasize the dominant visual modality, which we believe to have the strongest correlation with the label, the visual encoding outputted by the visual TCN is taken as the leader and is concatenated to the fused encoding. Experiments on AVEC2019, MAHNOB-HCI, and AffWild2 databases are carried out, where our LFAN achieves promising results compared to state-of-the-art methods.

The contributions of this work include the following.

- The LFAN, which is a multimodal deep neural network, is developed for CER. It consists of parallel TCNs and a cross-modal co-attention block. The TCNs learn the long-term dependency and the fusion block combines the temporal encodings using the co-attention mechanism. LFAN can achieve temporal dynamic learning and cross-subject generality.

- Our previous work [143] is extended with extensive ablation studies and experiments. The ablation studies explore the parameter/hyper-parameter settings and model properties. The quantitative experiments demonstrate the superiority of our LFAN over state-of-the-art methods. And the qualitative experiment visualizes the cross-modal co-attention.

This chapter is structured as follows. First, the methodology is formulated mathematically, followed by the illustration of the model architecture. Next, the feature synchronization, which is in addition to the implementation details introduced in Section 3.4, is introduced. After which, the results of the knowledge distillation are reported. Finally, we conclude this chapter.

6.2 Related Works

Previous studies [3, 98, 108, 117] adopt feature-level and/or decision-level fusion strategies for CER. For the former, the features from a single modality are extracted separately. They are then concatenated and fed to the fusion model to learn the cross-modal representation. For the latter, the predictions from each modality, instead of the extracted features, are fed to the fusion model. However, the feature-level fusion usually suffers from the curse of dimensionality due to the feature concatenation. Moreover, the decision-level fusion overlooks the complementarity and redundancy across different modalities, which is against the evidence [20] in neuroscience, suggesting that multimodal integration occurs at an early stage. An intermediate-level fusion strategy is therefore in need. Knowledge distillation is another technique to combine multimodal information. Unlike fusion, which requires the data from all the modalities are available for the training and testing sessions, knowledge distillation favors the situations where the information from certain domains/modalities is hard to acquire [171]. Studies show that the visual modality can teach other modalities such as audio [184] or EEG [4] and improve their performance.

Recently, a growing number of researchers have resorted to the attention mechanism and transformer [52] to achieve intermediate-level fusion. Le et al. [185] proposed the multimodal transformer networks (MTN) to generate conversational responses to the queries of humans for a video-grounded dialogue system. MTN consists of i)

transformer-encoders for representation learning, ii) transformer-decoders for reasoning over the learned representation using the multi-head attention mechanism, and iii) auto-encoder layers for the emphasis of query-related video features. Tsai et al. [186] employ the co-attention module to learn representations from paired modalities, which retrieves the temporal dependency of the unaligned modalities. Zadeh et al. [187] proposed to combine the uni-modal, bi-modal, and tri-modal feature sequences into one feature sequence and reconstruct them into multiple streams of the same dimension. Every individual stream is fed to its own transformer encoder, and all the streams' outputs are merged point-wise. Chen et al. [117] proposed the transformer-encoder with a multimodal multi-head attention framework to learn the intra/inter modality dynamics. The intra-modality transformer aims to learn the temporal dynamics within one modality, after which the inter-modal transformer is used to perform the fusion.

Inspired by the co-attention mechanism where the query from one modality can be combined with the key/value from other modalities, we proposed the leader-follower attentive network (LFAN) to perform the intermediate-level multimodal fusion. Our LFAN only utilizes TCNs to learn the temporal dependency of each modality. The latter is then fed to the fusion block to learn the cross-modal dynamics using the co-attention mechanism. Specifically, given n feature sequences from n modalities, the n query, n key, and n value vectors are concatenated to form one single query, key, and value, respectively. Upon which the attention operation is performed yielding the co-attention feature. The feature sequence from the visual modality is concatenated to the latter to emphasize its leading status. We argue that the temporal transformer, which is of $T \times T$ complexity, might not be necessary for CER, especially when the sampling window is large. We validated this hypothesis through ablation studies using the AVEC2019 databases.

6.3 Methodology

Suppose that we are given synchronous sequences of emotional cues $\mathbf{X} = \{\mathbf{X}^{(m)}\}_{m=1}^{M}$ from M modalities, where $\mathbf{X}^m \in \mathbb{R}^{T \times d_m}$, and the corresponding ground truth $\mathbf{Y} = \{y_1, y_2, \ldots, y_T\}$, where $\mathbf{Y} \in \mathbb{R}^T$. In each modality, the sequence $\mathbf{X}^{(m)} =$

$\{x_1^{(m)}, x_2^{(m)}, \ldots, x_T^{(m)}\}$ contains T samples, where T denote the length of the re-sampling window. Our goal is to find a function $f : \mathbf{X} \rightarrow \mathbf{Y}$ so that its output $\hat{\mathbf{Y}} = f(\mathbf{X}) \in \mathbb{R}^T$ minimizes some loss $L(\mathbf{Y}, \hat{\mathbf{Y}})$ against the ground truth $\mathbf{Y}$. Note that the causal constrain is applied, which requires that y_t depends only on $x_1, \ldots, x_t$ and not on any future inputs $x_{t+1}, \ldots, x_T$.

6.3.1 Temporal Modelling

To learn the temporal dependency of the given sequences, some feature extractors b are firstly used upon the raw input data, updating $x_t \leftarrow b(x_t)$ for each sample, so that their representability is increased. Depending on the data modalities, the actual extractors b are varied. For example, a convolution neural network (CNN) backbone trained on a large-scale 2D face database could be used to extract the deep facial feature, and the average band power can be used as the low-level EEG feature, etc.

Several deep neural networks are capable of the temporal modeling, such as the long-short term memory (LSTM) and the TCN. Systematical comparison [154] demonstrated that TCNs convincingly outperform recurrent architectures across a broad range of sequence modeling tasks. With the dilated and casual convolutional kernel and stacked residual blocks, the TCN is capable of looking very far into the past to make a prediction. Therefore, we employ the TCN as our temporal model.

A typical TCN consists of a stack of 1D dilated convolutions, and residual connections, which are formulated as:

$$TCN(\mathbf{X}^{(m)}) = Activation(\mathbf{X}^{(m)} + \sum_{i=0}^{k-1} f(i) \cdot \mathbf{X}_{s-d \cdot i}^{(m)}), \tag{6.1}$$

where k, s, and d denote the kernel size, stride, and dilation, respectively. $s - d \cdot i$ stands for the direction of the past. The receptive field is therefore determined by Eq. 3.2.

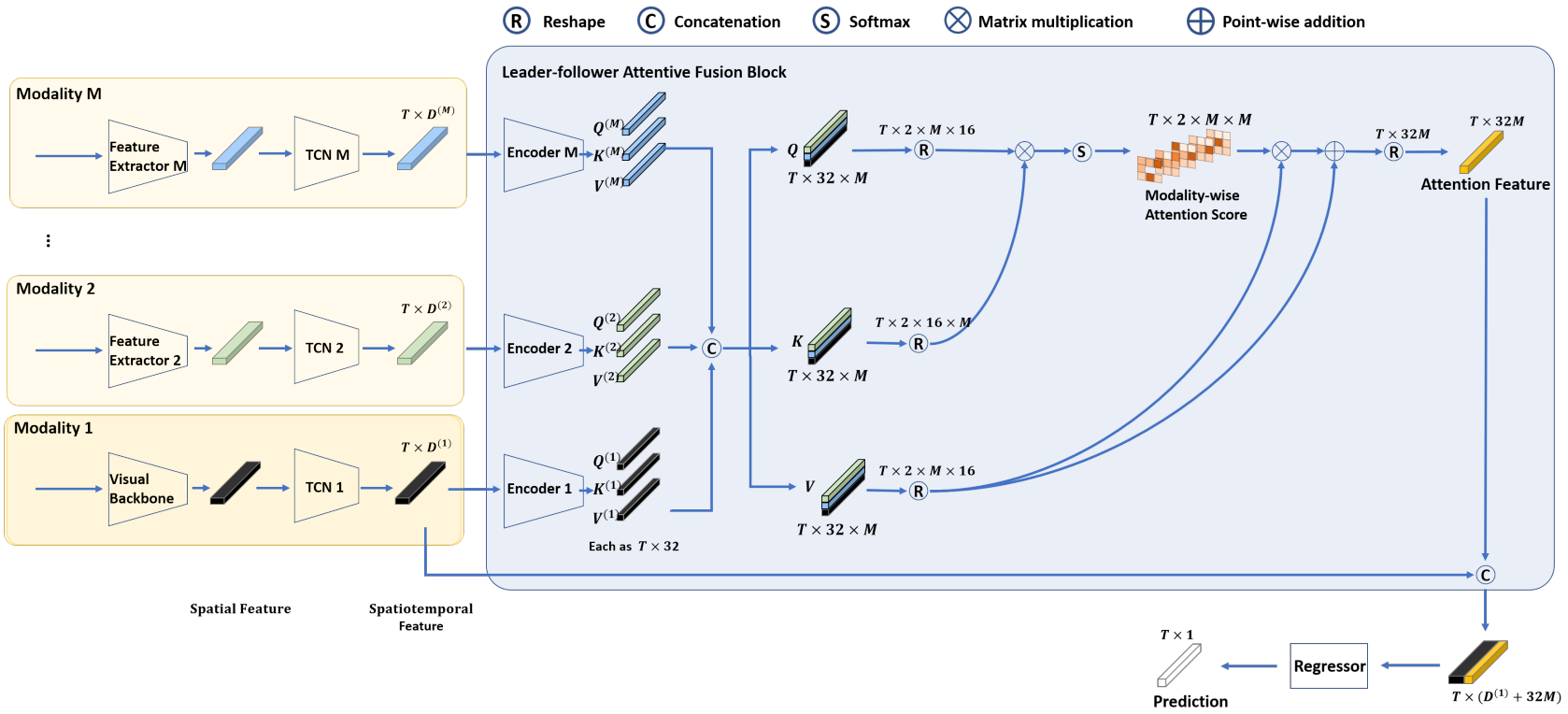

Figure 6.1: The architecture of our proposed model. The model consists of $M + 1$ components, i.e., the temporal modeling blocks and leader-follower attentive fusion block. For each temporal modeling block, it consists of the feature extractor and a TCN. The M branches yield M independent spatiotemporal feature vectors. They are then fed to the attentive fusion block. M independent attention encoders are used. For the i-th branch, its encoder consists of three independent linear layers, they adjust the dimension of the feature vector producing a query $\mathbf{Q}^{(i)}$, a key $\mathbf{K}^{(i)}$, and a value $\mathbf{V}^{(i)}$. They are then regrouped and concatenated to form the cross-modal counterparts. For example, the cross-modal query $\mathbf{Q} = [\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots]$. An attention score is obtained by Eq. 6.4.

6.3.2 Leader-follower Attentive Fusion

The leader-follower attentive fusion combines the cross-modal encoding with the temporal encoding of the leading modality. The former contains the dynamics among M modalities so that it could look to important and expressive ones at each time step t. The latter forces the model to prioritize the most dominant modality.

Given the temporal encoding $\mathbf{X}^{(m)}$ from the m-th modalities produced by the m-th TCN, the leader could be chosen empirically. For example, for a database whose continuous annotations are only available for those time steps with the subject's face appeared, the visual modality would be the leader. We represent the 1-st modality as the leader for convenience.

To produce the follower, an encoder

$$\begin{aligned} Encoder(\mathbf{X}^{(m)}) =& \mathbf{Q}^{(m)} \in \mathbb{R}^{T \times d_K}, \mathbf{K}^{(m)} \in \mathbb{R}^{T \times d_K}, \\ & \mathbf{V}^{(m)} \in \mathbb{R}^{T \times d_K} \end{aligned} \tag{6.2}$$

is firstly employed to yield the query, key, and value vectors. After which, all the queries from the M modalities are concatenated:

$$\begin{aligned} \mathbf{Q} &= concat(\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots) \in \mathbb{R}^{T \times M \times d_K} \\ \mathbf{K} &= concat(\mathbf{K}^{(1)}, \mathbf{K}^{(2)}, \ldots) \in \mathbb{R}^{T \times M \times d_K}. \\ \mathbf{V} &= concat(\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \ldots) \in \mathbb{R}^{T \times M \times d_K} \end{aligned} \tag{6.3}$$

The follower, i.e., the co-attention feature $\mathbf{A}$ is obtained by

$$\mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_K}})\mathbf{V}. \tag{6.4}$$

The final encoding is obtained by concatenating the leader with the follower. Eventually, the prediction is yielded using a regressor, which is a fully connected layer:

$$\hat{\mathbf{Y}} = FC(concat(\mathbf{X}^{(1)}, \mathbf{A})). \tag{6.5}$$

6.3.3 Model Architecture

The model architecture is illustrated in Fig. 6.1. The model contains M parallel branches for the M modalities, and the leader-follower attentive fusion block to combine the M modalities. The first modality is taken as the leader modality. The rest modalities are taken as the follower modality. For each modality, the input would first go through the feature extractor, producing the spatial features that are temporally independent to each other in the sequence. The spatial features are then fed to the TCN for spatiotemporal learning, producing the $T \times D^{(i)}$ spatiotemporal features, where i denotes the i-th modality. Now that we have M spatiotemporal features from the M branches, they are then fed to the fusion block. In which, there are M encoders, each contains three independent linear layers, mapping the modality-wise features to the query $\mathbf{Q}^{(i)}$, the key $\mathbf{K}^{(i)}$, and the value $\mathbf{V}^{(i)}$, respectively. Subsequently, all the queries, keys, and values are concatenated, obtaining the combined $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, each of which is $T \times 32 \times M$ dimensional. The standard attention operation is then carried out, producing the attention feature that is $T \times 32M$. Finally, the leader's spatiotemporal feature is concatenated with the attention feature forming the leader-follower features sized in $T \times (D^{(l)+32M}$, where the superscript l denotes the leader modality. The model specification is determined in Section 6.5.

6.4 Implementation Details

In this section, we first introduce the three databases used for the ablation studies and experiment. We then introduce the preprocessing and training details.

6.4.1 Databases

Three databases are employed in our work. The AVEC2019 database [76] (a.k.a. the SEWA database) consists of in-the-wild audiovisual recordings of the spontaneous behavior of 200 participants. Friends or relatives from German, Hungarian and Chinese cultures are paired to communicate through a dedicated video chat platform based on the participants' webcams and microphones. To elicit responses, the paired participants are required to discuss the advert they just watched for up

to three minutes. Several native speakers are invited to annotate the corresponding participants' video chat w.r.t. the emotional dimensions of valence and arousal. The final annotations are combined into a single gold-standard using the evaluator-weighted estimator-based approach. All the video recordings are sampled at 50 fps and the annotations are of 10 Hz resolution.

The MAHNOB-HCI database is a multimodal database recorded in response to affective stimuli with the goal of emotion recognition and implicit tagging research [2]. The participants are asked to watch 20 film clips, during which the synchronized recording of facial videos, audio signals, eye gaze data, EEG signals, and other physiological signals are recorded. A subset [3] of the original MAHNOB-HCI database, including 24 participants and 239 trials, is then chosen to be continuously labeled in valence. Their averages are taken as the final labels. The video recording is at 60 fps. The EEG signals are sampled using a helmet with 32 electrodes and a sampling frequency of 256 Hz. The annotations are of 4 Hz resolution.

The AffWild2 [66] database contains 564 Youtube videos of spontaneous facial behaviors of daily life in arbitrary conditions, unlike the other two databases whose scenarios are only limited to the speaking participants sitting in front of the webcam. Some scenarios include subjects giving an interesting speech in ceremonies, participating in interviews, reacting to something that brings them happiness, etc. Overall, the subjects' age, ethnicity, profession, head pose, illumination conditions, and occlusions are in a wide range. Four experts annotate the videos in valence and arousal. The final annotations are determined by firstly performing the median filtering for the annotation of a video from one annotator, and then averaging the 4 median-filtered annotations. All the videos and the corresponding annotations have the same temporal resolution, which is about 30 Hz.

6.4.2 Preprocessing

6.4.2.1 Feature Extraction

Features from the visual, audio, linguistic, and physiological modalities are extracted as follows. For the visual modality, the facial appearance and landmarks are extracted. A resnet50 is employed as the appearance extractor. It is pre-trained using the MS-Celeb-1M database [150] as a facial recognition task and

then fine-tuned using the FER+ database [188] as an emotion classification task. The OpenFace toolkit is employed as the facial landmark extractor, which attempts to extract 68 facial landmarks for each face detected in each frame. To synchronize the visual modality to the continuous annotation, the subsampling with a factor of ψ and random offset of ν is employed, where ψ denotes the ratio of the video frame rate to the annotation resolution, and ν denotes that the ν-th sample from each segment corresponding to one annotation point is sampled. For example, the video and annotation of the AVEC2019 database are in 50 fps and 10 Hz, respectively. So that during the training, only 1 frame for every 5 frames is sampled, and the offset ν is randomly set to an integer within $[0, 4]$ for each epoch, resulting in an equal amount of video frames and annotations. During the testing, the offset ν is fixed to 0. The deep CNN feature and the facial landmarks are 512-D and 136-D, respectively.

The audio preprocessing firstly converts all the videos to mono with a $16K$ sampling rate in the wav format. The VGGish features are then extracted using the pretrained Vggish model[2]. The mfcc feature is extracted using the OpenSmile toolkit as a low-level audio feature. To synchronize the audio modality to the continuous annotation, the hop length for extracting the log-mel spectrum is set to be $1/framerate$ of the video frame, so that the resulting Vggish feature is synchronized with the latter. After which, the same random subsampling technique is employed when loading Vggish for training. As for the mfcc feature, its initial frequency is fixed to 100 Hz due to the OpenSmile Setting. For each time step corresponding to the i-th annotation point, the nearest mfcc point to that time step is sampled. The Vggish and mfcc features are 128-D and 29-D, respectively.

The linguistic preprocessing is carried out as follows. The mono wav file obtained from the audio preprocessing is fed to a pretrained speech recognition model from the Vosk toolkit[3], from which the recognized words and the word-level timestamp are obtained. The recognized words are then fed to a pretrained punctuation and capitalization model from the Nvidia Nemo toolkit[4]. After which a pretrained BERT model from the Pytorch library is employed to extract the word-level linguistic features. The linguistic features are obtained by summing together the last

[2]https://github.com/harritaylor/torchvggish

[3]https://alphacephei.com/vosk/models/vosk-model-en-us-0.22.zip

[4]https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/punctuation_and_capitalization.html

four layers of the BERT model [189]. To synchronize, the word-level linguistic features are populated according to the timestamp of each word and each frame. Specifically, a word usually has a larger time span than that of a frame. Therefore, for one word, its feature is repetitively assigned to the time steps of all the frames within the time span. After which, the same random subsampling technique is employed when loading the BERT feature for training. The BERT feature is 768-D.

The EEG preprocessing is carried out as follows. Given the EEG signal of a trial, the first and last 30s of the recording which do not correspond to stimuli watching are excluded according to the database manual [5]. The signals from the 32 electrodes are then re-referenced to the average reference to enhance the signal-to-noise ratio. The default API *set_eeg_reference* from MNE toolkit [6] is used for the average reference. After which, the average band power on $(0.3-5Hz)$, $(5-8Hz)$, $(8-12Hz)$, $(12-18Hz)$, $(18-30Hz)$, and $(30-45Hz)$ is calculated. The window size and hop size for band power calculation are 2s and 0.25s, respectively. The resulted $6 \times 32 = 192$-D band power features at the frequency of $4Hz$ are therefore synchronized with the continuous valence labels. Note that the EEG preprocessing was carried out following the baseline method [3], which employed only the average reference and band-pass filtering.

6.4.2.2 Data Partitioning

For the AVEC2019 database, the data are partitioned to the training, validation, and test sets by the contest organizer. There is no subject overlap across different partitions. Since the annotations for the test set are not available, we only adopt the training and validation sets. Specifically, the training set is adopted as is. It contains 68 trials from 34 German and 34 Hungarian subjects. For the original validation set, it contains 28 trials from 14 German and 14 Hungarian subjects. We randomly select 7 German and 7 Hungarian subjects as the actual validation set and the remaining 7 German and 7 Hungarian subjects as the test set. Note that the validation and test sets we mentioned for the AVEC2019 database are our newly partitioned ones throughout the paper unless otherwise specified.

[5] `https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf`
[6] `https://mne.tools/stable/index.html`.

For the MAHNOB-HCI database, two data partitioning schemes are used: (i) trial-level random shuffling (TRS, 10-fold) [3] and (ii) leave-one-subject-out (LOSO, 24-fold) [4]. TRS focuses on the trial level and overlooks from which subject the trial comes. It first randomly shuffles the 239 trials and then splits the 239 trials into 129, 86, and 24 trials for training, validation, and test, so that the test set contains 10% of the data and the training and validation sets contain the 60% and 40% of the remaining data. LOSO focuses on the subject level. For the i-th fold, trials from the i-th subject are taken as the test set. All the trials from the remaining 23 subjects are randomly shuffled, with 80% and 20% being the training and validation sets, respectively.

For the AffWild2 database, the data are partitioned to the training, validation, and test sets by the contest organizer, resulting in 341, 71, and 131 trials, respectively. There is no subject overlap across different partitions. The 6-fold cross-validation is employed. By evenly splitting the training set into 5 folds, we have 6 folds in total with a roughly equal trial amount, i.e., $68 \times 4 + 69 + 71$ trials. Note that the 0-th fold is exactly the original data partitioning.

6.4.2.3 Training and Parameter Settings

The training settings for the three databases are slightly different from each other. By default, the batch size is set to 2, and for AffWild2 it is set to 8. The resampling window length and hop length are 300 and 200, respectively. I.e., the dataloader loads consecutive 2×300 feature points to form a minibatch, with a stride of 200. For any trials having feature points smaller than the window length, zero padding is employed. The facial landmarks, mfcc, Vggish, and BERT features are normalized to $mean = 0$ and $std = 1$. The Adam optimizer with a weight decay of 0.001 is employed. The CCC loss written in Eq. 3.4 is used as the loss function. The *ReduceLROnPlateau* scheduler with a patience of 5 and factor of 0.1 is employed based on the validation CCC. The learning rate (LR) and minimal learning rate (MLR) are set to $1e-5$ and $1e-7$, respectively. The maximal epoch number and early stopping counter are set to 100.

The training for the MAHNOB-HCI database sets the resampling window length and hop length to 96 and 24, respectively, as the annotation resolution is at 4 Hz so that the trial lengths, determined by its scarce label points, are greatly smaller

than those from other databases. The visual backbone model is partly updated during the training for the AffWild2 database so that the yielded CNN features are refined to some extent compared to the counterparts for other databases generated by a frozen backbone. Since back then we were attending the ABAW contest, doing so can further boost the prediction performance. Specifically, following the Resnet50 naming convention, three groups, i.e., the output layer, the layer4, and the last three blocks of layer3 of the Resnet50 [190] are manually selected. When $Epoch = 0$, the output layer is unfrozen. When the learning rate decreases to below $1e-7$, the layer4 is unfrozen and the scheduler is reset. The last group is unfrozen when next time the learning rate is below $1e-7$.

The actual training time is largely varied based on different setting on i) the input dimension of the modality, ii) whether a backbone is employed, and iii) the dataset size and partitioning. For example, for the visual modal, we can choose to feed the video frames or extracted CNN features to the networks with or without a backbone. The dimension of video frames and CNN features are and , where the former is about 2 times larger than the latter. Meanwhile, the employment of the the Resnet50 backbone could result in extra 23 million parameters. In practical, using a Resnet50 backbone with the video frame input lead to about 6 times of training time comparing with the counterpart without the backbone. As for the dataset size and partitioning, the AVEC2019 dataset is partitioned by the owner with the smallest size, whereas MAHNOB-HCI is not originally partitioned. The employment of 24-fold LOSO partitioning could result in 23 times more data and training time.

The training time for our proposed models varies greatly depending on several factors such as input dimension, backbone usage, and dataset size and partitioning. For the visual modality, we experimented with feeding either $40 \times 40 \times 3$ dimensional video frames or extracted 512 dimensional CNN features with or without a Resnet50 backbone, resulting in different input dimensions and number of parameters. The use of a Resnet50 backbone with the video frame input led to a six-fold increase in training time compared to the counterpart without the backbone. Additionally, the size and partitioning of the dataset also affect the training time. AVEC2019 is the smallest in size because the dataset is originally partitioned by the owner. Whereas MAHNOB-HCI is the largest because the LOSO partitioning is employed, resulting in 23 times more data and training time.

6.5 Ablation Study

Several questions are crucial to validate our LFAN. First, how to determine the window, hop, and kernel size for the TCN? Second, is the visual modality a better leader for our LFAN? And finally, is the LFAN a better fusion strategy than others? To investigate, the AVEC2019 database with our newly partitioned validation and test sets is used. Specifically, we use the validation set to determine the best modality as the leader, as well as the window, hop, and kernel size. After which, by using the best setting for our LFAN, we compare the test results using different fusion strategies.

Table 6.1: The performance in CCC of our LFAN using different kernel, window, and hop sizes. The batch size is set to 2. The bold fonts denote the best results.

Window size / hop size		60/40	120/80	180/120	240/160	300/200	360/240	420/280	480/320
Kernel size 3	Valence	0.592	0.601	0.608	0.601	0.591	0.602	0.598	0.604
	Arousal	0.567	0.585	0.659	0.622	0.658	0.636	0.639	0.614
	Mean	0.580	0.593	0.633	0.612	0.625	0.619	0.619	0.609
Kernel size 5	Valence	0.596	0.596	0.604	0.599	0.606	0.604	0.607	**0.611**
	Arousal	0.578	0.650	0.678	0.650	**0.688**	0.662	0.670	0.676
	Mean	0.587	0.623	0.641	0.625	**0.647**	0.633	0.639	0.644
Kernel size 7	Valence	0.591	0.593	0.594	0.590	0.590	0.596	0.590	0.591
	Arousal	0.582	0.610	0.612	0.630	0.658	0.653	0.613	0.606
	Mean	0.587	0.602	0.603	0.610	0.624	0.625	0.602	0.599

For the first experiment regarding the parameter settings for the window size, hop size, and kernel size, we employed the grid search, with the result shown in Table 6.1. Note that the receptive field of the TCN is determined by Eq. 3.2. Therefore, the actual receptive fields are 61, 121, and 181 when the kernel size k equals 3, 5, and 7, respectively. We see that, overall, when kernel size is 5, under the same window and hop size, the mean CCCs are always higher than their counterparts. Particularly, when window and hop size are 300 and 200, respectively, the highest mean CCC is observed. The final parameter settings are listed in Table 6.3.

For the second experiment regarding the importance of leader modality, we compare three different LFANs using the deep CNN, the VGGish, and mfcc feature as the leader, respectively, as shown in Table 6.2. It is shown that the deep CNN feature is the most expressive and secures the highest CCC when batch size is 1, 2, or 4. Our LFAN achieves the highest mean CCC when the batch size is 1. The specification of our LFAN is listed in Table 6.3.

Table 6.2: The performance in CCC of our LFAN using different leaders and batch sizes. The bold fonts denote the best results.

Batch size		1	2	4	8
Audio-leader (Vggish)	Valence	0.587	0.584	0.572	0.575
	Arousal	0.599	0.649	0.639	0.625
	Mean	0.593	0.617	0.606	0.600
Audio-leader (mfcc)	Valence	0.584	0.574	0.563	0.574
	Arousal	0.671	0.663	0.641	0.610
	Mean	0.628	0.619	0.602	0.592
Visual-leader (CNN)	Valence	**0.614**	0.606	0.599	0.594
	Arousal	**0.691**	0.688	0.669	0.663
	Mean	**0.653**	0.647	0.631	0.629

Table 6.3: The specification of LFAN. Our LFAN consists of M TCNs for temporal modeling on the M modalities in parallel, and the leader-follower attentive fusion block to fuse the M temporal encodings. I/O: input and output size. Channel: the per-layer kernel number to define a TCN.

TCNs			Leader-follower attentive fusion	
Modality	I/O	Channel	Module	I/O
CNN	512 / 128	(256, 256, 128, 128)	K encoder Q encoder V encoder	From any output size of upstream TCNs to 32
Vggish	128 / 64	(128, 128, 64, 64)		
mfcc	29 / 32	(32, 32, 32, 32)		
Landmark	136 / 64	(128, 128, 64, 64)		
BERT	768 / 128	(256, 256, 128, 128)	Regressor (a linear layer)	128+Mx32 / 1
EEG bandpower	192 / 168	(128, 128)		

For the third experiment, we compare our LFAN against its variant with different fusion strategies. Given the outputs from the M TCNs where each TCN belongs to one modality, the following fusion strategies/blocks are employed. The first one, termed CAT, is straightforward concatenation. The second one termed TCN utilizes an extra TCN for fusion. The fusion TCN is fed by the concatenated outputs of the per-modal TCNs. The third one, termed TLFAN, employs the self-attention on the output of each modal, before feeding to the leader-follower attentive block. It actually makes the LFAN attentive not only in the modal direction but also in the temporal direction. The fourth one, termed as W/o leader, employs only the follower feature, i.e., the co-attention feature $\mathbf{A}$ from Eq. 6.4 without the leader concatenation. The results are shown in Table 6.4.

It is seen that the highest CCC for valence is 0.652, produced by LFAN when the deep CNN, Vggish, and mfcc features are used with a batch size of 1, while the highest CCC for arousal is 0.626, produced by TLFAN using the three features with

Table 6.4: The validation of the leader-follower attentive block in CCC. BS: batch size. C: deep CNN feature. V: Vggish feature, M: mfcc feature, L: facial landmark. The bold fonts indicate the best results.

Batch size 1						
Modality	Emotion	CAT	TCN	TLFAN	W/o leader	LFAN
C+V	Valence	0.641	0.644	0.638	0.645	0.641
	Arousal	0.554	0.541	0.607	0.612	0.625
	Mean	0.598	0.593	0.623	0.629	0.633
C+M	Valence	0.612	0.624	0.627	0.609	0.63
	Arousal	0.503	0.518	0.509	0.487	0.504
	Mean	0.558	0.571	0.568	0.548	0.567
C+V+M	Valence	0.64	0.644	0.648	0.637	**0.652**
	Arousal	0.544	0.533	0.571	0.621	0.625
	Mean	0.592	0.589	0.610	0.629	**0.639**
Unimodal	C: 0.613/0.516/0.565, V: 0.306/0.276/0.291, M: 0.108/0.148/0.128					
Batch size 2						
Modality	Emotion	CAT	TCN	TLFAN	W/o leader	LFAN
C+V	Valence	0.633	0.647	0.644	0.645	0.632
	Arousal	0.549	0.542	0.608	0.61	0.622
	Mean	0.591	0.595	0.626	0.628	0.627
C+M	Valence	0.609	0.611	0.613	0.616	0.609
	Arousal	0.526	0.518	0.509	0.493	0.501
	Mean	0.568	0.565	0.561	0.555	0.555
C+V+M	Valence	0.615	0.641	0.643	0.631	0.638
	Arousal	0.54	0.537	**0.626**	0.609	0.622
	Mean	0.578	0.589	0.635	0.620	0.630
Unimodal	C: 0.606/0.508/0.557, V: 0.306/0.276/0.291, M: 0.108/0.152/0.130					
Batch size 4						
Modality	Emotion	CAT	TCN	TLFAN	W/o leader	LFAN
C+V	Valence	0.636	0.641	0.642	0.624	0.628
	Arousal	0.592	0.587	0.541	0.53	0.623
	Mean	0.614	0.614	0.592	0.577	0.626
C+M	Valence	0.607	0.626	0.602	0.601	0.612
	Arousal	0.5	0.515	0.503	0.482	0.505
	Mean	0.554	0.571	0.553	0.542	0.559
C+V+M	Valence	0.628	0.648	0.639	0.621	0.639
	Arousal	0.573	0.548	0.535	0.54	0.618
	Mean	0.601	0.598	0.587	0.581	0.629
Unimodal	C: 0.613/0.512/0.563, V: 0.297/0.277/0.287, M: 0.099/0.133/0.116					
Batch size 8						
Modality	Emotion	CAT	TCN	TLFAN	W/o leader	LFAN
C+V	Valence	0.633	0.636	0.632	0.599	0.63
	Arousal	0.581	0.575	0.532	0.511	0.574
	Mean	0.607	0.606	0.582	0.555	0.602
C+M	Valence	0.607	0.62	0.603	0.583	0.616
	Arousal	0.497	0.515	0.502	0.497	0.501
	Mean	0.552	0.568	0.553	0.540	0.559
C+V+M	Valence	0.623	0.642	0.617	0.608	0.634
	Arousal	0.521	0.534	0.519	0.506	0.614
	Mean	0.572	0.588	0.568	0.557	0.624
Unimodal	C: 0.602/0.512/0.557, V: 0.300/0.281/0.291, M: 0.100/0.129/0.115					

a batch size of 2. Overall, LFAN achieves the highest mean CCC of 0.639, when the batch size is 1. Comparing the results from modalities C+V against those from C+V+M, we can observe that the concatenation, TCN, TLFAN, and partly of w/o leader (when batch size equals 1 and 2) have a slight drop, which implies that they are unable to complement the three features for a gain. The results from modalities C+M are outperformed by those from the other two combinations suggesting that these two modalities alone are sub-optimal for CER. Comparing the results from TCN fusion against those from our LFAN, we can see that for valence prediction, the TCN fusion can produce comparable results. For example, for modalities C+V, TCN fusion beats our LFAN in all four batch size settings. It is outperformed by LFAN when it goes to arousal prediction or the three modalities scenario. It suggests that our LFAN can maintain a reliable performance on the two emotional dimensions. Comparing the results from TLFAN against our LFAN, we see that the temporal attention cannot bring gains over the LFAN but even degrades it, especially for arousal prediction when batch size equals 4 or 8. Comparing the results from w/o leader against our LFAN, we can see that except for the modalities C+M with batch size 2, there is always a gain in mean CCC for all the scenarios, which proves the importance of emphasizing the leading visual feature.

6.6 Results and Analysis

In this section, we report results on AffWild2 and MAHNOB-HCI databases.

6.6.1 Results on AffWild2

The results come from our attempt at the ABAW3 contest, in which we utilized the CNN, Vggish, and BERT features. As we mentioned before, we carried out the 6-fold cross-validation to make full use of the data and submitted the results from the original partition (fold 0) and three other folds with the best validation CCC, as shown in Table 6.5. The highest CCCs from all the participators are listed for comparison, as shown in Table 6.6.

In TRM [145], the visual and audio modalities, including the facial expression, eGeMAPS [41], ComParE [195], VGGish [42] and the wav2vec2.0 [43] are extracted.

Table 6.5: The CCC results from the 6-fold cross-validation on the validation and test sets. Fold 0 is exactly the original data partitioning provided by ABAW3. Since 5 submissions are allowed, there are no test results on Fold 2 and 3.

Emotion	Partition	Method	Fold 0	**Fold 1**	Fold 2	Fold 3	Fold 4	Fold 5
Valence	Validation	Baseline	0.310	–	–	–	–	–
		Ours	0.450	**0.559**	0.469	0.531	0.539	0.448
	Test	Baseline	0.180	–	–	–	–	–
		Ours	0.490	**0.520**	–	–	0.479	0.511
Arousal	Validation	Baseline	0.170	–	–	–	–	–
		Ours	0.651	**0.671**	0.564	0.562	0.631	0.618
	Test	Baseline	0.170	–	–	–	–	–
		Ours	0.584	**0.602**	–	–	0.580	0.587

Table 6.6: The overall test results in CCC on AffWild2 database. The bold fonts indicate the best results. The highest CCCs from all the participators are listed.

Method	Modality	Valence	Arousal	Mean
TRM [145]	V+A	**0.606**	0.596	**0.601**
Ours	V+A+L	0.520	**0.602**	0.561
Nguyen et al. [142]	V	0.450	0.445	0.448
Savchenko [191]	V	0.417	0.454	0.436
Karas et al. [192]	V+A	0.418	0.407	0.413
JCA [193]	V+A	0.374	0.363	0.369
Zhang et al. [194]	V+A+L	0.300	0.244	0.272
Baseline [158]	V	0.180	0.170	0.175

The features from different modalities are concatenated and fed to a linear layer, after which a transformer encoder is employed to learn the temporal dependencies before prediction using a fully connected layer. In Nguyen et al. [142] [only utilized the facial expression] the extracted deep appearance features are fed to a GRU block [196] and transformer block [52] in parallel. They are then concatenated and sent to another GRU module, followed by the temporal attention operation. In Savchenko [191], the embeddings and scores from the visual backbone are directly used to predict valence and arousal through two fully connected layers. In Karas et al. [192], a 2D CNN and a 1D CNN are used to extract the visual and audio features, after which the features are fused using the cross-modal attention [138], followed by another temporal attention, before feeding to the fully connected layers. In Praveen et al. [193], several backbones such as the I3D [197], R3D [198] and fusion strategies, such as concatenation and cross-attention [199], are explored.

Table 6.7: The comparison results in CCC of our LFAN against unimodal counterpart [4] on MAHNOB-HCI database. The bold fonts indicate the best values. C: deep CNN feature. E: EEG band power. L: facial landmark.

Modality	LOSO		TRS	
	Validation	Test	Validation	Test
C	0.785±0.054	0.679±0.188	0.794±0.061	0.747±0.080
E	0.512±0.107	0.382±0.252	0.527±0.125	0.442±0.150
L	0.613±0.112	0.482±0.277	0.607±0.085	0.535±0.159
C+E	**0.800±0.050**	**0.684±0.190**	**0.814±0.049**	**0.761±0.088**
C+L	0.792±0.048	0.592±0.182	0.806±0.055	0.751±0.087
C+E+L	0.782±0.057	0.683±0.191	0.802±0.053	0.750±0.093

The gap of our LFAN against Situ-RUCAIM3 [145] in the valence dimension might be due to the following facts. First, the backbone of Situ-RUCAIM3 was pretrained using combined 2D facial image databases, including the AffectNet [200], RAF-DB [201], and FER+ [188], whereas we used the latter only. Second, the visual input of Situ-RUCAIM3 is of 112×112 resolution, and the sampling window length is 250. Compared to the input fed to IFAN, which is $40 \times 40 \times 300$ dimensional, the larger image size conveys more details by which the backbone could extract more accurate visual encoding, though it requires over 6.5 times of VRAM and is only practical with multiple GPUs. Nguyen et al. [142] utilized only the visual modality. Savchenko [191] only used visual-spatial embeddings and scores to predict, without considering the temporal dependencies.

6.6.2 Results on MAHNOB-HCI

We compare the results of LFAN against our previous work [4], which employed unimodal input, i.e., the deep CNN features and EEG band power from MAHNOB-HCI for CER. Note that to carry out a fair comparison, the numbers of epochs for training are set to 30 for both methods and trained from scratch. The results are listed in Table 6.7.

We can see that when only one modality is inputted, the CNN feature could achieve the best results in both the LOSO and TRS scenarios. When it comes to two modalities, the CNN combined with EEG features secured the first place, whereas the employment of the landmark feature could deteriorate the performance in LOSO

compared to that of the standalone CNN feature. Adding the landmark as the third modality could yield no improvement over the CNN+EEG instance.

It can also be seen that, generally, results from TRS scenarios outperformed those from the LOSO counterparts. It is understandable since the random shuffling employed by TRS would split the data from the same subject to training, validation, and test sets. Compared to the cross-subject data, its counterpart from the same subject has greater consistency. The model trained in this manner has actually seen the test data to some extent and would inflate the test performance.

6.6.3 Visualization

A qualitative evaluation is carried out to reveal the importance of cross-modal attention. Two representative trials, i.e., *127-30-1280x720* and *video87* from the validation set of the AffWild2 databases, are chosen. The two trials are representative because they both refer to situations where relatively mild facial expressions concurred with strong audio/linguistic cues, i.e., through loud voice, emotional vocal speech, or both, which could be good indicators of high arousal. We take the deep CNN features as the visual cues (V), Vggish features the audio cues (A), and BERT features the linguistic cues (L). They are fed to the trained LFAN for arousal prediction. Meanwhile, the unimodal variant, which utilizes only CNN features is employed for comparison. The intermediate attention matrix $\mathbf{A}$ from Eq. 6.4 is visualized. Note that the dimension of $\mathbf{A}$ is $M \times M \times T$, where T denotes the sequence length of the input and M denotes the number of modalities we utilized. Therefore we have $\mathbf{A} \in \mathbb{R}^{3\times3\times T}$ in this experiment. To visualize $\mathbf{A}$ we reshaped it to $\mathbb{R}^{9\times T}$ so that taking adjacent three rows as a group, the first, second, and third groups represent the extent to which V, A, or L looks to the three modalities, respectively. The visualization, as shown in Fig. 6.2, is limited to several representative windows for clarity.

In *127-30-1280x720*, our LFAN achieved the CCC of 0.506, which outperformed the unimodal counterpart with the CCC of 0.349. In window a and b, the predictions of LFAN are of stronger consistency with the labels. The V modality, i.e., the first three rows of its attention map, has a greater weight towards the L modality, as indicated in row V→L, compared to row V→ V and V→A. Even greater favor can be observed for row A→L and L→L. In short, LFAN downplayed the V modality in

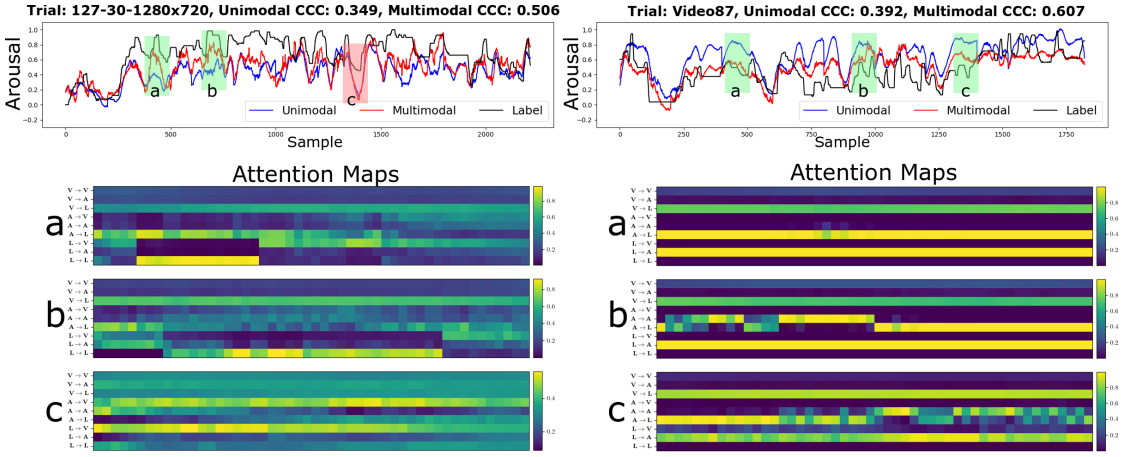

Figure 6.2: The attention maps from LFAN for two representative trials. In this experiment, we fed the deep CNN features (V), the Vggish features (A), and the BERT features (L) to LFAN. And the V modality was fed to the unimodal variant for comparison. The first two line graphs are the visualization of the prediction and labels for the two trials. The attention maps below correspond to the selected windows from the line graphs. It is the visualization of $\mathbf{A}\mathbb{R}^{M\times M\times T}$, where T denotes the sequence length of the input and M denotes the number of modalities we utilized. Since the V, A, and L modalities were utilized, and T was limited to 50 through windowing, we have $\mathbf{A} \in \mathbb{R}^{9\times 50}$ for each attention map. Taken adjacent three rows as a group, the first, second, and third groups represent to what extent does V, A, or L looks to the three modalities, respectively.

window a and b, by which it achieved a greater consistency with the label than that from its unimodal variant. As for Window c, both two methods yielded almost the identity prediction. And its attention map shows that LFAN paid more attention to the V modality. Specifically, the first three rows and row A→V and L→V are all much brighter than the counterpart from the other two windows. Such a misjudgment degraded the performance of LFAN. In *video87*, our LFAN achieved the CCC of 0.607, which outperformed the unimodal counterpart with the CCC of 0.392. From the attention map of the three windows, we can see that the A and L modalities own higher priority from IFAN.

6.7 Discussion and Conclusion

CER aims to map a subject's sequential recordings to the sequential outputs in a real-valued space axed by valence and arousal. Two challenges hinder the development of a reliable CER deep learning model. The first challenge is how to model the temporal dynamics of the emotion cue. Unlike emotion classification,

where a trial is labeled categorically, in CER the emotion cue at a specific time step is the effect of the emotional roller coaster over all the previous steps. Thus, it is crucial to consider the temporal dynamics of the sequential cues in order to predict reliably. The second challenge is how to counter the nontrivial subject bias and preserve the model generality across subjects. The emotion process involves a great number of factors ranging from an individual's physiological status, experience, and mood, to cultural/moral influence, resulting in a large subjective bias and degenerating the representation learning.

We use a large resampling window and multimodal fusion to counter the two challenges. Unlike previous methods that consider only about 100 sampling points and below, we take 300 sampling points as an example and feed them to the TCN for temporal modeling. With the dilated convolutional kernel and stacked residual blocks, the TCN is capable of looking very far into the past to make a prediction. Emotions are dynamic and evolve over time, and capturing these temporal dynamics is important for accurate and robust emotion recognition. By feeding the network with a long sequence, the model can learn how the emotional state changes over time. In addition, the temporal context can help to disambiguate ambiguous emotional states that may be difficult to recognize based on a narrower context. For example, a person may exhibit a neutral facial expression, but the preceding or following emotional states may provide additional information that helps to identify the emotional state more accurately. Also, by learning the temporal dynamics from a broader context, the model can improve its ability to predict future emotional states. This can be particularly useful in applications such as affective computing or human-robot interaction, where it is important to predict the emotional state of a person in real-time and adapt the system's behavior accordingly.

As for the multimodal fusion, It follows the assumption that incorporating multimodal data would yield superior predictions than that from unimodal data, thanks to the complementary nature among different modalities. The visual modality can provide information about facial expressions, body language, and visual cues related to the emotional state. The audio modality can capture tone of voice, intonation, and other auditory cues related to emotions. The linguistic modality can provide information about the content of speech, such as sentiment or emotional content. The physiological modality can capture neural, physiological, or

biochemical signals associated with emotional states. By combining these modalities, the model can capture more comprehensive and nuanced information about the emotional state than any single modality alone. However, the combination of modalities can also be challenging, as different modalities may have different levels of importance or relevance at different time steps. This is where the proposed fusion module comes into the picture.

We proposed the leader-follower attentive fusion block, which combines the learned encodings using the cross-modal co-attention mechanism and yields the cross-modal encoding. To emphasize the dominant visual modality, which we believe in having the strongest correlation with the label, the visual encoding outputted by the visual TCN is taken as the leader and is concatenated to the cross-modal encoding. The attentive module works by assigning weights to each modality based on its relevance to the emotional state at a particular time step. This allows the model to selectively focus on the most informative modality, while downweighting or ignoring less informative modalities. By attending to the most relevant modality, the model can effectively leverage the strengths of each modality and improve the accuracy of the combined model.

By using the AVEC2019 database, we investigated different settings regarding the window, hop, and kernel size. The results suggested that by using a kernel size of 5, a window of 300, and a hop size of 200, our LFAN could produce the highest mean CCC. We showed that the visual modality plays an important role as the leader modality. We also tested the superiority of LFAN over other fusion strategies, including by i) the direct concatenation, ii) an extra TCN, iii) the cross-modal encoding without the visual leader, and iv) the LFAN with temporal self-attention. Overall, our proposed LFAN achieved superior performance against all other variants. Experiments using the MAHNOB-HCI and AffWild2 databases showed that our LFAN could produce promising results when compared to state-of-the-art methods.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The primary objective of this thesis is to enhance continuous emotion recognition (CER) by utilizing multi-modality. Two significant challenges need to be addressed to achieve this objective, which are learning the long-range temporal dynamics of emotion information and preserving the cross-subject generality. These challenges are essential to handle human emotion, which is an ongoing event with temporal continuity and subject variability. The thesis explores three different directions to address these challenges, and their consolidated view is illustrated in Fig. 7.1.

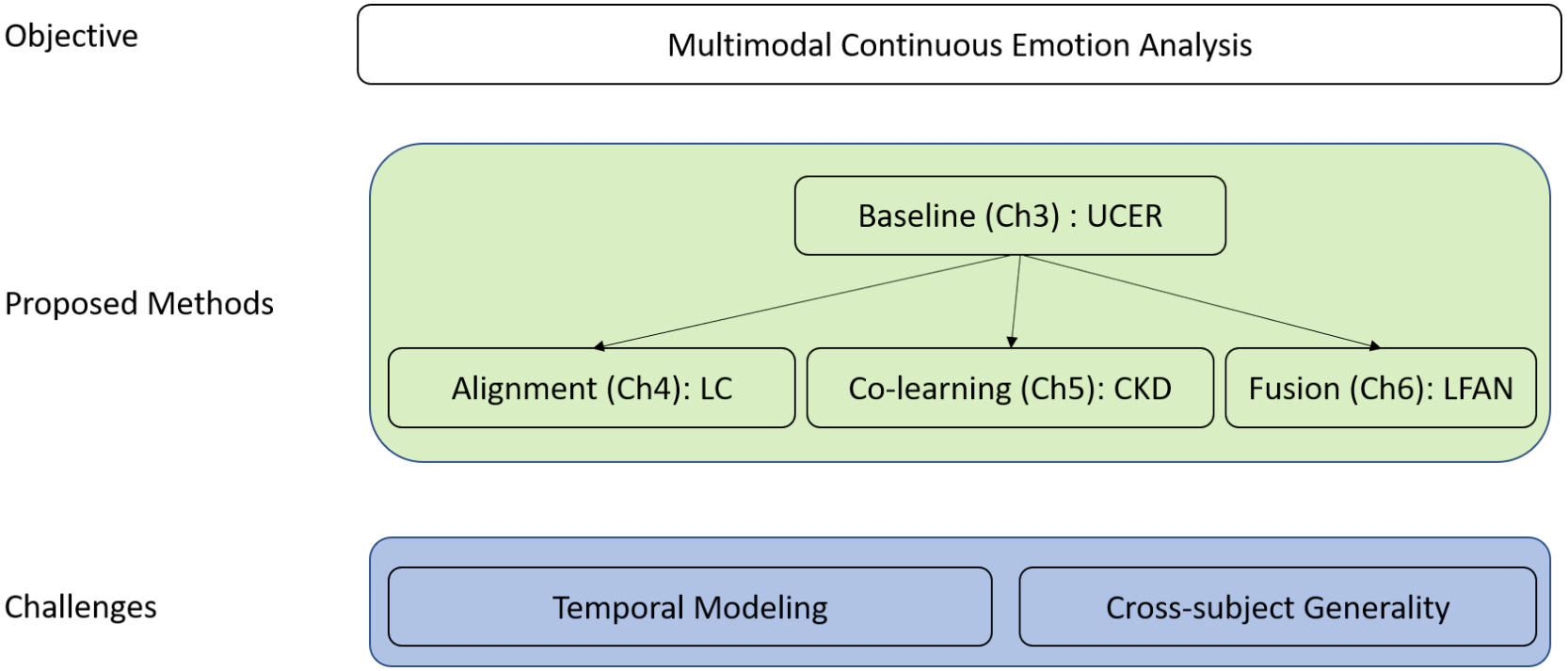

Figure 7.1: The flowchart of the thesis. UCER: unimodal continuous emotion recognition. LC: label correction. CKD: cross-modal knowledge distillation. LFAN: leader-follower attentive network.

In Chapter 3, we proposed a deep neural network for unimodal CER, which employs the temporal convolutional network (TCN) for long-range temporal modeling. Our model features dilated and casual convolutional kernels, along with stacked residual blocks, allowing it to make predictions by looking far into the past. The large resampling window, which contains more time steps (e.g., 96 points for EEG band power data and 300 for video frames), increases global expressiveness and is favored by long-range temporal modeling. This approach also improves generality as longer sequences limit the total batch numbers and updating frequency of the optimizer. Our experimental results demonstrate that the proposed unimodal CER model outperforms baseline methods on RMSE, PCC, and CCC for both TRS and LOSO scenarios. In addition, we visualized the contribution of each brain lobe and EEG band towards the emotion process and found that all four lobes can be active on the six bands, with the β_2 and γ bands (18-30Hz and 30-45Hz) contributing the most to the human emotion process compared to other bands.

Chapter 4 describes our efforts to enhance the deep learning-based emotion classification accuracy by utilizing discretized continuous emotion annotations for model training. To achieve this, we designed a thresholding scheme that converted the continuous real values into three-class categorical labels. As the continuous labels, which were annotated based on the subjects' facial expressions, carried information from the visual modality, this scheme was equivalent to incorporating visual information into the training process. We used facial landmark and EEG average band power as inputs for the visual and EEG modalities, respectively, and compared the emotion classification accuracy with and without the discretized emotion traces. Results showed that applying the thresholding scheme improved the classification accuracy of the EEG modality by approximately 3%, with accuracies of 51.9% and 54.96% before and after the scheme, respectively. However, the visual modality achieved 53.78% accuracy without the scheme, and applying the scheme did not lead to any improvement. Our findings suggest that the gain in EEG modality accuracy may be due to the incorporation of visual information into the artificial categorical labels.

In Chapter 5, we investigated the potential of cross-modal knowledge distillation (CKD) for improving the performance of EEG-based CER by leveraging the visual modality. CKD is a promising approach for addressing the limitations of EEG, such as low information-to-noise ratio and subject bias, by utilizing the complementarity

of visual cues that have a high spatial resolution and cross-subject generality. The teacher and student models were trained on the visual and EEG modalities, respectively, and the temporal embeddings from the trained teacher were used as dark knowledge. To fit the knowledge between the teacher and student, the L1 loss was employed, and the student's prediction was fitted onto the continuous labels using the concordance correlation coefficient (CCC) loss. Experimental results showed that the proposed CKD approach significantly improved the performance of the student model on root mean square error (RMSE), Pearson correlation coefficient (PCC), and CCC compared to the student model without CKD.

Lastly, in Chapter 6, we leveraged multimodal complementarity through intermediate feature fusion. In addition to the temporal modeling discussed in Chapter 3, we addressed cross-subject generality by combining the multimodal emotion cue using our proposed leader-follower attentive network (LFAN). Our LFAN first learns the per-modality long-term dependencies and then combines the learned encodings using the cross-modality co-attention mechanism, with the visual modality as the leader concatenated to the fused encoding. The LFAN achieved runner-up and in the ABAW3 [158] challenge and was expanded with extensive ablation studies and experiment comparisons on the AVEC2019 [76] and MAHNOB-HCI [2, 3] databases. The results showed that our LFAN achieved promising results compared to state-of-the-art methods.

Finally, in Chapter 6, we exploit the multimodal complementarity through intermediate feature fusion. In addition to the temporal modeling we addressed in Chapter 3, the cross-subject generality was dealt with by combining the multimodal emotion cue using the proposed leader-follower attentive network (LFAN). Our LFAN aims to learn the per-modality long-term dependencies first and then combine the learned encodings using the cross-modality co-attention mechanism. The fusion takes the visual modality as the leader, which is concatenated to the fused encoding. The proposed LFAN won the 2nd and 3rd places in the ABAW3 [158] and ABAW5 [202] challenges. It was expanded with abundant ablation studies and experiment comparisons on the AVEC2019 [76] and MAHNOB-HCI [2, 3] databases. The results showed that our LFAN achieved promising results against state-of-the-art methods.

7.2 Limitation and Future Work

The proposed works have the following limitations.

7.2.1 Multitask Learning

The current works overlook the possibility of utilizing shared representations learned from a collection of related tasks for performance improvement. Multitask learning (MTL) reflects the learning process of human beings. A human baby learning to walk does not only acquire the skill of walking itself, but also accumulates general motor skills for balance and intuitive physics. The learned knowledge can benefit him/her anytime when it is required to learn more complex motor tasks. However, in the current works, the networks only carry out task learning one at a time.

In the context of CER, several tasks are available for MTL, considering that the data annotations are provided as basic emotions, facial action units, continuous valence, and arousal traces. It is observed that there is relatedness among different emotion cues. For example, the seminal work [203] found that there is a statistical relationship between categorical (basic and compound) emotions and facial action units. And in [204], the authors utilized such relatedness to address the missing labels on MTL and further boost the model performance on CER. And in [122], the authors generated missing labels based on empirical statistics within several databases and improved the CER performance over the single-task counterpart. The experiments in [205] found that the valence and arousal manifested a V-shaped, asymmetric relation. It means that the more valence a person feels, the higher level of arousal he or she will experience, and vice versa.

In sum, utilizing the MTL on CER can increase the data availability and efficiency, and thus further improve the performance against the single-task counterpart. It is interesting to expand our work with MTL.

7.2.2 Training Calibration

In Section 2.7, we have discussed the differences among RMSE, PCC, and CCC. RMSE suffers from unboundedness and convexity and fails to capture the correlation of the vector being measured. PCC fails to distinguish between linearity and identity, which means that the PCC of two overlapped vectors is the same as that from the same vectors having an offset. CCC overcomes the inferiors and is capable to penalizes deviation from the identity relationship, and has drawn increasing interest in the CER community.

However, we argue that the CCC loss (ρ_c) formulated in Eq. 3.4 may not be the optimal choice to supervise the sequence learning of CER. One reason is the expensive computation. Computing ρ_c necessitates the computations of the standard deviations, covariance, and mean between the sequential prediction and labels, and latter operations such as squaring summing, and the division, let alone the derivative for back-propagation. Another reason is the insufficient point-to-point calibration. Given the sequential predictions and labels, our goal is to fit the former onto the latter. Usually, the deviation between the i-th points from them is varied. Intuitively, the points from the prediction having a greater deviation towards their corresponding labels are "hard" examples, which require stronger supervision. And weaker supervision is needed for "easy" examples. This follows the same idea of the focal loss [206] in the area of object detection.

There are two possible improvements on ρ_c. The first one seeks to simplify it to a form that is computational-friendly. In [207], the authors proposed a set of loss functions that can achieve sequence learning and be the alternatives to CCC loss. Another one seeks to combine multiple loss functions in a random manner, such as the shake-shake regularization [208]. In [209] the authors use a compound of RMSE, PCC, CCC and the sign agreement with the shake-shake regularization. Doing so utilizes the advantages of each loss function and also alleviates the over-fitting. Moreover, we are also interested to design a focal version of ρ_c that emphasizes hard example learning.

7.2.3 Uncertainty

The current works did not take the emotion uncertainty into consideration. The issue of uncertainty dwells in several perspectives, including emotion categorization, emotion quantification, and multimodal fusion.

7.2.3.1 Emotion categorization

The hypothesis for emotion understanding that is widely accepted claims that each basic emotion has uniquely identifiable facial expressions that are universal across human cultures [53], on which the current works were based. However, a growing number of theoretical frameworks hypothesize that people express instances of emotion in situation-specific ways. Multiple types of context, including the external environment and internal status, could influence the facial movement for emotion expression. As a result, there exist intra-category variation and inter-category similarity. The intra-category variation refers to the differences for every person to express the same emotion, while the inter-category similarity refers to the similarities among different emotion expressions. Based on the above context-sensitivity hypothesis, however, it is problematic to label the emotion into a specific category. Yet many facial expression databases (e.g., FER2013 [210], AffectNet [200] and RAF-DB [211]) have done so, which cause the mislabelled annotation, degraded data quality, and deteriorate the recognition credibility.

Methods to account for the uncertainty of this topic include the following. She et al. [212] proposes proposed an auxiliary multi-branch learning framework to discover the label distribution of samples, and an elaborate uncertainty estimation module to reflect the ambiguity extent of paired samples. Zhang et al. [213] proposed relative uncertainty learning. It is based on the relative difficulty and feature mixup of samples. The uncertainty acts as weights to mix different facial features, encouraged by an add-up loss. Wang et al. [214] suppressed the uncertainty using a self-attention mechanism, a ranking regularization, and a relabeling scheme. The self-attention is applied across the mini-batch to weight each training sample so that the uncertain examples are assigned low weights. The ranking regularization splits the weights into two groups and forces a margin between their average. The relabeling module would modify the label if the maximum predicted probability of the sample exceeds the threshold. Some works also discard the deterministic

nature of deep neural network learning and resort to a random process, including Gaussian process [215], affective process [135, 216], and neural process [217, 218].

7.2.3.2 Emotion quantification

The dimensional emotion annotation process depends heavily on human annotators. It is carried out as follows. The annotation tool such as the FeelTrace [219] is used, which allows annotators to watch the subjects' recordings move their cursor or joystick within the 2D valence-arousal dimensional space to quantify the emotional state of the subject. Usually, four to six annotators are employed to conduct the annotation, and the "final" annotation is determined by a certain type of averaging method, such as the CCC-centering [74–76], estimator weighted evaluator (EWE) [220], dynamic time warping barycentre averaging (DBA) [221], generic-canonical time warping (GCTW) [222], and rater aligned annotation weighting (RAAW) [141]. The fused annotation is coined as gold-standard. It is challenging to achieve inter-annotator agreement in dimensional emotion analysis [223] due to varied reaction delays [86] and personal bias among the human annotators. Methods dealt with reaction delays [89], personal bias [109], and inter-annotator agreement [141] are reported.

7.2.3.3 Multimodal Fusion

It is generally expected and accepted that incorporating multimodal information should produce improvement in CER over the unimodal counterpart. However, it may not always be true. Though impressive improvement is seen, negligible or null improvements [224, 225] and even negative effects [226, 227] are also reported. Though we have employed cross-modal co-attention to enhance the feature fusion, there is still considerable ground to be covered to counter such downsides.

Evaluating modality-wise uncertainty/confidence and weighting each modality accordingly has been a classic machine learning topic before the advent of the deep learning era. It is intuitive to pay more attention to the confident modality while downplaying the uncertain ones. The kernel entropy component analysis [228], multi modal-hidden Markov models [229] are two of the conventional methods in

this topic. As for the deep learning-based methods, Subedar et al. [230] utilize Bayesian networks for uncertainty-aware audiovisual fusion. Tian et al. [231] fused the softmax scores from CNNs to avoid the degradation jeopardizing the multimodal fusion. Tellamekala et al. [144] proposed the calibrated and ordinal latent distribution fusion. It learns the latent distribution of the per-modality encodings and then constrains their variance so that the encodings can better represent the information for recognition. The modality-wise uncertainty scores are capable of indicating the difference between the sequential ground truth and the predictions.

List of Author's Publications

Awards

- **2nd place**, "Valence-Arousal Estimation Challenge," *The 3rd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW3).*
- **3rd place**, "Valence-Arousal Estimation Challenge," *The 5rd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW5).*

Journal Articles

- **Su Zhang**, Chuangao Tang and Cuntai Guan, "Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition," *Pattern Recognition* (2022): 108833.
- **Su Zhang**, Wanjing Zhao, Xuying Hao, Yang Yang, and Cuntai Guan, "A context-aware locality measure for inlier pool enrichment in stepwise image registration," *IEEE Transactions on Image Processing* (2019): 4281-4295.
- Yi Ding, Robinson Neethu, **Su Zhang**, Qiuhao Zeng, and Cuntai Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Transactions on Affective Computing* (2022): 1-1.
- **Su Zhang**, Yi Ding, Chuangao Tang and Cuntai Guan, "Leader-follower Attentive Network for Continuous Emotion Recognition," Submitted to IEEE Transactions on Affective Computing.

Conference Proceedings

- **Su Zhang**, Yi Ding, Ruyi An and Cuntai Guan, "Continuous Emotion Recognition using Visual-audio-linguistic information: A Technical Report for ABAW3," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Workshop), 2022.*

- **Su Zhang**, Yi Ding, Ziquan Wei and Cuntai Guan, "Continuous emotion recognition with audio-visual leader-follower attentive fusion," in *IEEE/CVF Conference on Computer Vision (Workshop), 2021.*

- **Su Zhang** and Cuntai Guan, "Emotion recognition with refined labels for deep learning," in *42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2020.*

Bibliography

[1] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, 120: 108102, 2021. xix, 55

[2] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011. xx, 24, 26, 27, 48, 56, 62, 63, 66, 72, 80, 87, 98, 116

[3] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2015. xx, 5, 26, 27, 28, 48, 49, 50, 51, 56, 63, 64, 66, 67, 80, 87, 92, 98, 100, 101, 116

[4] Su Zhang, Chuangao Tang, and Cuntai Guan. Visual-to-eeg cross-modal knowledge distillation for continuous emotion recognition. *Pattern Recognition*, page 108833, 2022. xxv, 43, 75, 92, 101, 108

[5] Michel Cabanac. What is emotion? *Behavioural Processes*, 60(2):69–83, 2002. 1

[6] Randolph R Cornelius. Theoretical approaches to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. 1, 14, 16

[7] Elwin Marg. Descartes'error: emotion, reason, and the human brain. *Optometry and Vision Science*, 72(11):847–848, 1995. 2

[8] Leda Cosmides and John Tooby. Evolutionary psychology and the emotions. *Handbook of Emotions*, 2(2):91–115, 2000. 2

[9] John Tooby and Leda Cosmides. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. 2008. 2

[10] Kathleen Ries Merikangas, Jian-ping He, Marcy Burstein, Sonja A Swanson, Shelli Avenevoli, Lihong Cui, Corina Benjet, Katholiki Georgiades, and Joel Swendsen. Lifetime prevalence of mental disorders in us adolescents: results from the national comorbidity survey replication–adolescent supplement (ncs-a). *Journal of the American Academy of Child & Adolescent Psychiatry*, 49 (10):980–989, 2010. 3

[11] Rajendra Kale. The treatment gap. *Epilepsia*, 43:31–33, 2002. 3

[12] M Thirunavukarasu. Closing the treatment gap. *Indian Journal of Psychiatry*, 53(3):199, 2011. 3

[13] David G Myers. Psychology, 2004. 4, 57, 79, 90

[14] Tom Hollenstein. This time, it's real: Affective flexibility, time scales, feedback loops, and the regulation of emotion. *Emotion Review*, 7(4):308–315, 2015. 4, 5

[15] Kristína Czekóová, Daniel J Shaw, Eva Janoušová, and Tomáš Urbánek. It's all in the past: temporal-context effects modulate subjective evaluations of emotional visual stimuli, regardless of presentation sequence. *Frontiers in Psychology*, 6:367, 2015. 4

[16] James J Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998. 4

[17] Sander L Koole and Lotte Veenstra. Does emotion regulation occur only inside people's heads? toward a situated cognition analysis of emotion-regulatory dynamics. *Psychological Inquiry*, 26(1):61–68, 2015. 4

[18] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980. 4, 15, 45

[19] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020. 5, 61

[20] Jean-Luc Schwartz, Frédéric Berthommier, and Christophe Savariaux. Audio-visual scene analysis: evidence for a" very-early" integration process in audio-visual speech perception. In *Seventh International Conference on Spoken Language Processing*, 2002. 6, 92

[21] Kenneth S Saladin and Leslie Miller. *Anatomy & physiology*. WCB/McGraw-Hill New York, 1998. 10

[22] John E Hall et al. Guyton and hall textbook of medical physiology. *Philadelphia, PA: Saunders Elsevier*, 107:1146, 2011. 11

[23] MW Eysenck and MT Keane. Cognitive psychology; a student's handbook. hove, east sussex, 2010. 11

[24] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

[25] Michael W Eysenck and Mark T Keane. *Cognitive psychology: A student's handbook*. Psychology press, 2015. 11

[26] NG Müller and RT Knight. The functional neuroanatomy of working memory: contributions of human brain lesion studies. *Neuroscience*, 139(1):51–58, 2006. 12

[27] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier, 2013. 14

[28] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003. 14, 21

[29] Charles Darwin. The expression of the emotions in man and animals. In *The Expression of the Emotions in Man and Animals*. University of Chicago press, 2015. 14

[30] Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962. 14

[31] Silvan Tomkins. *Affect imagery consciousness: Volume II: The negative affects*. Springer publishing company, 1963. 14

[32] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier, 1980. 14

[33] Nico H Frijda et al. *The emotions*. Cambridge University Press, 1986. 14

[34] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17 (3):715–734, 2005. 15

[35] Chen Yu, Paul M Aoki, and Allison Woodruff. Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027*, 2004. 15

[36] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001. 15

[37] James R Averill. A constructivist view of emotion. In *Theories of Emotion*, pages 305–339. Elsevier, 1980. 16

[38] W Douglas Frost and James R Averill. Differences between men and women in the everyday experience of anger. In *Anger and Aggression*, pages 281–316. Springer, 1982. 16

[39] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 16

[40] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16 (6):582–589, 2001. 18

[41] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015. 18, 106

[42] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 18, 106

[43] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 18, 106

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 18

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 19

[46] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic dependency-based n-grams as classification features. In *Mexican International Conference on Artificial Intelligence*, pages 1–11. Springer, 2012. 19

[47] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972. 19

[48] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 19

[49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013. 19

[50] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 19

[51] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 19

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 19, 92, 107

[53] Tuan Le Mau, Katie Hoemann, Sam H Lyons, Jennifer Fugate, Emery N Brown, Maria Gendron, and Lisa Feldman Barrett. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature Communications*, 12(1):1–13, 2021. 21, 119

[54] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 22

[55] Christian Keysers, Evelyne Kohler, M Alessandra Umiltà, Luca Nanetti, Leonardo Fogassi, and Vittorio Gallese. Audiovisual mirror neurons and action recognition. *Experimental Brain Research*, 153(4):628–636, 2003. 22

[56] Ken W Grant and Philip-Franz Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108(3):1197–1208, 2000. 23

[57] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004. 23

[58] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. 23

[59] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. 24, 63

[60] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011. 24, 62

[61] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003. 24

[62] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 488–500. Springer, 2007. 25, 26

[63] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011. 26, 30, 31, 40

[64] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013. 26, 27, 31, 40

[65] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, 2019. 26, 28

[66] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 26, 29, 32, 98

[67] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *arXiv preprint arXiv:2101.06053*, 2021. 26, 29, 33, 52

[68] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011. 30

[69] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 449–456, 2012. 31

[70] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2013. 31

[71] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2014. 31

[72] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1335–1336, 2015. 31

[73] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2016. 31, 52

[74] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017. 31, 120

[75] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13, 2018. 31

[76] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019. 32, 52, 97, 116, 120

[77] Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. State of mind: Classification through self-reported affect and word use in speech. 2018. 32

[78] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W Schuller, Iulia Lefter, et al. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 35–44, 2020. 33, 52

[79] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993. 33

[80] Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality*, 37(1):48–75, 2003. 34

[81] Bjorn Schuller. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective Computing*, 2(4):192–205, 2011. 36, 40

[82] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011. 36, 40

[83] Maxim Sidorov and Wolfgang Minker. Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 81–86, 2014. 36

[84] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40, 2014. 36, 41

[85] Albert C Cruz, Bir Bhanu, and Ninad S Thakoor. Vision and attention theory based sampling for continuous facial emotion recognition. *IEEE Transactions on Affective Computing*, 5(4):418–431, 2014. 36, 38

[86] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1299–1311, 2014. 36, 41, 120

[87] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80, 2015. 36, 41

[88] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72, 2015. 36, 41

[89] Soroosh Mariooryad and Carlos Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, 2014. 36, 41, 120

[90] Rui Xia and Yang Liu. A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1):3–14, 2015. 36, 40

[91] Arman Savran, Houwei Cao, Ani Nenkova, and Ragini Verma. Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities. *IEEE Transactions on Cybernetics*, 45(9):1927–1941, 2014. 36, 41

[92] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 97–104, 2016. 36

[93] Shizhe Chen and Qin Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 571–575, 2016. 36, 41

[94] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1748–1761, 2015. 36, 38

[95] Ankit Goyal, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE, 2016. 36, 41

[96] Hongying Meng, Nadia Bianchi-Berthouze, Yangdong Deng, Jinkuang Cheng, and John P Cosmas. Time-delay neural network for continuous emotional dimension prediction from facial expression sequences. *IEEE Transactions on Cybernetics*, 46(4):916–929, 2015. 36, 38

[97] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 890–897, 2017. 36, 41

[98] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Björn Schuller. Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*, 65:76–86, 2017. 36, 41, 92

[99] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5005–5009. IEEE, 2017. 36, 40

[100] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26, 2017. 36, 41

[101] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017. 36, 41

[102] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2017. 36

[103] Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–25, 2017. 36, 39

[104] Arianna Mencattini, Eugenio Martinelli, Fabien Ringeval, Björn Schuller, and Corrado Di Natale. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing*, 8(3):314–327, 2016. 36, 40

[105] Zhaocheng Huang and Julien Epps. A pllr and multi-stage staircase regression framework for speech-based emotion prediction. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5145–5149. IEEE, 2017. 36, 40

[106] Zixing Zhang, Jing Han, Eduardo Coutinho, and Björn Schuller. Dynamic difficulty awareness training for continuous emotion prediction. *IEEE Transactions on Multimedia*, 21(5):1289–1301, 2018. 36, 41

[107] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 65–72, 2018. 36, 41

[108] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 57–64, 2018. 36, 38, 92

[109] Brandon M Booth, Karel Mundnich, and Shrikanth S Narayanan. A novel method for human bias correction of continuous-time annotations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3091–3095. IEEE, 2018. 36, 120

[110] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018. 36, 40

[111] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, and Jiangyan Yi. End-to-end continuous emotion recognition from video using 3d convlstm networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6837–6841. IEEE, 2018. 36

[112] Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Compensation techniques for speaker variability in continuous emotion prediction. *IEEE Transactions on Affective Computing*, 12(2):439–452, 2018. 37, 39

[113] Suowei Wu, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. In *2019 International Conference on Multimodal Interaction*, pages 40–48, 2019. 37, 38

[114] Anna Mitenkova, Jean Kossaifi, Yannis Panagakis, and Maja Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 37, 38

[115] Jie Guo, Bin Song, Peng Zhang, Mengdi Ma, Wenwen Luo, et al. Affective video content analysis based on multimodal data fusion in heterogeneous networks. *Information Fusion*, 51:224–232, 2019. 37, 41

[116] Jinming Zhao, Ruichen Li, Jingjun Liang, Shizhe Chen, and Qin Jin. Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 37–45, 2019. 37, 41

[117] Haifeng Chen, Yifan Deng, Shiwen Cheng, Yixuan Wang, Dongmei Jiang, and Hichem Sahli. Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 19–26, 2019. 37, 41, 80, 92, 93

[118] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition. *IEEE Transactions on Multimedia*, 21(12):3150–3163, 2019. 37, 40

[119] Meshia Cédric Oveneke, Yong Zhao, Ercheng Pei, Abel Dìaz Berenguer, Dongmei Jiang, and Hichem Sahli. Leveraging the deep learning paradigm for continuous affect estimation from facial expressions. *IEEE Transactions on Affective Computing*, 2019. 37, 39

[120] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6060–6069, 2020. 37, 39

[121] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 37, 42

[122] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 37, 39, 117

[123] Bagus Tris Atmaja and Masato Akagi. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *ICASSP 2020-2020IEEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4482–4486. IEEE, 2020. 37

[124] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020IEEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020. 37

[125] Zhaocheng Huang and Julien Epps. An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech. *IEEE Transactions on Affective Computing*, 11(4):653–668, 2018. 37, 40

[126] Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. Multimodal recurrent attention networks for facial expression recognition. *IEEE Transactions on Image Processing*, 29:6977–6991, 2020. 37, 42

[127] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3567–3574, 2021. 37, 42, 89

[128] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3603–3608, 2021. 37, 42

[129] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. Multitask multi-database emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3637–3644, 2021. 37, 39

[130] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 37, 39

[131] Licai Sun, Mingyu Xu, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 15–20. 2021. 37

[132] Salam Hamieh, Vincent Heiries, Hussein Al Osman, and Christelle Godin. Multi-modal fusion for continuous emotion recognition by using auto-encoders. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 21–27. 2021. 37

[133] I Li et al. Technical report for valence-arousal estimation on affwild2 dataset. *arXiv preprint arXiv:2105.01502*, 2021. 37, 40

[134] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021. 37, 42

[135] Enrique Sanchez, Mani Kumar Tellamekala, Michel Valstar, and Georgios Tzimiropoulos. Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9074–9084, 2021. 37, 40, 120

[136] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671, 2021. 37, 39

[137] Soheil Khorram, Melvin G McInnis, and Emily Mower Provost. Jointly aligning and predicting continuous emotion annotations. *IEEE Transactions on Affective Computing*, 12(4):1069–1083, 2019. 37, 39

[138] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020. 37, 42, 107

[139] Ercheng Pei, Yong Zhao, Meshia Cedric Oveneke, Dongmei Jiang, and Hichem Sahli. A bayesian filtering framework for continuous affect recognition from facial images. *IEEE Transactions on Multimedia*, 2022. 37, 39

[140] Ercheng Pei, Meshia Cedric Oveneke, Yong Zhao, Dongmei Jiang, and Hichem Sahli. Monocular 3d facial expression features for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:3540–3550, 2020. 37, 39

[141] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W Schuller. Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 75–82. 2021. 37, 42, 120

[142] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial expression analysis in video. *arXiv preprint arXiv:2203.12891*, 2022. 37, 39, 107, 108

[143] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2376–2381, 2022. 37, 42, 89, 92

[144] Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *arXiv preprint arXiv:2206.05833*, 2022. 37, 42, 121

[145] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Yuanyuan Deng, Ruichen Li, Yannan Wu, Jinming Zhao, et al. Multi-modal emotion estimation for in-the-wild videos. *arXiv preprint arXiv:2203.13032*, 2022. 37, 106, 107, 108

[146] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 37, 42

[147] Wang Kay Ngai, Haoran Xie, Di Zou, and Kee-Lee Chou. Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. *Information Fusion*, 77:107–117, 2022. 37, 42

[148] Gilderlane Ribeiro Alexandre, José Marques Soares, and George André Pereira Thé. Systematic review of 3d facial expression recognition methods. *Pattern Recognition*, 100:107108, 2020. 44

[149] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 44

[150] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 44, 50, 98

[151] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 44

[152] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 45

[153] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018. 45, 54, 59, 78

[154] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 46, 91, 94

[155] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 50

[156] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 52

[157] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[158] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 52, 107, 116

[159] Diane M Beck. The appeal of the brain in the popular press. *Perspectives on Psychological Science*, 5(6):762–766, 2010. 57

[160] Arthur P Shimamura. Bridging psychological and biological science: The good, bad, and ugly. *Perspectives on Psychological Science*, 5(6):772–775, 2010.

[161] William R Uttal. *The new phrenology: The limits of localizing cognitive processes in the brain*. The MIT press, 2001. 57

[162] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 67

[163] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 67

[164] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 67

[165] Kaoru Amano, Naokazu Goda, Shin'ya Nishida, Yoshimichi Ejima, Tsunehiro Takeda, and Yoshio Ohtani. Estimation of the timing of human visual perception from magnetoencephalography. *Journal of Neuroscience*, 26(15): 3981–3991, 2006. 71

[166] Richard Jiang, Anthony TS Ho, Ismahane Cheheb, Noor Al-Maadeed, Somaya Al-Maadeed, and Ahmed Bouridane. Emotion recognition from scrambled facial images via many graph embedding. *Pattern Recognition*, 67:245–251, 2017. 76

[167] Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, and Manmohan Chandraker. Cross-domain similarity learning for face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15292–15301, 2021. 76

[168] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 76

[169] Yixin Wang, Shuang Qiu, Xuelin Ma, and Huiguang He. A prototype-based spd matrix network for domain adaptation eeg emotion recognition. *Pattern Recognition*, 110:107626, 2021. 76

[170] Soraia M Alarcao and Manuel J Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017. 76

[171] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021. 77, 78, 92

[172] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020IEEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020. 78

[173] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8427–8436, 2018. 78

[174] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5032–5039. IEEE, 2016. 78

[175] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2827–2836, 2016. 78

[176] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 79

[177] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE, 2019. 79

[178] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 79

[179] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. 79

[180] Siddharth Roheda, Benjamin S Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2926–2930. IEEE, 2018. 79

[181] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 79

[182] Nishant Sankaran, Deen Dayal Mohan, Nagashri N Lakshminarayana, Srirangaraj Setlur, and Venu Govindaraju. Domain adaptive representation learning for facial action unit recognition. *Pattern Recognition*, 102:107127, 2020. 79

[183] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv e-prints*, pages arXiv–1412, 2014. 81

[184] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 292–301, 2018. 92

[185] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*, 2019. 92

[186] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 93

[187] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*, 2019. 93

[188] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced

label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016. 99, 108

[189] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 27–34, 2020. 100

[190] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 102

[191] Andrey V Savchenko. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv preprint arXiv:2203.13436*, 2022. 107, 108

[192] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. *arXiv preprint arXiv:2203.13285*, 2022. 107

[193] Gnana Praveen Rajasekar, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro Koerich, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. *arXiv preprint arXiv:2203.14779*, 2022. 107

[194] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 107

[195] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, pages 2001–2005, 2016. 106

[196] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 107

[197] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 107

[198] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action

recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 107

[199] R Gnana Praveen, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 107

[200] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 108, 119

[201] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017. 108

[202] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 116

[203] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 117

[204] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 117

[205] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4):917, 2013. 117

[206] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 118

[207] Vedhas Pandit and Björn Schuller. The many-to-many mapping between the concordance correlation coefficient and the mean square error. *arXiv preprint arXiv:1902.05180*, 2019. 118

[208] Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017. 118

[209] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021. 118

[210] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013. 119

[211] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 119

[212] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021. 119

[213] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021. 119

[214] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 119

[215] Ting Dang, Brian Stasak, Zhaocheng Huang, Sadari Jayawardena, Mia Atcheson, Munawar Hayat, Phu Le, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 27–35, 2017. 120

[216] T Mani Kumar, Enrique Sanchez, Georgios Tzimiropoulos, Timo Giesbrecht, and Michel Valstar. Stochastic process regression for cross-cultural speech emotion recognition. *Proc. Interspeech 2021*, pages 3390–3394, 2021. 120

[217] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018. 120

[218] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018. 120

[219] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. 120

[220] Björn W Schuller. *Intelligent audio analysis*, volume 3. Springer, 2013. 120

[221] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011. 120

[222] Feng Zhou and Fernando De la Torre. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, 2015. 120

[223] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1 (1):68–99, 2010. 120

[224] Simina Emerich, Eugen Lupu, and Anca Apatean. Emotions recognition by speechand facial expressions analysis. In *2009 17th European Signal Processing Conference*, pages 1617–1621. IEEE, 2009. 120

[225] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012. 120

[226] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011. 120

[227] Zahra Khalili and Mohammad Hassan Moradi. Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of eeg. In *2009 International Joint Conference on Neural Networks*, pages 1571–1575. IEEE, 2009. 120

[228] Zhihong Zeng, Jilin Tu, Ming Liu, and Thomas S Huang. Multi-stream confidence analysis for audio-visual affect recognition. In *International Conference on Affective Computing and Intelligent Interaction*, pages 964–971. Springer, 2005. 120

[229] Zhibing Xie and Ling Guan. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013. 120

[230] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2019. 121

[231] Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated

input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE, 2020. 121