Time-Continuous Dimensional Affect Estimation from Limited Labelled Data

Ravikiran Parameshwara

A Thesis Submitted for the Degree of Doctor of Philosophy of the University of Canberra

July 11, 2024

Faculty of Science and Technology

University of Canberra
Bruce, Australian Capital Territory

Abstract

Affect, the subjective experience of emotion, feeling, or mood, plays a fundamental role in human cognition, social interactions, and overall well-being. One of the approaches conceptualises emotions into discrete categories, such as *happiness, sadness, anger, fear, disgust, surprise*, and *contempt*, while the other approach represents emotions on continuous dimensions, namely *valence* (degree of pleasantness or unpleasantness) and *arousal* (degree of excitation or calmness). The categorical models oversimplify the complex nature of emotions, failing to capture the variations within each category. Further, emotions are often experienced on a spectrum, and categorical models may not adequately represent this range. Since the dimensional models view emotions as existing on a continuum rather than discrete categories, they provide greater flexibility in capturing the complexity of emotions.

The integration of various psychological and neuroscientific perspectives, coupled with advancements in machine learning and deep learning approaches have enriched the understanding of emotions. Automatic emotion inference aims at developing computational methods employing affective data capturing facial expressions, speech signals, and physiological responses. Since emotions evolve dynamically over time, a *time-continuous* modelling allows to capture the fluctuations, trajectories, and transitions of emotions, providing a more accurate representation.

This thesis focuses on estimating time-continuous dimensional human affect computationally. Specifically, the aim is to infer emotions from facial images/videos by employing computer vision algorithms. However, these algorithms require massive amounts of data for training. Collecting affective data is a serious challenge, due to the subjective and dynamic nature of emotions, making it difficult to obtain consistent and reliable self-reported emotional responses. Additionally, privacy concerns, and the need to capture time-continuous emotional states further complicates the process of collecting accurate affective data. Considering these challenges, a preliminary study is performed to examine the influence of limited labelled data on affect inference. The results reveal that the learnt facial features corresponding to valence and arousal

are not generalisable across subjects. Therefore, towards building a generalised affect inference model, a robust method employing multi-task contrastive learning is proposed. This framework aims to capture affect (dis)similarity, valence and arousal differentials between a pair of facial images are captured for learning effective affect representations. Further, an integration of the Action Units and facial landmarks is proposed for obtaining a focused input where affect is prominent.

While the collection of time-continuous affective data is resource-intensive, affect annotation poses a further challenge. Affect annotation is a time-consuming, costly, and copious process, as it requires skilled annotators to carefully examine each sample. Emotions can be perceived differently by individuals, making it difficult to achieve consistent annotations, resulting in a low consensus among annotators. Hence, a few-shot learning-based approach is proposed for dynamic valence and arousal labelling. Few-shot learning reduces the annotation burden, and increases adaptability to new target domains. The experimental results demonstrate that using the proposed approach, efficient labelling can be performed with few labelled samples (size $< 6\%$ of the dataset).

Additionally, performance of the proposed few-shot learning-based approach is further enhanced by incorporating a non-local neural network, which captures the temporal variations of affect in a video. A cross-dataset evaluation demonstrate that the adopted affect inference methodology is transferable and generalisable. Enhanced human-computer interaction through reduced annotation cost and time can be viewed as a broader impact of this thesis.

to Appa, Amma & Souji...

Acknowledgements

As I conclude this incredible journey of pursuing my Ph.D., I find myself reflecting on the numerous individuals who have played pivotal roles in shaping this academic adventure. This thesis is a product of their unwavering support and encouragement at various stages of my life. Befittingly, this piece in my thesis is dedicated to express my gratitude towards them.

First and foremost, my deepest gratitude towards my Ph.D. supervisors at the University of Canberra (UC) Dr. Ibrahim Radwan, Prof. Roland Goecke and Dr. Ramanathan Subramanian. I have learnt the true meaning of being an academic by witnessing them at work. The weekly meetings were joyous times to ponder, discuss, and debate on the research questions. Their wealth of knowledge and insightful guidance have been invaluable, shaping not only my research but also perspectives about life in general. In addition to their sage mentorship, I am grateful for their unconditional support, which allowed me to attend conferences, explore other labs, and engage in collaborative efforts. These opportunities have broadened my academic horizons and enriched my overall research experience. Ibrahim, thank you for your unconditional support and encouragement when the going was tough. Roland, thanks for trusting me on the project, and showing me what dedication truly means. Ram, thanks for believing in me and introducing me to UC, and being an exceptional advocate of student's well-being. Together, you all epitomise scholarly life, and I am fortunate to have witnessed and learnt from the best. Your generosity and support will be forever etched in my journey towards scholarly excellence.

Beyond my supervisory panel, I extend my sincere appreciation to my collaborators Akshay Asthana (Seeing Machines, Canberra), Iman Abbasnejad (Seeing Machines, Canberra), and M. Murugappan (KCST, Kuwait) for enhancing the depth and scope of my doctoral research. Thanks to the Australian Research Council through Discovery Project scholarship for the generous financial support. My gratitude towards Intersect Australia, Sydney, for trusting me to deliver technical workshops on Python and Machine Learning. Thank you, Kyle, Anastasios and Aidan!

The early seeds of curiosity about Science and Mathematics were during my Bachelor's degree at St. Joseph's College, Bengaluru, India. The love for the subjects wouldn't have been possible without the company of my friends Thashwin, Abhishek and Chandrakiran. I extend my heartfelt thanks to them. The road towards Ph.D. was first paved by them as we attended every Science talk and workshop at Bengaluru. During one such event, I had the privilege of meeting Prof. C.S. Aravinda (TIFR-CAM, Bengaluru), whose guidance has been a continuous source of inspiration to me. I extend my gratitude towards him for providing a reference letter that has been invaluable in my academic pursuit.

I am indebted to my Master's lecturers at Christ University (CU), Bengaluru, for steering me towards a career in research. Special thanks to Dr. Fr. Joseph Varghese, Prof. S. Pranesh, Dr. Mayamma Joseph, and Dr. Smitha Nagouda for providing intricacies of research, instilling in me the discipline of perusing research papers, shaping my mindset, and offering a glimpse into the lifestyle associated with this pursuit. Their guidance has been invaluable in preparing me for the world of research. I would like to thank Venkatesh, Sowmya, Libin, and Johnson, for making the time spent at CU more memorable.

My tenure at IIT Ropar, India, has been instrumental in imparting both the positive and negative aspects of research. I express my gratitude for the unwavering support of my friends at IIT Ropar–Amit Kumar, Simran Setia, Shubhada Aute, Neeru Dubey, Shreya Ghosh, and Varsha Bhat. Their camaraderie has been an invaluable source of encouragement, fostering a collaborative and enriching academic environment. I am grateful for the shared moments of learning and growth that we have experienced together.

The period dedicated to my doctoral studies at the University of Canberra stands out as the most memorable chapter of my academic journey. Thanks to all my lab mates at the Human-Centred Technology Research Centre, specially, James Ireland, David Hinwood, and Ghazal Bhargshady for the thought-provoking discussions ranging from neural networks to academic perks! This thesis would be incomplete without extending my special gratitude to Jason Weber for his invaluable technical assistance, on and off campus. Beyond the academic sphere, Australia has been very kind, making life very smooth. Thanks to Mangala, Manjunath, Bhavana, and Bharath Yalandur for graciously opening their home to us. Special thanks to Vani, Lohith,

and my little buddies Sanath and Mahi, for making Canberra a home away from home.

In the course of this journey, a constant element that experienced both my highs and lows was my family. They have weathered my lows with unwavering support and echoed my highs with shared joy. A heartfelt thanks to my parents, Sadhana and H. K. Parameshwara, for instilling in me core values, humility, and dignity. I cannot imagine how they managed to raise their son to be a first-generation master's and Ph.D. student, all while managing a local bakery shop. Any credit attributed to my hard work and patience today is a direct reflection of their sacrifices and unwavering support. Thank you, *Amma* & *Appa* for everything! Special thanks to my sister and brother-in-law–Chaitra and Ganesh–for their constant support and encouragement. Thank you, Chaitra and Bhava; your encouragement has made a significant impact! Thanks to their little princess Parnika for showering joy to our family. I extend special gratitude to my in-laws, Lalitha and Narayana. Atte, Mava, you hold a place in my heart akin to my parents, thank you for your unconditional support, specially during my doctoral studies! Support by Manasa, Shivu, Sumanth and Nagaveni Hegde will never be forgotten. Special gratitude towards my friends Ganesh, Koushik, Yashas, Chaitanya, Prasad, and Shreesha.

One individual who merits an entire volume of acknowledgment is my wife, Soujanya Narayana. Our companionship, which began during our master's studies, has fostered a deep understanding unparalleled by any other. Undertaking our Ph.D. journeys together at UC has proven to be an immense blessing for both of us. Recognising that words may fall short in fully capturing the extent of her support, I express my heartfelt gratitude to my wife for all that she has done. *Souji*, I cannot comprehend how you manage everything perfectly and synchronously, how you cook deliciously and code efficiently, how you support me morally and technically! I am deeply indebted to your encouragement to pursue my doctoral studies. Thanks for laying the foundation stone for this thesis!

Thank you all!

Ravikiran Parameshwara

January 16, 2024

Canberra, ACT

Publications

During the course of this study, the following refereed journal and conference papers were published.

- **Ravikiran Parameshwara**, Ibrahim Radwan, Ramanathan Subramanian, and Roland Goecke. *Examining Subject-Dependent and Subject-Independent Human Affect Inference in Limited Data*. In Proceedings of the IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pages 1–6, 2023.
  DOI: 10.1109/FG57933.2023.10042798.

- **Ravikiran Parameshwara**. *Determining Affect Intensity on a Continuous Range*. In Doctoral Consortium of the IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), 2023.

- **Ravikiran Parameshwara**, Ibrahim Radwan, Akshay Asthana, Iman Abbasnejad, Ramanathan Subramanian, and Roland Goecke. *Efficient Labelling of Affective Video Datasets via Few-Shot & Multi-Task Contrastive Learning.* In Proceedings of the 31st ACM International Conference on Multimedia (MM'23), pages 6161–6170. 2023.
  DOI: 10.1145/3581783.3613784

Related to but not part of the thesis:

- Soujanya Narayana, Ibrahim Radwan, **Ravikiran Parameshwara**, Iman Abbasnejad, Akshay Asthana, Ramanathan Subramanian, and Roland Goecke. *A Weakly Supervised Approach to Emotion-change Prediction and Improved Mood Inference*. In Proceedings of the 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, MA, USA, pages 1–8. 2023.
  DOI: 10.1109/ACII59096.2023.10388146.

- Harshit Malik, Hersh Dhillon, **Ravikiran Parameshwara**, Ramanathan Subramanian, Roland Goecke. *Examining the Influence of Personality and Multimodal Behavior on Hireability Impressions*. In Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics, and Image Processing (ICVGIP'23), 2023.
  DOI: 10.1145/3627631.3627658

- **Ravikiran Parameshwara**, Soujanya Narayana, Murugappan Murugappan, Ramanathan Subramanian, Ibrahim Radwan, Roland Goecke. *Exploring EEG-based Affective Analysis & Detection of Parkinson's Disease*. Submitted to Intelligent Computing (accepted, in press).

Abbreviations

AAM	Active Appearance Model
AC	Affective Computing
AFEW	Acted Facial Expressions in the Wild
AFEW-VA	Acted Facial Expressions in the Wild [with] Valence [and] Arousal
AffectNet	Affect from the InterNet
Aff-Wild	Affect-in-the-Wild
AI	Artificial Intelligence
ANOVA	Analysis of Variance
AU	Action Unit
BLSTM	Bidirectional Long Short-Term Memory
CCC	Concordance Correlation Coefficient
CE	Cross Entropy loss
CK	Cohn-Kanade Database
CK+	Extended Cohn-Kanade Database
CNN	Convolutional Neural Network
CNS	Central Nervous System
DL	Deep Learning
DNN	Deep Neural Network
ECG	Electrocardiogram
EDA	Electrodermal activity
EEG	Electroencephalogram
EMMA	EMotion and Mood Annotations
FACS	Facial Action Coding System
FAE	Facial Affect Estimation

FAN	Face Alignment Network
FC	Fully Connected
FER-Wild	Facial Expression Recognition from the *wild* web
FSL	Few-Shot Learning
GAN	Generative Adversarial Network
GD	Gradient Descent
HCI	Human-Computer Interaction
HUMAINE	HUman-MAchine Interaction Network of Excellence database
HSD	Honestly Significant Difference
ICC	Intra-class Correlation Coefficient
LBP	Local Binary Pattern
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
ML	Machine Learning
MLP	Muti-Layer Perceptron
MSE	Mean Squared Error
MT-CLAR	Multi-Task Contrastive Learning for Affect Representation
MTCNN	Multitask Cascaded Convolutional Neural Networks
MTL	Multi-task learning
NLNN	Non-local Neural Networks
OMG	One-Minute Gradual Emotion dataset
PCC	Pearson Correlation Coefficient
ReLU	Rectified Linear Unit
RECOLA	REmote COLlaborative and Affective intercations
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RVM	Relevance Vector Machine
SAGR	Sign Agreement
SD	Subject-dependent

SEWA-DB	Automatic Sentiment Estimation in the Wild Database
SGD	Stochastic Gradient Descent
SI	Subject-independent
SL	Supervised Learning
SN	Siamese Network
SOTA	State-of-the-art
SVM	Support Vector Machine
SVR	Support Vector machines for Regression
t-SNE	t-distributed Stochastic Neighbor Embedding
ViT	Vision Transformer
2D	Two-dimensional
3D	Three-dimensional
5FCV	5-Fold Cross Validation

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

At the core of human experience lies *affect* — a fundamental aspect of emotions, sentiments and mood that shape our interactions with the world. Based out of psychological, cognitive, and neuroscience principles, affect provides a sophisticated perspective to perceive, process, and respond to information [Posner 05]. The complex interactions of affect forms the bedrock of our subjective experiences, posing a significant challenge and a valuable opportunity for exploration within the field of computing.

In layman's terms, affect refers to a simple feeling, and to be *affected* is to feel something [Hogg 07]. More formally, psychologists define it as a feeling state that is a primary constituent of the human mind [Wundt 02, Titchener 09]. Constituting a fundamental aspect of

human experience, it plays an integral role in many psychological theories and studies. From a neuroscience perspective, affect is associated with the functionalities of different brain regions. Neuroscientists examine how neural processes provoke and lead to affective experiences [Calvo 15]. Deliberating on a biological viewpoint, there is large agreement that affective experiences are created within the brain [Panksepp 11]. Further, affect is manifested through physiological changes in the body, including alterations in heart rate, hormonal levels, facial expressions, and other bodily responses. Research in neuroscience also affirms that cognitive and affective processes share overlapping neural structures [Bechara 00]. The *Amygdala*, the key structure of the human brain responsible for processing emotions, is also involved in cognition, particularly in the way affect influences cognitive processes. Cognitive theories postulate that affect can influence cognitive abilities such as information processing, ideas and memory, and conversely, cognitive processes also contribute to generating affective responses [Forgas 08]. While the research on affect—encompassing psychology, biology, and cognitive neuroscience—is substantial, it is interesting to note that there is a considerable overlap and subtle disparities in their perspectives.

Early research on creating intelligent computers focussed on problem solving, reasoning, learning, and other tasks which are considered important for intelligence. Without being aware that these cognitive functions are influenced by affect, the idea of building affect-aware computers was neglected till late 20^{th} century. Building upon the established principles and theories of psychology, biology, and cognitive neuroscience, a novel dimension of making machines capable of interpreting and synthesising affect was introduced. *Affective Computing* (AC) was initiated with the aim of instilling human-like capabilities of observing, interpreting and generating affect features in computers [Picard 00, Tao 05]. In the affective computing paradigm, recognising, inferring and interpreting affect is formulated as a computational problem. A few among the several topics of investigation involve developing systems that can identify affect from facial expressions, voice, gesture, etc., creating technology capable of synthesising affect (for example, chat bots), modifying responses on interfaces based on the user's affective experience, and developing tools for assisting in mental health diagnostics. Overall, the concept of affect is multifaceted, and affective computing represents a rich and interconnected domain that

spans the subjective aspect of human experience to the objective aspect of technology development.

1.1 Motivation

Although the psychology literature offers a plethora of definitions of emotion[1], it is also agreed that emotions are mental states brought about by neurophysiological changes associated with behavioural responses [Ekman 94, Damasio 98, Panksepp 04]. *Emotional state* refers to the internal dynamics when an emotion is experienced. Emotions involve multiple components, such as subjective experience, cognitive processes, expressive behaviour, and physiological changes [Scherer 05]. For example, while helping a colleague at work, if someone says something derogatory, you may experience an emotional response. Neurophysiological processes prepare you for a new action, your heart pumps faster (a bodily change), you feel anger mounting (a subjective experience), you frown (a facial expression), and you may say something vindictive (a new action).

Emotions play a key role in cognitive processes such as decision making, perception, learning, and memory retention. Emotions are believed to be potent and beneficial drivers of decision making [Lerner 15]. Scientific evidence also shows that too little emotion can impair decision making [Bechara 00]. The impact of emotion on memory retention and the learning process is well established. Studies report that positive emotions contribute to academic achievements and facilitate the learning process [Um 12]. Moreover, negative emotional states impair the learning process and are known to be detrimental to memory [Vogel 16].

While humans can identify and synthesise emotions, computers have to *process* or be *trained* explicitly to grasp and respond to emotional cues. It is argued that for an intelligent Human-Computer Interaction (HCI), there is a need for computers to have a natural interaction with the user, which in turns requires the computer to be able to understand and recognise emotions [Sebe 05]. The mounting evidence of the importance of emotions in human cognitive functions and social interactions, and the need for affective HCI makes automatic affect inference the

[1]The terms *affect* and *emotion* are used synonymously in this thesis.

Figure 1.1: Discrete emotion classes. Images adopted from the Compound Facial Expression Database [Du 14].

cornerstone of affective computing.

1.2 Preliminaries

The earliest scientific study of emotion can be attributed to the philosophical enquiries of the Stoics (3rd century BC) [Graver 02]. Further, the publication of Descartes' treatise remains a noteworthy event in the history of psychology, where he regards emotion as the result of physical conditions [Irons 95]. The pioneering work by Darwin [Darwin 72] has largely been an inspiration for the *basic emotions* that we are conversant today. Darwin's work was followed by the remarkable advancement by William James, in his critique, *What is an Emotion?*, where he posited that the relation between our body and emotions is bidirectional [James 84]. While bodily disturbances are conventionally considered byproducts of emotions, these variations are the source for the emotion itself. This claim was further criticised by Cannon's work [Cannon 27], where he argued that the body cannot cause emotion, as the internal changes are too slow and can occur in both emotional and non-emotional states. These reciprocal arguments set the foundations for what is now conversant as *affect sensing*. The most influential work in emotion research was by Ekman, Sorenson & Friesen [Ekman 69], where they attempted to demonstrate the universality and discreteness of emotions, in addition to finding specific biological correlations of specific emotions.

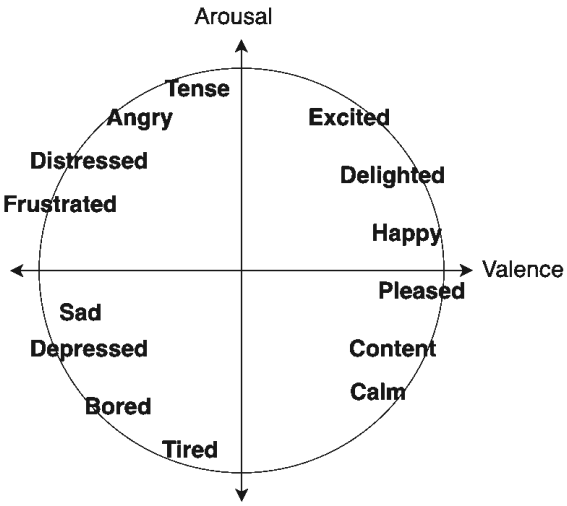


Figure 1.2: Illustration of Russell's circumplex model of emotion [Russell 80].

1.2.1 Categorical and Dimensional Emotions

The numerous attempts by psychologists explaining the origin and function of affect have lead to intense research on this topic. The discrete emotion theory revolved around finding that emotions are discrete, measurable, and physiologically distinct [Darwin 72, Tomkins 62, Ekman 69]. In [Ekman 69], it was stated that certain emotions appeared to be universally recognised, even in preliterate cultures where people had minimum exposure to media, and could not learn from depicted emotions. The *basic* emotions, namely *happiness*, *sadness*, *fear*, *anger*, *surprise*, *disgust*, and *contempt* are described as discrete as they are distinguishable by facial expressions and biological processes [Ekman 69, Ekman 92]. An illustration of facial expressions of basic emotions is shown in Figure 1.1.

While representing emotions as discrete and fundamentally different constructs is one viewpoint, the other doctrine was to characterise emotions on a dimensional basis. The dimensional model of emotion aims at conceptualising human emotions by defining their position on two or three dimensions. The earliest dimensional model proposed that emotions can be described by three dimensions, as *pleasurable-unpleasurable*, *arousing-subduing*, and *strain-relaxation* [Wundt 02]. In another dimensional model, the similarity between two expressions was represented by their closeness in a two-dimensional space with axes *pleasantness* or *unpleasantness*, and a combination of *sleep-tension* and *attention-rejection* [Abelson 62]. While

many such dimensional models were developed, only a few are commonly accepted.

Russell's *Circumplex model* of emotion is a widely acknowledged dimensional model, which suggests that emotions are distributed in a two-dimensional circular space, with *valence* and *arousal* dimensions [Russell 80]. As depicted in Figure 1.2, valence in the horizontal axis, refers to the degree of pleasantness or unpleasantness, and arousal in the vertical axis, represents the degree of excitement or calmness. The centre of the circle represents a neutral valence and an intermediate level of arousal.

1.2.2 Modalities

Modality refers to the channel through which an emotion is expressed or perceived. Each modality offers unique information about the emotional expressions. Research on understanding emotions has been done integrating information from these diverse modalities.

- **Face:** The face serves as one of the fundamental modalities for expression and recognition of emotions, thus playing a key role in interpersonal communications. The human face is rich in emotional cues, conveying a spectrum of emotional expressions ubiquitously seen in diverse cultures and societies [Ekman 71]. The significance of facial expressions in emotion inference is deeply rooted in evolutionary biology, where the ability to interpret facial signals provided a survival advantage for early humans in social interactions [Darwin 72]. The Facial Action Coding System (FACS), a taxonomy developed to catalog facial expressions by assigning numerical codes to various facial muscle movements, is a fundamental tool in studying emotions [Ekman 78]. FACS breaks down distinct facial expressions into configurations called *Action Units* (AUs). Neuroscientific studies highlight the dedicated neural circuits in the brain, such as the *fusiform* face area and the superior temporal *sulcus*, specifically attuned to processing facial information, underscoring the intrinsic importance of the face in emotion perception [Kanwisher 97, Haxby 00].

  In the context of affective computing, leveraging facial cues for emotion inference has become integral to the development of emotion-aware technologies. The face, as a modality, offers real-time and non-intrusive insights into an individual's emotional state, making

it invaluable for applications ranging from HCI to healthcare and social robotics. The widespread adoption of facial expression analysis in affective computing emphasises its effectiveness and applicability in capturing the nuances of human emotions [Tao 05].

- **Voice:** Although much research focuses on investigating the role of the face in emotional expression, there are vocal correlates of emotional expression in *speech* as well. The human voice carries adequate affect information through variations in pitch, tone, rhythm, and intensity. Conversely, emotions have significant repercussions on voice production mechanisms and, ergo, on vocal characteristics [Sundberg 11]. As a consequence, much research has been dedicated to identify the acoustic correlates for different vocal expressions using a variety of standard algorithms namely mel-frequency cepstral coefficients (MFCCs), intensity, fundamental frequency of phonation, etc. [Juslin 05].

- **Gesture and posture:** Gesture is a form of non-verbal communication, which include a movements of hands, face, or other body parts, while posture refers to another non-verbal form of communication, which describes the position or configuration of the body. The Body Action Coding Scheme, including descriptive characteristics of upper body, shoulders, head, and arms, was developed associating distinct postural behaviours with emotions [Wallbott 98]. Body motion and posture are known to highlight and intensify the emotion expressed through face and voice [de Gelder 15].

- **Physiological signals:** *Physiological signals* such as electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response, pupilary dilation, etc. are known to assist in emotion recognition [Picard 00]. For example, in a person experiencing anger, the respiration rate and heart rate may increase. The physiological signals change when people face different situations, and are in response to the Central Nervous System (CNS) of the human body, which influences emotions [Shu 18].

1.2.3 Static vs Time-Continuous

Static modelling of affect involves analysing data at single or specific instances in time. For example, examining the physiological signals at a specific moment to infer affect, or using

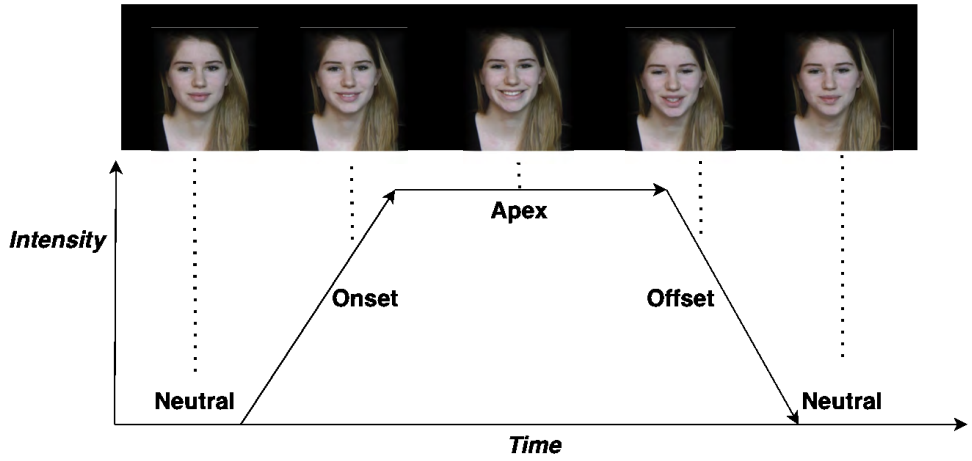


Figure 1.3: Illustration of temporal phases of neutral-onset-apex-offset. Image sequence (video) from the UvA-Nemo Smile Database [Dibeklioglu 12, Dibeklioğlu 15].

images to infer affect through facial expressions. However, affective states are dynamic that unfold and evolve over time as responses to stimuli [Ekman 78, Puccetti 21].

Affective states are dynamic responses that vary around a "baseline" or *homeostatic* set point [Puccetti 22]. Unexpected or aversive stimuli lead to shifts from the homeostatic baseline, which further results in adaption of the human behavioural response. Thus, relative to the human's evolving response, the emotional states continually change in the onset, apex, offset stage, and decaying to the baseline state. For instance, , as illustrated in Figure 1.3, *Neutral* is when the face is relaxed and there are no muscle movements, *onset* is the initial relatively brief phase when the facial muscle activity begins, *apex* is when the facial expression is at maximum intensity, and finally, *offset* is when the expression starts to subside as the muscles begin to relax [Ekman 69].

Emotional dynamics are an outcome of neuronal activity [Panksepp 11], that yield physiological, behavioural, and cognitive processes associated with emotion. Thus, variability of emotional dynamics across individuals is fundamentally attributed to evolutionarily conserved neural circuit motifs [Berridge 19]. While modelling static data provides valuable insights [Li 22], understanding the temporal dynamics of affect is crucial for a comprehensive analysis of affective states.

1.2.4 Affect Estimation Systems

Affective computing encapsulates the idea of expanding the intersection between affect and technology, by developing systems and devices that can recognise, interpret and synthesise affect. It is inspired by the observations and theories of psychology, neuroscience, and cognitive science, and emphasises on the integral role of affect in intelligent human behaviour. With a broad aim of integrating emotional intelligence into machines, this area of research seeks to develop systems capable of understanding human emotions, thus facilitating natural and effective human-computer interactions. Consequently, it encompasses efforts to (a) automatically infer affect from various cues, and (b) synthesise affect and respond relevantly, leading to engaging interactions.

The insights obtained from psychological theories on understanding affect and categorising it from facial expressions, voice, physiological signals, etc., can be used to teach computers to infer human affect from images and videos obtained from cameras, speech waveforms gathered from microphones, etc. For example, with vision input (image or video), gestures, facial and lip movements can be extracted. Likewise, with audio input, linguistic information can be extracted. At this juncture, the recent advances in *Artificial Intelligence (AI)* techniques can be leveraged to design affect estimation frameworks. *Machine Learning (ML)*, a branch of AI, involves computational methods to enable machines to learn patterns from data, and to make predictions or decisions based on that learning. Overcoming the need for explicitly providing a computational model to the machines, ML algorithms enable machines to make data-driven recommendations and decisions.

While ML algorithms provide enormous advantages, a common challenge is the *curse of dimensionality*, where the higher the dimensions used to represent the data, the less effective the ML algorithms become [Bengio 05]. For instance, in images, high dimensionality arises from the large number of pixels, where each pixel contains colour information and its position within the image, and in audio files, high dimensionality emerges from the audio's sampling rate, amplitude, frequency, and time. As the dimensionality increases, the accompanying problems include (a) increased sparsity, where as the dimension increases, the data spreads

out, and the majority of the data points become distant from each other, (b) *overfitting*, where high-dimensional spaces can lead models to capture noise or irrelevant features, reducing their generalisation capability, and (c) the need for more data, where as the number of dimensions increases, more data is needed to maintain a representative sample.

In order to mitigate the problem of the curse of dimensionality, strategies such as feature selection and dimensionality reduction methods are employed. These methods are used to reduce the number of input variables or features in a dataset while retaining as much useful information as possible. ML algorithms with a limited number of layers or depth in their structure, called *shallow* architectures can then be applied for efficient modelling. However, shallow architectures, typically consisting of only a few layers and, hence, fewer parameters, have limited capacity to capture complex patterns or hierarchical representations present in the data. Since affect estimation involves nuanced understanding of emotions or dealing with highly complex emotional cues, shallow architectures might not be able to capture the patterns efficiently [Bengio 07]. This leads to the adoption of *deep* architectures, which enable the models to learn intricate representations hierarchically. The obvious difference between shallow and deep architectures is the *depth*, which refers to the number of subsequent computational layers. While there is no universal rule on the number of layers to differentiate shallow and deep architectures [Goodfellow 16], a cutoff of three or more is generally used [Hinton 06, LeCun 15].

Inspired by the human brain's structure and its way of processing information, deep architectures mimic the brain's ability to learn hierarchical representations of data. Just like neurons in the brain, artificial neurons in the *deep neural networks (DNNs)* receive inputs, process them using weights and activation functions, and produce an output signal. The potential of deep learning algorithms to perform end-to-end learning, discover meaningful and discriminative features, and their capacity of being scalable to handle large datasets has led to their successful adoption for affect estimation [Rouast 19]. Estimating affect from various modalities (refer Section 1.2.2) involves dealing with data that has *spatial* or *temporal* structure. Hence, specialised architectures of DNN, called *Convolutional Neural Network (CNN)* and *Recurrent Neural Network (RNN)* are employed [Kossaifi 20b, Toisoul 21]. The methodology in this thesis is inspired by the literature of DNN for robust affect estimation .

1.3 Challenges

The impending challenges of the field, as described below, have influenced how the research questions are framed and the methodology chosen.

- **Data acquisition:** Data acquisition forms a principal element in the field of affective computing. Subjectivity and diversity influence the variability and richness of the data, thus posing a challenge. Building robust and unbiased models requires datasets that span diverse cultures, age groups, social backgrounds, and demographics, further heightening the complexity. Collecting affective data further raises concerns of privacy and ethical considerations. It is extremely important to protect the identity of individuals sharing affective data by implementing anonymisation techniques.

- **Annotation:** Obtaining accurate labels of affective states in a time-continuous space is a crucial part of the affect estimation system. Additionally, annotation requires experts and is a laborious task, impacting scalability and efficiency. Achieving high inter-annotator agreement is a challenge due to subjective interpretations and the perception of affect, especially when dealing with continuous dimensional data. While many researchers take into account the agreement and correlation measures (for example, [Grimm 08]), others have resorted to self-assessment reports (for example, [Haag 04]). Overall, deriving ground truth labels by modelling the inter-annotator agreement levels remains a challenge in the field. A lack of standard annotation guidelines has also lead to inconsistent and ambiguous annotation practices.

- **Temporal modelling:** A challenge in affect estimation using time-series data is modelling the temporal dynamics. The strict sequence of temporal phases, for example, neutral-onset-apex-offset in facial expressions (refer Section 1.2.3), is visible in posed expressions, but the distinctions in phases is hard to discern in spontaneous expressions, as the subject's behaviour is more natural and unpredictable. While considering the granular differences is crucial for affect modelling, it is challenging to capture the intricate dynamics.

- **Data attributes:** Attributes such as pose, illumination, and occlusion, pose significant challenges. Changes in orientation, position, or posture can distort features crucial for accurate analysis. Different lighting conditions such as low light, high shadows, variation in illumination, alter the appearance of the subject. Partial or full occlusion of subjects obstructs crucial visual information.

- **Ethical concerns:** Safeguarding emotional data to prevent unauthorised access is critical to ensure confidentiality. Ensuring fair and unbiased representation of affect across diverse population rises as a sensitive issue that requires careful management, to avoid stereotypes and discrimination. Further, responsible usage of the data to avoid exploitation for commercial or manipulative purposes is a significant challenge.

- **Government regulations:** While regulations are crucial for ensuring ethical and responsible deployment, the dynamic and evolving nature of regulatory frameworks for affective computing raises concerns about potential constraints on the conductance and practice of affective computing research. For example, the current AI Act proposed by the European Union (EU) [Union 24] has a critical impact on the affective computing research [Iren 23]. The systems that use biometric-based data to infer emotions are considered *high-risk*, and are permitted subject to compliance with regulatory requirements. Striking a balance between fostering innovation in affective computing and addressing societal concerns through effective regulation poses a challenge for researchers in the field.

1.4 Aim

The aim of this research is to propose methodologies focused on temporal modelling of affect from faces, particularly within the constraints of the limited availability of labelled data.

The modality of the face is favoured over voice, body gestures, text, and other channels for dimensional affect estimation due to several compelling reasons. Firstly, the human face conveys a rich array of emotional information through its various expressions, which are both subtle and complex [Ekman 71]. The immediate and direct feedback provided by facial ex-

pressions makes them particularly useful for real-time affect estimation, crucial in interactive settings [Zeng 07]. Additionally, facial expressions of basic emotions are largely *universal* across different cultures, offering a more reliable and consistent method for emotion inference compared to modalities like voice or gesture, which may vary significantly across cultural contexts [Ekman 92]. Analysing facial expressions is also non-intrusive, leveraging cameras and computer vision techniques to monitor emotions without physical contact, unlike physiological signals that require sensors [Cohn 10]. Moreover, facial analysis can be robust in varied environmental conditions where voice and gestures might suffer from background noise or occlusions [Pantic 00]. Finally, advancements in machine learning and computer vision have led to the development of sophisticated tools and algorithms for facial recognition and analysis, enhancing the accuracy of affective estimation [Toisoul 21]. While other modalities offer valuable emotional cues, the face remains a primary and effective modality for affect estimation due to its rich informational content, immediacy, universality, non-intrusiveness, robustness, and the advanced state of facial analysis technologies.

Another focus in the proposed research is the temporal modelling of affect.Temporal modelling is paramount in capturing the dynamic nature of affective states over time, providing a more comprehensive understanding of how emotions evolve and manifest. This research recognises the significance of temporal dynamics, seeking to establish frameworks that improve the accuracy and reliability of affective state recognition systems.

This temporal perspective enriches the computational understanding of affect, offering a more realistic portrayal of how emotions unfold in dynamic real-world scenarios. Moreover, limited labelled data settings are a practical consideration, as in many real-world applications, acquiring large labelled datasets can be challenging and resource-intensive. This research aims to address this practical constraint by developing methodologies that are effective even when training data is limited, thus increasing the feasibility and applicability of affective computing in diverse contexts.

This research is poised to benefit the Affective Computing community and beyond by contributing valuable insights and tools for improved affect recognition in scenarios where labelled data is scarce. The developed methodologies have the potential to enhance the adaptability and

generalisability of affective computing systems, making them more applicable in real-world, data-constrained settings. Beyond the immediate applications in affective computing, the findings of this research may also have broader implications in related fields, such as HCI, healthcare, and emotion-aware technologies.

1.5 Research Questions

Based on the challenges listed and the aim stated above, this thesis aims to propose novel methodologies for facial affect estimation. Specifically, this thesis addresses the following central research question:

How can time-continuous dimensional human affect be automatically estimated using limited labelled video data?

The following incremental research questions guide towards fulfilling the overall aim of this thesis.

1. Do subject-specific idiosyncrasies play a role in time-continuous automatic affect estimation?
2. Can learning affect differences be useful for learning affect representation?
3. How can affect be reliably estimated from limited labelled data?
4. How can the temporal context be used for better affect estimation?
5. Are the proposed methodologies generalisable across various datasets and subjects?

1.6 Contributions

The above research questions are addressed in this thesis as follows:

Contribution 1: Firstly, the thesis explores whether there exists subject-specific idiosyncrasies when modelling the time-continuous affect estimation. While studies have established the existence of *signature* facial expressions corresponding to the basic categorical emotions,

individual differences in emoting facial expressions nevertheless exist; factoring out these idiosyncrasies is critical for effective emotion inference [Ekman 11, Barrett 19]. Chapter 4 explores continuous human affect recognition using AFEW-VA [Kossaifi 17], an 'in-the-wild' video dataset with limited data, employing *subject-independent* (SI) and *subject-dependent* (SD) settings. The SI setting involves the use of training and test sets with mutually exclusive subjects, while training and test samples corresponding to the same subject can occur in the SD setting. A novel, dynamically-weighted loss function is employed with a CNN-Long Short-Term Memory (LSTM) architecture to optimise dynamic affect prediction. The results indicate that the features of valence and arousal learnt by the model are not generalisable across subjects. Visualisations convey that the features of the subject-independent framework are not as discriminative as the subject-dependent setting.

Contribution 2: Next, in order to estimate affect from limited labelled data for subject-dependent and subject-independent settings, learning robust affect representations is explored through a pair of images as input, as detailed in Chapter 5. Multi-Task Contrastive Learning for Affect Representation (MT-CLAR) combines multi-task learning with a Siamese network (SN) trained via contrastive learning to infer from a pair of expressive facial images (a) the (dis)similarity between the facial expressions, and (b) the difference in valence and arousal levels of the two faces. This foundational model (foundational to the research described in this thesis) can be fine-tuned and used for other downstream affect related tasks. Extensive experiments are performed on the choice of design of the architecture. To improve subject-independent affect estimation, a landmark-driven Action Units attention module is introduced. Results show that MT-CLAR is a general purpose affect system, and can be effectively used for affect inference from limited data via few-shot learning framework.

Contribution 3: Leveraging apparent affect differences learnt from MT-CLAR, Chapter 6 introduces a novel few-shot learning framework to estimate affect from a sparse labelled data setting. Given one or a few labelled video frames (termed *support-set*), the framework labels the remainder of the video for valence and arousal. Experiments are performed on the AFEW-VA dataset with multiple support-set configurations; moreover, supervised learning on representations learnt via MT-CLAR are used for valence, arousal and categorical emotion pre-

diction on the AffectNet [Mollahosseini 19] and AFEW-VA datasets. The results show that valence and arousal predictions via MT-CLAR are very comparable to the state-of-the-art (SOTA), and SOTA results are significantly outperformed with a support-set ≈6% the size of the video dataset.

Contribution 4: Investigating the question on whether the temporal modelling enhances the image-based affect estimation system, Chapter 7 details the methodology. The time-continuous affect representation system is based and built on MT-CLAR, called Temporal MT-CLAR. A non-local neural network is employed to encode temporal information. Non-local neural networks capture long-range dependencies in spatial, as well as temporal dimensions.

Contribution 5: In order to validate the reusability and generalisability of different methods proposed for time-continuous affect estimation from limited data, investigations on different datasets are performed in Chapter 8. The foundation model, MT-CLAR, is tested on different datasets with varied pose, subjects, illumination condition, etc. The chapter also details the usage of MT-CLAR for different affect related tasks.

1.7 Thesis Outline

This dissertation comprises a total of nine chapters. The remaining chapters are organised as follows:

Chapter 2: This chapter provides a literature review of dimensional affect estimation systems.

Chapter 3: This chapter details the general methodology used in this research. The chapter discusses the overall workflow of automatic affect estimation systems. A list of datasets considered in this research is also detailed.

Chapter 4: While prior studies have established the existence of *signature* facial expressions corresponding to the basic categorical emotions, individual differences in emoting facial expressions nevertheless exist; factoring out these idiosyncrasies is critical for effective emotion inference. This chapter explores continuous human affect estimation on an 'in-the-wild' video dataset with limited data, employing SI and SD settings. A novel, dynamically-weighted loss

function is introduced to optimise the CNN-LSTM architecture for dynamic affect prediction.

Chapter 5: In order to surmount the challenge of affect estimation from limited video data, this chapter proposes MT-CLAR, a novel deep-learning based affect representation algorithm, establishing a framework wherein the formidable objective of affect estimation from limited video data can be achieved through MT-CLAR. The chapter also discusses variants of MT-CLAR to enhance generalisability.

Chapter 6: This chapter extends the image-based MT-CLAR framework for automated affect labelling for videos, where given one or few labelled video frames, MT-CLAR labels the remainder of the video for valence and arousal. The chapter also details comprehensive experiments with multiple support-set configurations.

Chapter 7: This chapter is dedicated to learning a spatio-temporal representation of affect, based on the frameworks designed in the prior chapters.

Chapter 8: The chapter provides details on the generalisation capability of the proposed methodologies. The affect representations learnt are tested on various datasets, comprising different subjects in varied environments (for example, pose, illumination, etc.).

Chapter 9: This chapter revisits the contributions of this thesis and concludes the current research. The chapter also discusses future directions, ethical impacts, and the broader impact to the research community and its stakeholders.

Chapter 2

Dimensional Affect Analysis: A Literature Review

Contents

The exploration of affect from a computational perspective holds a pivotal role in HCI. As affective computing draws inspiration from diverse domains such as psychology, computer science, and neuroscience, this chapter provides a comprehensive overview of prior research targeted at the analysis of dimensional affect. By surveying past endeavors, this chapter aims to identify research gaps, list challenges and opportunities that form the basis for the design and execution of the research described in this thesis. Figure 2.1 provides an overview of the topics considered for the literature survey and their relevance to the thesis.

In Section 2.1, dimensional and categorical representations of emotions are discussed. An

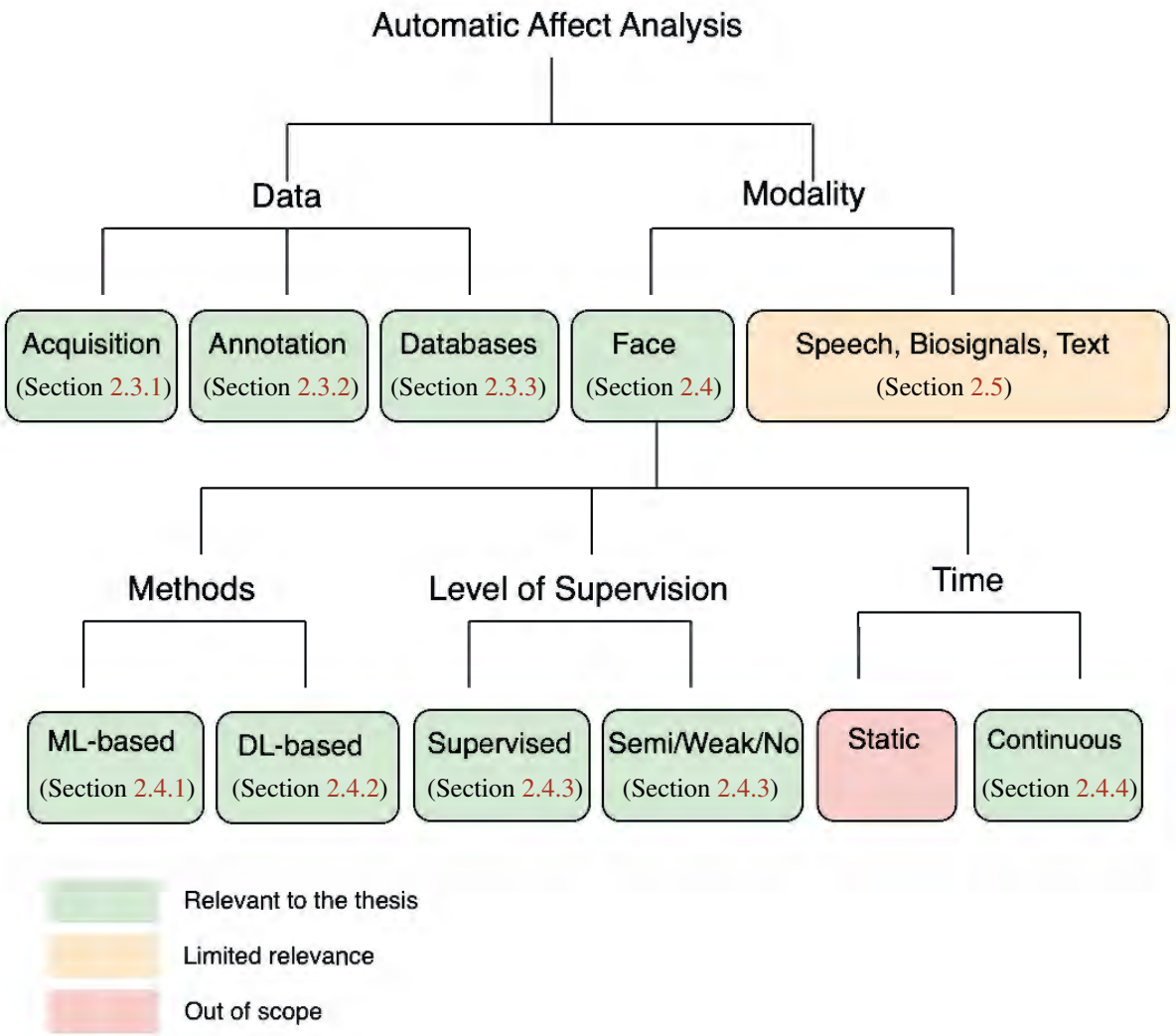


Figure 2.1: Schematic diagram of topics reviewed as part of the literature survey, along with their relevance to the thesis. *Best viewed in colour.*

important distinction is presented in Section 2.2, between time continuous and non-continuous input for affect analysis. Section 2.3 describes typical data acquisition procedure, annotation schemes, and relevant databases available. Section 2.4 reviews all the relevant affect estimation systems where face images/videos are the primary modality. On a related note, dimensional affect estimation from other modalities is reviewed in Section 2.5. Finally, in Section 2.6, research gaps are identified and discussed, which sets a motivation for methodologies discussed in the following chapters. Figure 2.2 provides a brief chronology of related works relevant to the thesis.

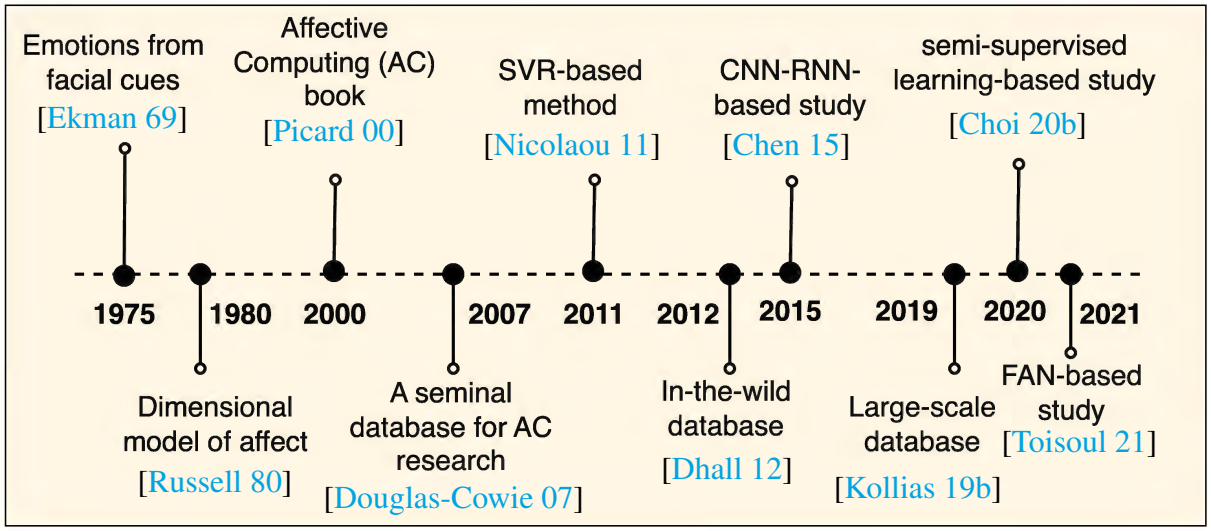


Figure 2.2: A brief chronology of representative works concerning time-continuous dimensional affect estimation. *Timeline not to scale.*

2.1 Dimensional and Categorical Emotions

Recent advances in affective computing deal with modelling human behaviour by considering their emotional state. The term *affect* refers to the underlying experience of emotion, mood or feeling, and the term *affective* is used in a broader sense than *emotional*. The description of affect using latent dimensions is attributed to the study conducted by Russell [Russell 80], with similar studies conducted by other psychologists [Larsen 87, Thayer 90, Watson 85]. The neurophysiological state present as raw feelings, evident in emotion and mood is termed as *core affect*. Core affect is rudimentary, simple and irreducible [Russell 03]. An analogy of core affect is feeling the body temperature. It exists and can be observed at any instance of time. The concept of temperature exists in scientific theory even before words such as *hot* or *cold*. Likewise, a person always has core affect. While there is no direct access to the connections that cause core affect, attributions and interpretations of core affect are made.

At any given instance, the raw feeling is describable as a single point using two bipolar dimensions [Russell 03]. A circular representation of emotions, as proposed by [Schlosberg 52], involved the dimensions of *pleasantness-unpleasantness* and *attention-rejection*. The similarity-dissimilarity of pairs of facial expressions was analysed using a multidimensional scaling procedure, in which similarity was represented by their closeness in a geometric space. This resulted in a two-dimensional (2D) space, with the axes of pleasantness-unpleasantness,

and a combination of attention-rejection and sleep-tension.

Amongst all the dimensional models, the *Circumplex of Affect* model is widely accepted, which states that affect is a linear combination of two dimensional constructs, namely *valence* and *arousal*. As shown in Figure 1.2, valence and arousal are represented along x and y axes, respectively. The three-dimensional (3D) emotional space of *pleasure – displeasure*, *arousal – non-arousal*, and *dominance – submissiveness* [Mehrabian 96] is commonly referred as the PAD (Pleasure-Arousal-Dominance) model.

The axes' polars are defined as:

- **Valence**, referring to the degree of pleasantness or unpleasantness.
- **Arousal**, referring to the degree of excitation or calmness.
- **Dominance**, capturing whether the individual feels in control or not in the environment.

While the dimensional models describe affect in a circular configuration, the other paradigm in the literature posits that emotions can be categorised into discrete and independent classes. The categorical approach claims that there exist a small number of emotions that are "basic", hard-wired in our brain and are recognised universally [Ekman 03]. The dominant theory of emotion in neuroscience research also posits that humans are evolutionarily endowed with a discrete set of basic emotions, and each emotion is independent of others in terms of behavioural, psychological, and physiological manifestations, and arises from the activation within unique neural pathways of the CNS [Posner 05].

The discrete theory of emotions is another approach to representing emotions [Darwin 72, Tomkins 62, Ekman 69]. Although the emotions belong to discrete categories, they represent a family of related affective states [Ekman 03]. The seven universal basic emotions are described as follows [Lewis 10]:

- **Anger** is characterised by feelings of displeasure, irritation, and hostility. It arises in response to perceived threat, frustration or injustice.
- **Fear** is an emotional response to perceived threat or danger. It involves a heightened state of arousal, increased vigilance, and a desire to escape or avoid the threatening stimulus.

- **Surprise** is a sudden and unexpected emotional reaction to an unforeseen event or stimulus.
- **Sadness** is an emotional state characterised by feelings of sorrow or unhappiness. It arises when someone or something important is lost, although varies greatly based on the personal notion of loss.
- **Disgust** is an emotional reaction to offensive or repulsive stimuli. It often arises when perceived repulsion from senses (vision, smell, taste, etc.), actions, or even from ideas.
- **Contempt** is the feeling of dislike or superiority over other person, group of people, or ideology.
- **Happiness** is a positive emotional state characterised by feelings of joy, contentment, and satisfaction.

The existence of basic emotions is disputed by some researchers, and is a long debated problem in the emotion theory literature [Ortony 90, Stein 92, Panksepp 92]. However, the field of affective computing does not hinge on the existence of basic emotions, rather on the representation of emotions as discrete categories or as a continuum.

Early studies on basic emotions state the existence of a core facial configuration reflecting the emotional state of a person [Ekman 11]. In contrast, other scientific frameworks posit that expressions of the same emotion vary substantially across individuals and situations [Barrett 19]. For example, the typical expression of anger (eyebrows furrowed, eyes wide, lips tightened) might sometimes be accompanied with additional facial movements such as a widened mouth, while in other instances, a facial movement might be missing with respect to the prototype. Such variations are considered to be a meaningful part of an emotional expression, because facial movements are functionally tied to other factors such as external context and the person's internal affective state.

There are advantages of the dimensional model over its categorical counterpart. The circumplex model allows modelling of emotions along with their intensity. It also enables similarities and distinctions among various emotion categories [Gunes 13]. On the other hand, categorical

models also have advantages over dimensional models, by simplifying the complex emotional space and being highly interpretable. As noted earlier, categorical emotions are derived from universally shared human experiences, fostering in real-world applications for end users. Thus, it is noted that the choice of dimensional or categorical model of emotion is not universal. In the context of this thesis, dimensional representation of emotion is adopted as it offers more nuanced understanding of the complex nature of emotions than the categorical representations.

In recent years, there has been a notable surge in efforts to model categorical and dimensional emotions for automatic affect estimation [Li 22]. Since the dimensional approach poses a greater challenge and complexity for affect estimation systems, many approaches focus on discretising the continuous dimensional space. A frequently employed approach involves transforming the challenge of estimating dimensional emotions as a regression problem into a two-class classification problem (positive vs negative emotions, or high valence vs low valence emotions) (*e.g.* [Caridakis 06]), or a three-class classification problem (positive, neutral, and negative emotions) (*e.g.* [Yu 04]), or a four-class classification problem (classifying 2D valence-arousal space to four quadrants, namely Quadrant I: high arousal, positive valence (for example, joy), Quadrant II: high arousal, negative valence (for example, anger), Quadrant III: low arousal, negative valence (for example, sadness), and Quadrant IV: low arousal, positive valence (for example, calm)) (*e.g.* [Fragopanagos 05]). Furthermore, [Wöllmer 08] discretises the valence-arousal space using a Conditional Random Field to predict the quantised labels.

On the other hand, a continuous representation of emotions is achieved by various researchers, including the widely accepted Circumplex model [Russell 80]. From an affective computing perspective, [Hoffmann 12] maps discrete emotion classes onto the dimensional space. Recently, [Zhou 23] maps categorical emotions of speech to points on the 2D valence-arousal space using self-supervised learning.

Thus, to comprehend human emotions comprehensively, both categorical and dimensional models are essential. Categorical models offer insights into universally recognised expressions. However, relying solely on categories may oversimplify the complex nature of emotions. By integrating dimensional models, such as the valence-arousal framework, a more comprehensive understanding emerges, capturing both the specificity of categories and the continuous, dynamic

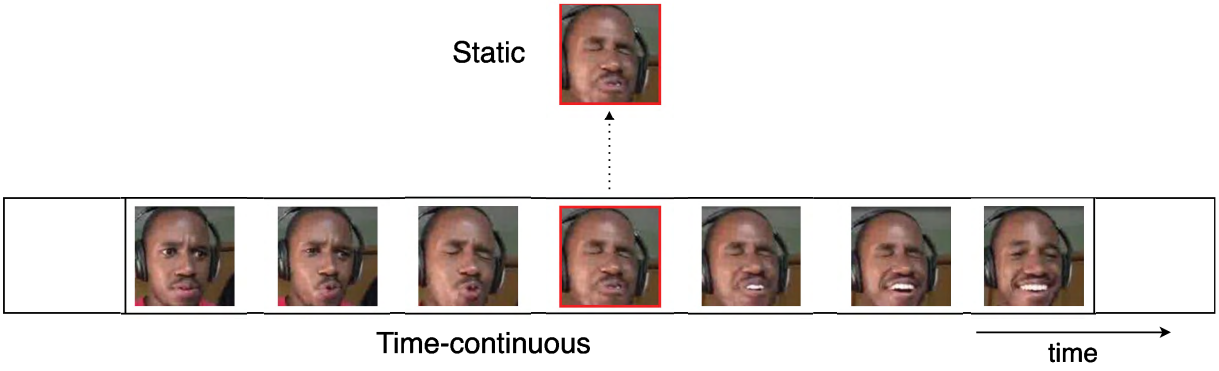


Figure 2.3: Temporal Dynamics of Emotion: The lower frames capture the evolving emotional expressions in a video sequence, while the isolated frame above might convey a different emotional context when viewed independently. Courtesy: the frames are part of the Aff-Wild2 database [Kollias 19b].

aspects of human emotional experiences.

2.2 Time Continuous vs Non-continuous Input

Affect evolves dynamically over time when perturbed by external events [Hamaker 15]. Additionally, since affect drives human behaviour, understanding the dynamics of affect is crucial, as it subsequently shapes human experiences. Intensive Longitudinal Data (ILD), capturing day-to-day, moment-to-moment, or even second-to-second measurements are essential for comprehending affective processes [Bolger 03]. ILD data can comprise univariate or multivariate measurements [Hamaker 15]. While univariate processes focus on general development over time [Moberly 08], multivariate processes refer to the influence between variables within the same occasion [Bringmann 13]. Another important factor for understanding affect dynamics is whether the intraindividual process are stationary or nonstationary [Hamaker 15]. While stationary processes are characterised by fluctuations over time, nonstationary processes can be modelled by choosing parameters that change slowly over time.

Considering time-continuous input for affect analysis presents several distinctions compared to non-continuous input. In non-continuous input scenarios, data is typically segmented, restricting it to encompass a singular affective event, such as an image containing an expression, body pose, or pain. Instead, in the time-continuous scenario, the input contains a sequence of affective events, capturing changes over time, as illustrated in Figure 2.3. Funda-

mentally, the latter is a valid setting as the affective experiences are not static entities but are influenced by an array of internal and external factors [Barrett 99]. The need for temporal affect dynamics is also echoed from studies in facial mimicry, face recognition, and neuroimaging [Sato 07,Wehrle 00,Mühlberger 10]. For instance, in a neuroimaging study, [Mühlberger 10] compared the brain response of the starting and stopping of the same emotional expression. The study revealed that the activity in the brain networks for the onset and endpoint of emotional facial expressions were different, thus, highlighting the importance of temporal modelling of facial expressions for social communication.

The inclusion of temporal dynamics contributes to a more comprehensive portrayal of affective experiences, encompassing the fluidity and transitions within emotional states. Neglecting temporal aspects poses a risk of oversimplification in models, potentially constraining their relevance in real-world scenarios by not adequately capturing the intricate nature of human emotions [Gunes 13]. Also, continuous affective input, by nature, demands swift and adaptive analysis to provide timely insights into evolving emotional states. This real-time responsiveness is critical for applications ranging from HCI to healthcare [Parameshwara 22], where timely recognition of emotional cues is imperative for effective system responses [Gunes 13].

2.3 Data and Databases

At the core of affect analysis lies the design and development of affect datasets. There have been numerous datasets proposed in the space of affective computing over the years. For automatic affect estimation systems, these datasets are indispensable for the training and development of machine learning and deep learning models. In this section, the typical data acquisition and annotation procedures and available databases are reviewed.

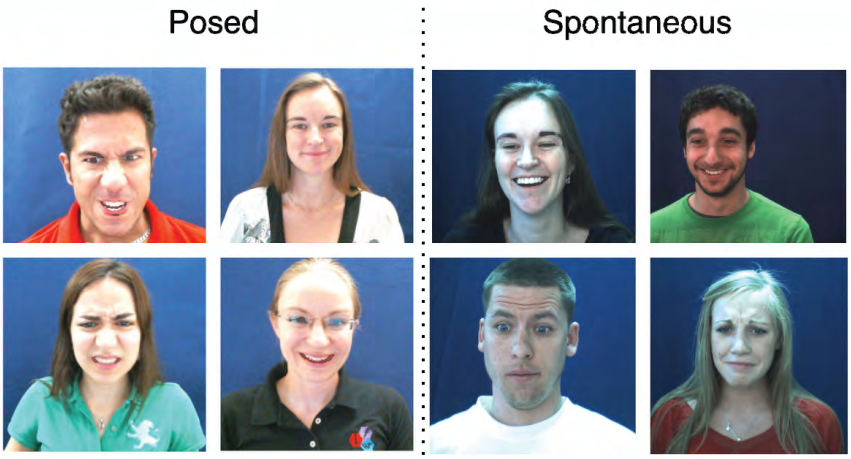


Figure 2.4: Subjects portraying posed (left) and spontaneous (right) facial expressions. Images are extracted from video frames of DISFA+ [Mavadati 16] (left) and DISFA [Mavadati 13] (right) database.

2.3.1 Data Collection

Posed vs Spontaneous Emotional States

To ensure accurate interpretation of emotions in natural and real-world scenarios, owing to the complementary nature, it is important to balance the use of posed and spontaneous data. Numerous studies have been conducted to examine the differences between posed and spontaneous emotional expressions [Borod 83, Cohn 04, Ekman 04]. *Posed* expressions are deliberate expressions of an individual, which are clearly intended, or requested of him/her. On the contrary, *spontaneous* expressions are involuntary unintended reactions to evocative situations [Myers 76]. Figure 2.4 illustrates apparent differences between posed and spontaneous facial expressions.

Zygomatic major, a muscle that controls facial expression drawing the mouth's corner upward and outward, and *orbicularis oculi*, the muscle responsible for closing the eyelids, contract during spontaneous smiles, while during a posed smile, the zygomatic major alone contracts [Ekman 82]. Behavioural research indicates that posed and spontaneous expressions differ in terms of speed, trajectory, movement of facial muscles, and the total duration of onset and offset. It is observed that the trajectory often appears smoother for spontaneous expressions, and the total duration of the phases of expression is longer. While for posed expressions, the duration of the phases is shorter and the onset is more abrupt [Ekman 82, Ekman 04].

Both posed and spontaneous emotional states have their own significance and utility. Posed emotions occur in controlled settings, offering predictability and consistency for research or modelling purposes. On the other hand, spontaneous emotional states reflect genuine emotional responses occurring naturally in real-life situations, providing a more authentic representation of human emotions. While posed audio and visual display of affect can be automatically detected with reasonably high accuracy, there has been a shift to the automatic estimation of spontaneously displayed affect [Zeng 07]. This is because deliberate expressions differ in visual appearance [Jia 21], audio profile [Juslin 18], and timing [Cohn 04]. Several studies are performed for the detection of spontaneous facial expressions [Li 23, Pfister 11, Dibeklioglu 12] and vocal expressions [Zhang 21, Feng 23, Zhang 19].

Controlled vs Uncontrolled Environment

Early studies examining emotions and their expressions focused on *acted* emotion expressions. That is, deliberate expressions were observed in experiments where an actor or actress was asked to exhibit an emotion. In such scenarios, the person consciously tries to feel and communicate a particular emotional state. This results in physical behaviour, which deviates from their spontaneous counterparts, as spontaneous emotions occur more naturally. Spontaneous facial expressions are characterised by subtle, minimal facial deformations, which are difficult to track, and frequent out-of-plane head movements whose effects are difficult to remove [Kossaifi 17]. Recent studies model affect under real-world conditions, where affect is mostly recorded in uncontrolled setting or *in-the-wild*, where there is little or no control over the lighting conditions, less constrained movement of subjects, etc. Figure 2.5 shows examples from a few representative databases for the comparison of controlled and uncontrolled environments.

In the human facial analysis, realistic data plays an important role. Affective data recorded in tightly controlled laboratory experiments is not a true representation of the real world [Dhall 12]. Recent studies focus on the analysis of incorporating spontaneous affective states using facial expressions [Cohn 04, Bartlett 05, Valstar 06], and vocal intonations [Batliner 03, Lee 05]. Also, multiple studies explore the differences between spontaneous and posed behaviour [Cohn 04, Valstar 07]. In [Buller 94, Buller 96], the study suggests that truthful and deceptive behaviour

Figure 2.5: Comparison of environments for affect observation: On the left, a controlled environment provides a structured setting for affect analysis, while on the right, an uncontrolled environment introduces real-world complexities. Images extracted from different databases: (clockwise from top left) EMMA [Katsimerou 16], AFEW [Dhall 12], SEWA-DB [Kossaifi 19], DISFA+ [Mavadati 16].

differs on the number of head movements, or the lack of accompanying gestures [DePaulo 03].

2.3.2 Annotation

Most of the affect detection systems rely either directly or indirectly on the variations of supervised learning algorithms to identify relationships between the machine-readable signals (for example, facial expressions, physiological signals, etc.) and affective states. These algorithms require *ground truth* information in the form of annotations (or labels) for affective data. Acquiring annotations is a non-trivial task for affect detection, as unlike supervised algorithms in other domains, such as biometrics forecasting, it is not possible to acquire affective ground truth information. Affect must be approximated as it is a psychological and latent variable, rather than a physical property (for example, height or weight) or identity property (for example, denoting person X).

The *source* of annotation refers to the type and number of individuals performing the annotation [D'Mello 16]. The most common approach is *observer*-based, where observer refers to the person other than the subject expressing the emotion. Affect annotation is considered complicated [Metallinou 13], as it demands empathetic skills and an understanding of the description

of affect. One of the methods could involve a small number of *skilled* observers, or experts at a relatively high cost per observer. For instance, a psychiatrist/psychologist could provide annotations for depression. Another method is obtaining annotations from a large number of *novices*, or unskilled observers, at a relatively low cost. For instance, individuals recruited from *crowd-sourced* platforms such as Amazon Mechanical Turk [Morris 15].

Annotations can be gathered by the subject themselves, where they self-report their affective state either by following the *emote-aloud* protocol [Scotty D. Craig 08], or by filling questionnaires. This is particularly suitable for annotating physiological signals, where the subject's own experience has to be captured. The Self-Assessment Manikin (SAM) is a commonly used non-verbal pictorial assessment tool used to measure the subject's valence, arousal, and dominance associated with the affective state [Bradley 94]. As another setup, annotations can denote the emotions perceived, recognised, or interpreted by observers rather than the actual emotional experiences of the subject. Observers are presented with visual cues such as facial expressions, body language, or gestures in images, videos, or real-time interactions, or vocal cues such as speech patterns or vocal intonations to infer perceived emotions.

Over time, several annotation tools have evolved with different capabilities. Though originally designed for linguistic annotations, ELAN [Brugman 04] is also used in annotating affective data in audio and video recordings, and allows for multimedia annotations on multiple layers called *tiers*. The annotation can be a sentence, a word, or a comment of any feature observed in the multimedia content. Anvil [Kipp 01], an Extensible Markup Language (XML)-based annotation tool, offers multi-layered annotation based on a user-defined coding scheme. Anvil allows importing ELAN files, 3D viewing of motion capture data, coding agreement analysis, and managing the corpora of annotation files. The FEELtrace [Cowie 00] annotation tool is designed to record perceived emotional content via self-reports of observers of a time-continuous stimulus. A two-dimensional space appears on the interface, allowing the annotators to watch the audiovisual recordings and move their cursor to rate their perception of the emotion on a scale of $[-1, 1]$. As a successor of FEELtrace, GTrace [Cowie 12] (General Trace), is designed to allow users to craft their own scales with minimum effort along a particular affective dimension. GTrace is an interface where the annotator can watch an audiovisual recording and move

a cursor simultaneously to provide affect annotations.

When discrete emotions are to be annotated, researchers adopt a 10-point Likert scale capturing the valence and arousal (0 as low valence/arousal, 9 as high valence/arousal), or different levels between -1 and +1 [Gunes 13]. Another challenge is obtaining a high inter-rater agreement, which refers to the level of consistency between the raters when annotating the same set of data. Commonly used statistical measures, which quantify the inter-rater agreement, include Cohen's Kappa [Cohen 60], Intra-class Correlation Coefficient (ICC) [Bartko 66], and Fleiss' Kappa [Fleiss 04].

Some of the affective phenomena such as pain, are fundamentally subjective and difficult to assess [Cohn 10]. They pose a challenge as they can be typically measured only by self-assessment reports. Overall, obtaining appropriate ground truth from both categorical and dimensional annotations, modelling the inter-rater agreements, and finding the signals which better correlate with self-assessment remain as challenges in the field [Fernandez Rojas 23].

2.3.3 Databases

One of the earliest systems to taxonimise human facial movements is the *Facial Action Coding System* [Friesen 78]. It was designed to systematically categorise facial movements known as *Action Units*, corresponding to specific facial muscles' contractions or movements. Although not a traditional database, FACS laid the foundation for understanding facial expressions and their relation to emotions. FACS is purely descriptive, and does not have any labels associated. Besides associating with the AUs, discrete emotion categories and continuous emotion dimensions of valence and arousal values are also employed to annotate the data samples.

With the rise in affective systems in the early 20th century, concurrently, numerous affective databases were developed. Figure 2.6 illustrates evolution of data points (images) of prominent affect databases. One of the databases, which is widely used even today, is the ***Cohn-Kanade (CK) database*** [Kanade 00]. The CK database is an AU-coded facial expression database, comprising 2105 digitised image sequences. Facial behaviour of 210 adults (69% female, 31% male) aged between 18 and 50 years, was captured in an observation room equipped with two cameras. The subjects were instructed to performed 23 facial displays, which included single

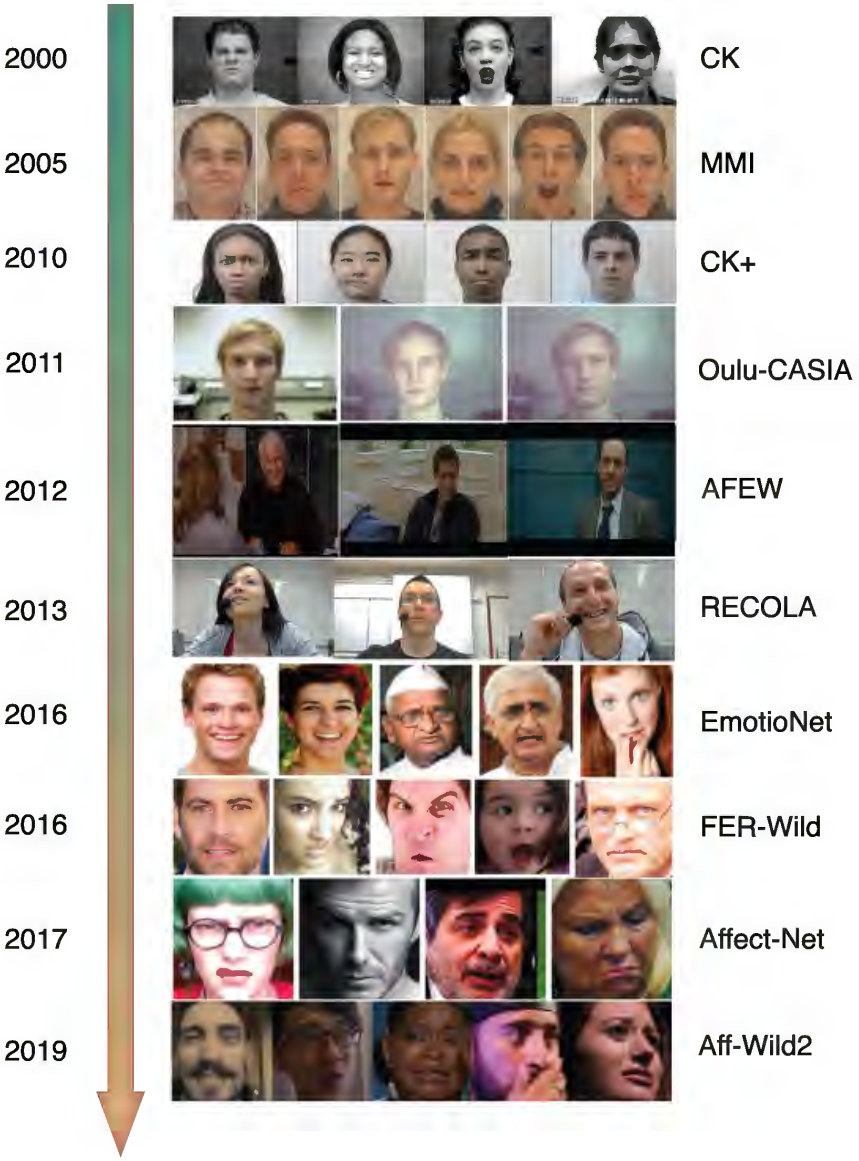


Figure 2.6: Evolution of Affect Databases: A visual timeline showcasing images from prominent diverse emotion databases spanning the early 2000s to the present day. The progression highlights the transformative changes in affective data collected over time reflecting the evolving landscape of affective computing research. *Timeline not to scale. Best viewed in colour.*

action units or a combination of action units, each beginning from a neutral or nearly neutral face. To date, 1917 image sequences are annotated with AUs, except AU 13 (corresponds to an elevated angle of the mouth). In 2010, an extended version of the CK database was released, called the **CK+ or Extended CK database** [Lucey 10]. CK+ has an additional 593 image sequences from 123 subjects augmented to the CK database. These 593 additional data are posed sequences with 10 to 60 frames, incorporating onset to peak formation of the facial expressions. Similar to the CK database, these sequences are annotated with AUs. Additionally, they are annotated with a nominal emotion label depending on the subject's impression of the categories contempt, anger, disgust, happiness, sadness, surprise, and fear.

Another example of a facial expression database based on FACS is the ***MMI*** [Pantic 05] database. It is a facial expression database with ≈1500 samples of both static images and image sequences of faces in frontal and profile view. 19 subjects (44% female, 56% male) in the age range of 19 to 62 years participated in the acquisition process. The images captured are of various facial expressions of emotions, single AU activation, and multiple AU activation. Image sequences are of variable length, ranging from 40 to 520 frames. About two-thirds of the image samples and 169 sequences are FACS coded by two experts.

There are databases where facial expressions are collected beyond the visible spectrum. For example, ***Oulu-CASIA*** [Zhao 11] facial expression database comprises images collected beyond the visible spectrum in *near-infrared* (NIR) band. The data is collected from 80 subjects (73.8% males, 26.2% females) in the age range of 23 to 58 years. One part of the database is collected in Oulu with 50 subjects, most of whom were Finnish, and the other part in Beijing with 30 Chinese subjects. Subjects were shown example pictures, and were asked to depict facial expressions accordingly, which was captured using a NIR camera and a visible light spectrum (VIS) camera. These expressions were annotated for six expressions namely, surprise, happiness, sadness, anger, fear, and disgust.

While these lab-controlled datasets offered a foundation for facial affect recognition research, their limitation lies in the controlled nature of the environments, which may not fully reflect the spontaneous and dynamic nature of emotions in everyday life. There exists an extensive list of databases for the task for affect recognition/estimation [Khan 20]. Considering the

scope of this thesis, face affective databases that are either labelled with dimensional affect or recorded in uncontrolled or *in-the-wild* settings are briefly reviewed.

Following is the list of affective databases widely used for facial affect estimation. Table 2.1 gives an overview of the following databases: **HUMAINE** [Douglas-Cowie 07]: One of the early databases with dimensional affect is the HUman-MAchine Interaction Network of Excellence (HUMAINE) database. The database is an ensemble of videos capturing a person either acting or talking or both. The videos capture a large spectrum of expressions generally shown in everyday life. The database consists of 50 naturalistic, induced and acted videos. The emotions are annotated time-continuously for valence and arousal. Additionally, the annotators were instructed to self report the underlying perceived *mood*.

AFEW [Dhall 12]: Acted Facial Expressions in the Wild (AFEW) is an acted facial expression database curated from movies. The database captures varied facial expressions with subjects from various ethnicities, gender, ages, etc., displaying natural head movements and occlusions, thus making it one of the more challenging databases for automatic affect estimation. The videos chosen are based on movie subtitle parsing and a recommender system, before being annotated with categorical emotions by two experts. The database consists of 330 subjects, with age range 1–70.

RECOLA [Ringeval 13]: REmote COLlaborative and Affective interactions (RECOLA) is a multimodal database based on spontaneous interactions for a collaborative task. The tasks are performed remotely (in the laboratory) and the dyadic (two-party) interactions are held in French. The corpus contains 46 participants, divided into 23 dyadic teams. Each participant self-reported their current emotional state and were asked to individually think about a solution for the survival task. Next, participants engaged in a dyadic interaction on the survival task where the multimodal data is recorded (video with audio, ECG, and Electrodermal Activity (EDA)). From $\approx$ 9.5 hours of recording the interaction, 3.8 hours of the video data is annotated with time-continuous valence and arousal values by six annotators.

EmotioNet [Benitez-Quiroz 16]: The EmotioNet database comprises one million facial expression images sourced from the Internet. The images were obtained from their associated keywords, where all keywords are related to the word “feeling” from WordNet [Miller 95]. The

images (with faces extracted) are then automatically annotated with activated AUs and their intensities, employing a Kernel Subclass Discriminant Analysis (KSDA) approach. The images are then automatically labelled with categorical emotions (basic and compound emotions, 23 in total) based on the presence of specific AUs. For example, an image with active AUs 1, 2, 12, and 25 is labelled as "happily surprised". The annotation is validated on a subset of the dataset (10%) by experienced AU coders. In its current form, the dataset lacks dimensional affect labels.

FER-Wild [Mollahosseini 16]: The Facial Expression Recognition from the *wild* web (FER-Wild) dataset comprises 24,000 images collected from the World Wide Web (WWW). This is done through querying three search engines using 1250 emotion related keywords (examples, *happy face*, *laughing man*, etc.). The images are annotated with categorical emotions by two annotators. The dataset is imbalanced towards certain categories, such as *disgust* and *fear*, and is not annotated for dimensional affect.

EMMA [Katsimerou 16]: The EMotion and Mood Annotations (EMMA) database consists of 180 long videos (0.33 – 5.23 minutes, with an average duration of 2.03 minutes) based on *improved acting* of non-interactive daily routines. The actors were 15 Dutch acting students (9 female and 6 male), who were provided with situational context and scenarios for the *mood* induction procedure. Unlike the previously mentioned databases, the videos in this database are annotated via a crowd-sourcing platforms for mood (*positive*, *negative*, or *neutral*) and dimensional affect. Each video is annotated by at least ten annotators.

AffectNet [Mollahosseini 19]: Similar to [Mollahosseini 16], Affect from the InterNet (AffectNet) is an image-based database curated by querying three search engines with emotion-related keywords, such as "joyful girl", "blissful Spanish man", etc. The database comprises one million images, with 450,000 images annotated by experts for categorical and dimensional affect.

AFEW-VA [Kossaifi 17]: Acted Facial Expressions in the Wild with Valence and Arousal (AFEW-VA) dataset consists of 600 videos from feature films. This collection of videos is a subset of the AFEW dataset by [Dhall 12]. Principally differing with AFEW in terms of annotations, AFEW-VA is annotated with per-frame annotations of valence and arousal values

Table 2.1: Overview of dimensional and categorical affect databases. The 'Subjects' column indicates the total number of subjects in the dataset, along with male (M) and female (F) distribution. *Details provided are of the additional data added to Aff-Wild. The union data is generally referred to as *Aff-Wild2*.

Dataset	Year	Environment	Source	Data type	Number of samples	Subjects	Annotation type	Number of annotators	Affect Label
HUMAINE	2007	Controlled	Laboratory	Video	50 videos	-	Manual	-	Dimensional
AFEW	2012	Wild	Movies	Video	1,426 videos	330	Semi-automatic	2	Categorical
RECOLA	2013	Controlled	Laboratory	Video, ECG, EDA	3.8 hours of video recording	46 (M: 19, F: 27)	Manual	6	Dimensional
EmotioNet	2016	Wild	Web search engine	Image	1,000,000 images	-	Automatic	-	AUs and categorical
FERWild	2016	Wild	Web search engine	Image	24,000 images	-	Manual	2	Categorical
EMMA	2016	Controlled	Laboratory	Video	180 videos	15 (M: 6; F: 9)	Manual	10 per video	Dimensional
AffectNet	2017	Wild	Web search engine	Image	440,601 images	-	Manual	1 per image	Categorical and Dimensional
AFEW-VA	2017	Wild	Movies	Video	600 videos	240	Manual	2	Dimensional
SEWA-DB	2017	Wild	Video chat	Video	538 videos	398	Manual	23	AUs and dimensional
Aff-Wild	2017	Wild	YouTube	Video	298 videos (118,000 frames)	200 (M: 130; F: 70)	Manual	6	Dimensional
OMG	2018	Wild	YouTube	Video	567 videos (7,371 utterances)	-	Manual	5 per video	Categorical and Dimensional (utternace-level)
Aff-Wild2*	2019	Wild	YouTube	Video	260 videos (141,300 frames)	258 (M: 149; F: 109)	Manual	4	Categorical and Dimensional

for each video. Thus, with videos of variable length, from 10 frames to 145 frames, the dataset contains annotations for more than 30,000 frames. Unlike previously, where annotations were done using trace-style tools [Sneddon 11], AFEW-VA annotations are done on a frame-to-frame basis by two annotators. The valence and arousal values are in the range of -10 to 10 integer values, resulting in a total of 21 discrete values.

SEWA-DB [Kossaifi 19]: Automatic Sentiment Estimation in the Wild Database(SEWA-DB) is an audio-visual database containing 398 subjects from six cultures. Subjects were recorded either while watching advertisements or while discussing advertisements in a video chat. Out of the entire database, 538 clips (ranging 10-30 seconds) contain annotations for valence, arousal, and liking/disliking of the advertisement. The database also includes annotations for facial landmarks and Action Units.

Aff-Wild [Zafeiriou 17] and **Aff-Wild2** [Kollias 19b]: The Affect-in-the-Wild (Aff-Wild) database comprises 298 videos curated from YouTube, a video sharing website. The "reaction" word is used as a video retrieval keyword, as the videos would then contain subjects reacting to a variety of topics (reviewing movies, products, etc.). The database consists of 200 subjects

(130 male and 70 female), with a total duration of ≈ 30 hours and 1,180,000 frames. The videos are annotated by eight lay experts with dimensional affect (valence and arousal).

In 2019, an updated database, called as Aff-Wild2 was released. The additional data comprises an additional 260 videos, with 258 subjects (149 male and 109 female). To curate more videos, the videos were again retrieved from YouTube using the "reaction" keyword, along with other related words from the dimensional model of affect, as depicted in 1.2. The new dataset is annotated using four experts. The length of the videos range from 3 seconds to 15 minutes and 4 seconds. Thus far, this is the largest in-the-wild affect video database with dimensional affect annotation.

OMG [Barros 18]: One-Minute Gradual (OMG) Emotion dataset is composed of 567 videos collected from YouTube based on the keyword "monologue". Videos were separated into clips based on utterances, resulting in 7371 samples. Each utterance was annotated by at least five annotators from a crowd-sourcing platform. The database is annotated with dimensional and categorical affect.

2.4 Affect Estimation from Face

The face is one of the most expressive non-verbal channels for communicating emotions [Mehrabian 17]. Affect estimation from face has been extensively researched due to its multifaceted applications including enhanced HCI, aid in mental health assessment and therapy, learning and engagement, analysis of consumer experience, etc. Automatic Facial Affect Estimation (FAE) has seen significant advancements due to accelerated improvements in computer vision, deep learning techniques, and the availability of larger and more diverse datasets. Broadly, FAE typically involves using either *static* data, where only a single image is used to decode the facial cues, or *temporal* data, where a sequence of contiguous frames are used for inferring the emotion. The following subsections provide an overview of the progression of FAE using ML approaches, Deep Learning (DL) algorithms, varying levels of supervision, and temporal modelling frameworks. The following subsections describe related studies, which either focus on dimensional affect estimation or develop methods for in-the-wild settings.

2.4.1 Machine Learning-based Studies

The relevance of ML algorithms in affect estimation is driven by their ability to process patterns in data. Properly representing and engineering features from raw data significantly influences algorithm performance. The baseline approach for the dimensional affect estimation involves a form of static regression, for example linear regression [Gupta 14], partial least squares regression [Meng 13], or support vector machines for regression (SVR) [Nicolaou 11]. One of the early effective works on dimensional affect estimation is based on features characterising head movements, face appearance and voice. Nicolle *et al.* [Nicolle 12] model the dynamic information by computing the log-magnitude Fourier spectra of the temporal signals to describe the evolution of the visual cues, and employ a kernel regressor based on the Nadaraya-Watson estimator [Nadaraya 64].

Additionally, ML algorithms for FAE largely depend on a handcrafted feature extraction process. *Handcrafted features* refer to manually designed or predefined features extracted from raw data. In the context of FAE, these features are derived from images or videos of faces before being fed into machine learning models. Handcrafted features are categorised into *geometry-based features* or *appearance-based features*. Geometry-based features refer to facial attributes or characteristics derived from the spatial arrangement, shape, and geometry of facial elements, for instance, facial landmarks (such as the corners of the eyes), facial proportions (such as the ratio of eye width to face width), angle measurements (such as the angle of eyebrow arching), etc. Appearance-based features refer to visual characteristics and attributes derived from the overall appearance of the face [Wang 22], for instance, colour information (such as skin tone changes), pixel intensity patterns, etc.

In [Ghimire 13], the authors extract geometric features from faces in videos by employing facial landmark tracking, and perform facial expression recognition using Support Vector Machine (SVM) algorithm on the features selected using the AdaBoost techniques. A Kinect device is used to detect and track facial movements in [Sujono 15], followed by extracting key facial features from the Active Appearance Model (AAM) [Cootes 01]. Changes in the values of these key features are then observed using Fuzzy Logic. To tackle the variations in local

neighbourhoods of pixels, a Local Prominent Directional Pattern is proposed in [Makhmudkhujaev 19] that encodes the local shape by using local statistical information of a pixel neighbourhood.

The contractions of facial muscles produce changes in the appearance of permanent facial features namely lip, eyes, etc., and transient facial features such as any facial lines and furrows that are not present at rest. An automatic system to analyse subtle changes in facial expressions based on AUs by using the permanent and transient features is developed by Tian *et al.* [Tian 01]. A real-time emotion classification algorithm is proposed in [Happy 12] by applying Local Binary Pattern (LBP), which is a texture descriptor used for capturing local patterns and texture variations in different regions of an image. Another robust approach for dimensional and continuous prediction of emotions from naturalistic facial expressions is through a variant of the Relevance Vector Machine (RVM), called the Output-Associative RVM framework, which learns non-linear dependencies between the input and output affective data [Nicolaou 12].

In [Gu 12], the authors divide images into local regions and each region is subjected to Gabor-filter operations to obtain local features representing facial expressions. Principal Component Analysis (PCA) and Fisher Linear Discriminant analysis are applied to obtain salient information and the outputs are concatenated to form global features, which are further fed to classifiers.

Other studies have employed Nearest Neighbor [Fasel 00], Bayesian Networks [Cohen 03], and AdaBoost classifiers [Wang 04] to recognise and model facial expressions in terms of emotions. However, traditional ML algorithms require handcrafted feature engineering, which might not encompass all the relevant information. It is also a resource-intensive task, requiring significant time and effort. On the contrary, DL algorithms learn feature representations from raw data and automatically extract hierarchical features through multiple layers in the neural networks.

2.4.2 Deep Learning-based Studies

Deep learning models comprise multiple layers that learn abstract features. They can handle large volumes of data efficiently and are known for their scalability. The *feedforward step* in

neural networks is the process of passing input data through the network's layers to generate an output prediction or inference without any feedback or loops. The output generated by the network is compared to the actual or expected output (ground truth) using a *loss function*. This is followed by a *backward propagation* of errors, where the gradient of the loss function with respect to each weight and bias in the network is calculated. The network learns from its mistakes by adjusting the weights and biases in a way that minimises the prediction error. Through repeated iterations of forward and backward passes, neural networks can improve their performance.

Since 2013, competitions such as FER2013 [Goodfellow 13], Emotion Recognition in the Wild (EmotiW) [Dhall 15], and Affective Behaviour Analysis in-the-wild (ABAW) [Kollias 23] have facilitated the collection of large amounts of data from real-world scenarios, which have consequently promoted comprehensive research on FAE. Additionally, the formidable increase in processing capacities and state-of-the-art network architectures have led to a surge in adoption of DL algorithms for FAE.

Convolutional Neural Networks-based Studies

Convolutional Neural Networks are a class of deep neural networks, primarily used in computer vision tasks, particularly for analysing visual imagery. They are designed to automatically and adaptively learn spatial hierarchies of features directly from raw pixel data. CNNs include *convolutional layers* that apply convolution operations to input images. They consist of *filters* (also called *kernels*) that slide across the input image, detecting spatial patterns and features. *Pooling layers* reduce the spatial dimensions of the data by down-sampling or summarising the information. Nonlinear *activation functions* (for example, the Rectified Linear Unit (ReLU)) introduce non-linearity into the network, enabling it to learn complex relationships in the data. *Fully connected layers* connect every neuron from one layer to every neuron in the subsequent layer, integrating high-level features learned by the previous layer. CNNs were originally designed for the handwritten digit recognition task, but given their capacity and advancements in computational power, they had a significant impact on diverse computer vision tasks, such as object detection [Redmon 16], human action recognition [Tran 18], pose estimation [Radwan 13, Rad-

wan 19], etc.

CNN-based models are applied on various affect databases in [Breuer 17] to provide a computational justification for FACS. Additionally, the filters in the trained models were visualised to demonstrate the capability of CNNs for emotion detection tasks. In [Kossaifi 17], the authors used a fine-tuned AlexNet, one of the influential neural network architectures for computer vision tasks, on the AFEW-VA database. Following the advancements in tensor learning methods, [Mitenkova 19] incorporated Tensor Regression Layers [Kossaifi 20a] into the CNN-based architectures. The advantages of using tensor-based methods are a) they preserve the multi-modal structure of the affect data, and b) the number of parameters is lower owing to the low-rank structure imposed on the regression weights. In [Kossaifi 20b], the authors extended the previous tensor factorisation framework for multidimensional convolutions of a higher order. Another attempt on proposing a *light-weight* architecture is by [Handrich 20], where the authors propose an architecture based on YOLO (you-only-look-once) [Redmon 16], a popular algorithm primarily designed for the task of *object detection* in the computer vision literature. The authors show that the proposed approach simultaneously detects faces, categorical and dimensional affect from raw input image.

The authors of [Toisoul 21] integrated face alignment and emotion estimation tasks by proposing an end-to-end CNN-based network that jointly estimates categorical, dimensional affect, and facial landmarks, suitable for real-time applications. Recently, [Hassani 22] proposed a residual-based network architecture, called BReG-NeXt, using a function with bounded derivative instead of a traditional identity mapping in the residual units for categorical and dimensional affect estimation. However, different from other recent works, instead of improving the CNN architectures for dimensional affect estimation, the authors of [Kim 21] proposed an adversarial learning scheme with the CNN as the encoder to learn facial feature information. Further, the authors of [Kim 22a] propose a feature transformation-based affect representation learning. That is, the contrast between the transformed features and overall facial features is quantified through contrastive learning, while dimensional affect estimation on 2D valence-arousal space is done based on the angle and the intensity of the learnt emotion representations. Interestingly, the dimensional affect estimation task is modelled as an identity-matching prob-

lem in [Kim 22b], where the proposed method finds and utilises optimal identity pairs based on the emotion similarity, in turn using this information for dimensional affect estimation.

While CNNs have been successful for affect estimation [Narayana 22, Kollias 19b], training them from scratch may require large amounts of task-specific labelled data to achieve reasonable performance. When training models with limited data, there is a higher risk of *overfitting*, where the model learns specific patterns in the training data, but does not generalise well to new, unseen data. *Transfer learning* is a powerful technique that leverages pre-trained models, typically trained on large-scale datasets, and adapts them for specific vision tasks with limited data [Yosinski 14]. Transfer learning in computer vision commonly employs two strategies; (a) *fine-tuning*, where the pre-trained model is further trained on a new dataset (target task). While the early layers might capture general features such as edges, textures, or shapes, fine-tuning adjusts these layers to adapt to task-specific features while retaining the learned representations, (b) *feature extraction*, where the pre-trained model's learned features are used as fixed representations. Only the final layers (classification or regression layers) are modified or replaced and trained on a new dataset.

Researchers have leveraged pre-trained neural networks such as VGG [Simonyan 15], VGG-face [Parkhi 15], and ResNet [He 16] for FAE. These networks form the *backbone*, or the primary set of layers that forms the core of the network. In [Liu 19], the authors propose a method for identity-disentangled facial expression recognition by using a light-CNN for extracting the identity features, and VGG-face for embedding feature level measurements. The mined images are then fed to an adaptive deep metric learning framework to disentangle the identity factors in a face image. A novel identity-aware CNN is proposed in [Meng 17] to jointly estimate expression and identity related features by using similarity losses. The CNN comprises two branches with identical CNNs, each of them fine-tuned from a pre-trained CNN model, trained on the FER2013 dataset.

The *Neural Attention Mechanism*, inspired by human visual attention, allows the neural networks to focus selectively on specific parts of the input data [Xu 15]. This mechanism enables models to learn to allocate different weights or importance to different parts of the input during computation, allowing for salient features to emerge prominently. The authors

in [Wang 21] tackle the problem of insufficient utilisation of local information by following an oriented attention ensemble approach for inferring the discrete emotions. They employ a multi-branch pseudo-Siamese network with original image as input to one branch and weighted-masked input to the attention branch. In [Marrero-Fernández 19], the authors create probability maps for faces and use a CNN-based attention module for jointly modelling representation and classification of discrete emotions. Using the VGG-face network to extract features from facial features, the authors in [Xie 19] employ an attention-based Salient Expressional Region Descriptor to adaptively produce a unique attentive mask to locate expression-sensitive regions. An encoder-decoder network is further employed for expression classification.

While traditional attention modules are used in computer vision tasks to focus on spatial regions within images or other spatial data structures, *self-attention* is a mechanism used in neural networks to weigh the significance of different parts or elements within the same input sequence. Each element in the input sequence interacts with all other elements to determine how much attention should be placed on each element. Self-attention is a fundamental component of *Transformers* [Vaswani 17], which are a class of neural networks particularly effective for tasks involving sequential data.

Transformer-based Studies

The transformer architecture enables the model to capture dependencies among different positions in sequences efficiently. The encoder processes the input sequence using self-attention layers, and the decoder generates the output sequence based on the encoder's representations. The *self* in self-attention indicates that the attention mechanism is applied within the sequence itself, enabling the model to attend to different parts of the sequence while considering relationships between them. *Multi-head attention*, a crucial component within transformer-based architectures, extends the base self-attention by employing multiple parallel self-attention layers, or *heads*. Each attention head operates independently, learning different sets of attention weights for the input sequence [Vaswani 17].

Vision Transformer (ViT) [Dosovitskiy 21], one of the early studies employing transformer for computer vision tasks, handles structured data such as images by dividing them into patches

and processing them as tokens. ViT's success has led to further research and exploration of transformer-based architectures in vision-based affect estimation approaches. To infer dimensional affect from a single modality, the authors in [Chen 21] employ a 1-dimensional CNN followed by a transformer encoder. Further, to learn the interactions between multiple modalities (audio and video) and to model the temporal dynamics, they employ multi-head attention in the transformer encoder. In [Ma 23b], in-the-wild facial images are converted into sequences of visual words, and attentional selective fusion is performed to dynamically and adaptively combine LBP features and CNN features for facial expression recognition. A similar multimodal fusion of audio and video features is performed in [Meng 22a] for dimensional affect estimation, where an LSTM and Transformer are used as encoders.

In [Zou 23], the authors extract visual features using a pre-trained ResNet-50 model. Additionally, they adopt a transformer encoder with a multi-head attention framework to learn the distribution of both the spatial and temporal features for estimating dimensional affect. The Masked Auto Encoder (MAE)-Face [Ma 22], a self-supervised pre-trained model constructed based on ViT for facial representations, is fine-tuned in [Ma 23a] by employing a two-pass pre-training process and a two-pass fine-tuning process for affect estimation. In [Zhang 23b], a pre-trained MAE is used as a visual feature extractor, while Hubert [van Niekerk 22] and Wav2vec2 [Baevski 20] are employed for acoustic feature extraction. The audio and visual features are concatenated and fed to a transformer for affect estimation. Integration of transformers with multimodal data has led to more comprehensive emotion analysis. The transformers' design for handling sequential data makes them well-suited for understanding emotional expressions.

Generative Adversarial Network-based Studies

Generative Adversarial Networks (GANs) are a class of deep learning models which focus on *generating* realistic data samples that resemble the training set. *Adversarial* describes a training framework where two neural networks, often referred to as adversaries, are pitted against each other in a competitive process. Specifically in the context of GANs, adversarial represents the adversarial relationship between the *generator* and the *discriminator*. The generator is the

component of the GAN that aims to generate synthetic data (images, text, etc.) that resembles the real data from training set. On the contrary, the discriminator is the part that acts as a binary classifier, distinguishing between real data samples from the training set and fake data produced by the generator. These two components are trained concurrently in a competitive manner, as the generator aims to improve by generating more realistic data to deceive the discriminator, while the discriminator aims to become more discerning to accurately classify between real and fake data. This adversarial setup results in the generation of increasingly realistic data by the generator and better discrimination by the discriminator.

In the context for affect inference, GANs either generate synthetic data to augment existing datasets, synthesising emotional content for applications, learning representations of emotions from unlabelled data, or aiding in the enhancement of emotion recognition models. In [Kim 21], the authors propose an adversarial learning approach to infer person-independent dimensional emotions. They use a binarisation process to generate strong and weak emotions, which include facial expressions and inner emotions. Changes in the inner emotions, which are not revealed in the facial expressions, can be used for adversarial learning. Adopting StarGAN [Choi 18], a GAN-based model designed for image-to-image translation tasks for continuous emotion synthesis, VA-StarGAN is proposed in [Kollias 20] for dimensional emotion estimation. A conditional difference adversarial autoencoder (CDAAE) is used in [Zhou 17] for facial expression synthesis, where the facial image of a previously unseen subject is used to generate an image of the same subject with a target emotion or facial AU label.

To enable the expression intensity to be continuously adjusted from low to high, an Expression Generative Adversarial Network (ExprGAN) is developed in [Ding 18] for photo-realistic facial expression editing. The authors rule out using training data with intensity values, as ExprGAN has the capability to synthesise multiple diverse styles of target expressions, while controling the intensity from weak to strong. [Pumarola 18] propose GANimation, an AU-conditioned adversarial architecture to describe the anatomical facial movements describing an expression. Additionally, an attention model is embedded within the network to ensure the focus is on the regions specific for the expression. For a given CNN classifier, authors in [Zhu 18] construct a GAN model to generate supplementary data for emotion classification. The authors

use CycleGAN [Zhu 17], a GAN model which learns translations between unpaired domains using the cycle-consistency loss for image-to-image transformation.

GANs have made a significant impact on affect estimation research, aiding in the generation of emotionally expressive content. These synthetic datasets can augment limited datasets used for training emotion recognition models, providing more diverse examples and improving model performance.

2.4.3 Level of Supervision

Fully Supervised

In fully supervised learning, the training data comprises labelled examples for every input-output pair used during the model's training phase. The algorithms aim to learn a mapping or relationship between the input features and the output labels provided in the training data. They learn to predict the correct output for new, unseen input data based on the patterns learned from the labelled examples. For affect estimation, the model learns to either classify the emotion category, or predict the continuous valence and arousal values corresponding to the input data.

Several studies described above have adopted a fully supervised approach for continuous and categorical affect estimation [Fasel 00, Dhall 12, Dhall 15, Kossaifi 17, Meng 17, Kollias 18a, Kossaifi 19, Kollias 20, Kollias 18b, Ma 23a]. However, fully supervised learning relies on having a substantial amount of accurately labelled data for training. Curating and annotating large-scale datasets for affect estimation is a costly, tedious, and time-consuming process. Moreover, domain experts, or individuals with a deep understanding of affect have to be involved in annotating expressions. Manual annotation is also a subjective process, often leading to ambiguous annotations. For instance, a FACS coder requires over 100 hours of training to achieve minimal proficiency, and scoring each video takes approximately an hour [Friesen 78]. Annotating for continuous emotion labels makes it challenging, as some emotional expressions might contain a blend of multiple emotions, further complicating the annotation process. Further, ensuring consistency across annotators in assigning continuous valence values is difficult due to the inherent subjectivity and varying perceptions of emotional intensity.

As a result of these considerations, researchers are exploring approaches with reduced supervision in affect inference to overcome the limitations associated with solely relying on full supervision, thereby aiming for more scalable, generalisable, and cost-effective solutions.

Semi-/Self-/Weak-/Unsupervised-based Methods

To address the limitations of fully supervised learning, researchers explore unsupervised, semi-supervised, and weakly supervised learning paradigms that aim to utilise unlabelled or weakly labelled data to complement the limited labelled data. These approaches attempt to mitigate the reliance on vast amounts of precisely labelled data while improving model performance and generalisation.

Semi-supervised learning is a machine learning paradigm that aims to utilise both labelled and unlabelled data for training models [Berthelot 19]. Semi-supervised learning (SSL) is beneficial when labelled data is scarce or expensive to obtain. It allows for leveraging the potentially vast amount of unlabelled data to enhance the model's generalisation and performance. A pseudo label-based SSL is proposed in [Choi 20b] for estimating continuous emotions in videos. A pre-trained LSTM network is used to generate features for both labelled and unlabelled data. Further, a CNN-LSTM regressor is trained with the pseudo-labels for estimating affect. In [Kim 17], the authors perform an adaptive fusion of the features from an image-based network, a facial-landmark network, and an audio network. A semi-supervised learning-based 3D autoencoder is incorporated in the image-based network for emotion inference. A novel semi-supervised framework with metric learning is proposed in [Tran 23] to adapt existing pre-trained encoders for affective downstream tasks such as continuous emotion recognition, affective highlight detection, etc.

Weakly-supervised learning refers to the use of imperfect, or less precise sources of supervision or labels to train machine learning models. Weak-supervision includes *incomplete supervision*, where only a subset of training data is equipped with labels, *inexact supervision*, where the training data has coarse-grained labels, and *inaccurate supervision*, where the labels need not be the ground-truth [Zhou 18]. A weak-supervision-based hybrid deep neural network and bidirectional LSTM (BLSTM) is employed for continuous affect estimation in [Pei 19],

where labels of context frames are considered to alleviate the negative effects of noisy or incorrect labels.

Unsupervised learning is a machine learning paradigm where the model is trained on unlabelled data without explicit guidance or labelled examples. The algorithm attempts to find patterns, structure, or inherent relationships within the data on its own, without being provided with predefined output labels or target values. Common techniques in unsupervised learning include *clustering*, where similar data points are grouped together based on some similarity measure. Unsupervised learning is valuable in scenarios where labelled data is scarce or unavailable, allowing models to explore and learn from raw data without human annotation. A domain adaptation technique is presented in [Zen 14] for personalised facial expression classification. Alleviating the need for labelled data, the proposed method relies on a regression framework to map the data points corresponding to a subject and the parameters of his/her classifier. In [Ji 23], the authors propose an unsupervised cross-domain facial expression recognition, which adapts an iterative pseudo-label assignment method to provide pseudo labels in the target domain. The significant emotion features are enhanced by employing a facial-landmark guided region-attention learning.

Unsupervised learning has not been used as commonly as supervised or semi-supervised methods in FAE. It still is in its infancy due to several challenges inherent in the complex nature of emotions and the limitations of unsupervised techniques. As a result, while unsupervised learning holds promise in processing data and exploring underlying structures, its direct application in accurate and robust emotion estimation remains in an early developmental stage, necessitating further research and advancements to overcome these inherent challenges.

2.4.4 Temporal Modelling

As hinted in Section 2.2, understanding the temporal dynamics of affect is crucial for a comprehensive analysis of affective states. Although the majority of the works for dimensional affect estimation model affect as a static entity, there is a formidable number of works in the literature on temporal modelling of affect. Since valence and arousal are time-varying signals, early methods have relied on short-term temporal correlations. For example, the authors

in [Baltrušaitis 13] use Continuous Conditional Random Fields on top of SVR for dimensional affect regression. In [Nicolaou 12], the authors predict time-continuous dimensional affect using the proposed Output-Associative Relevance Vector Machine (OA-RVM) regression framework by learning inherent spatio-temporal dependencies between output and input. Meng *et al.* [Meng 16] propose a two-stage system, where the first stage involves traditional regression methods to classify each individual video frame, and the second stage involves a proposed time-delay neural network (TDNN) to model temporal relationships between consecutive predictions.

The advancements in the deep learning research cascaded the advancements in the dimensional emotion research. For instance, Chao *et al.* [Chao 14] utilise deep belief network for dimensional affect estimation in two stages. In stage I, a temporal pooling function in the network is employed to encode time-continuous information in the features, and in stage II, the prediction results from various modalities and the emotion temporal context are combined simultaneously.

Owing to its end-to-end learning capability, multiple works employed CNN-RNN architectures for time-continuous dimensional affect estimation. Here, the spatial patterns are encoded using a CNN encoder and temporal patterns are modelled using a RNN architecture (for example, [Chen 15, Khorrami 16, Zhao 18]). In [Chen 17], authors compare the effectiveness of non-temporal SVR model and temporal LSTM model, and empirically show that temporal modelling is better than its non-temporal model for dimensional affect estimation. Kollias *et al.* [Kollias 21] propose multi-level CNN-RNN architecture. Low-, mid- and high-level features from CNN are extracted and passed as input to the RNN block. The authors show that multi-level CNN-RNN helped in boosting the network's performance for arousal estimation. A similar hierarchical learning framework is proposed in [Mao 19], where the authors propose three-stage hierarchical learning for predicting time-continuous dimensional affect. In the first stage, a raw input image is fed to a feed-forward neural network to generate high-level representation. Next, in the second stage, the BLSTM learns context information of the feature sequences from the previous stage. In the third and final stage, in an unsupervised way, a BLSTM network is used to correct the initial recognition result to obtain the refined and final prediction.

Differently, Aspandi *et al.* [Aspandi 21] use GAN to learn meaningful spatial representations and use *curriculum learning* to enable temporal modelling using LSTM.

The authors in [Lee 18] propose a spatio-temporal attention based neural network for time-continuous dimensional affect estimation. Convolutional LSTM (ConvLSTM) is used as a spatio-temporal encoder-decoder network and spatio-temporal attention is formulated for 3D-CNN. Further, similar to [Chao 14], Hu *et al.* [Hu 21] propose a two-stage spatio-temporal attention, where the first stage generates an initial recognition result, which is fed as an input into the second stage for correction.

Recently, a transformer-based encoder is used to capture the temporal context information, with fully connected layers as the prediction heads for valence and arousal prediction [Meng 22b]. Extending beyond self-attention, the authors in [Praveen 22] propose a joint cross-attention fusion model to exploit inter-modal relationships, however the visual temporal information is encoded with 3D-CNN.

Different from the above works, the authors in [Tellamekala 19] propose constrained representation learning, where the focus is on learning 'temporally coherent' latent features. This is achieved through a regularised contrastive loss based on the temporal coherency principle [Hurri 02], which states that when processing temporal input, the representation changes as little as possible over time. In [Kossaifi 20b], the authors learn temporal information through transduction of spatial information learnt via a tensor factorisation technique.

2.4.5 Benchmark Strategies

The benchmark strategy for the AffectNet and AFEW-VA datasets is proposed in [Toisoul 21], where a single network is used to detect facial landmarks, and perform continuous and categorical emotions. A face alignment network is used to obtain features for the input image. Further convolutional blocks are applied on these features to perform the affect estimation task. A stochastic temporal context modelling framework is proposed in [Tellamekala 22] to perform emotion inference. Affective processes are used to model the temporal dynamics via a probabilistic latent global variable. The commonly used metrics for affect estimation tasks are discussed in detail in Section 3.3.

2.5 Dimensional Affect Estimation from Other Modalities

Affect is a complex and multifaceted aspect of human experience. Since affect is inherently multimodal, besides facial expressions, affect spans across various channels such as vocal intonations, body language, textual cues, and physiological signals. This section provides a brief overview of studies performing affect estimation from modalities other than face. Although this section emphasises the multimodal nature of human affect, these aspects are beyond the scope, as the proposed methodologies in this thesis use facial cues for affect estimation.

2.5.1 Speech

Emotion is conveyed through voice in multiple ways, leveraging the various elements of speech such as, tone and pitch, speed of speech, changes in prosody, intensity, etc. As human affect is a continuum, automatic affect estimation systems aim at inferring continuous affect. Emotional history is modelled in [Wöllmer 08] using LSTM-RNNs to capture long-range dependencies in prosodic, spectral, and voice quality features of acoustic samples for inferring continuous emotions along valence, arousal, and time dimensions. In [Wöllmer 10], the authors examine an emotion inference system that is able to cope with spontaneous, non-prototypical, and unsegmented speech. They obtain acoustic features from the openEar [Eyben 09] feature extractor along with linguistic features, and train an LSTM for dimensional affect estimation.

For inferring affect from spontaneous speech, the authors in [Grimm 07] employ a 3-dimensional emotion-space comprising valence, arousal, and dominance. They segmented the speech signal into utterances, extracted features and estimated the three emotion primitives using a SVR. In [Parthasarathy 17], a deep neural network with multi-task learning is employed to jointly model valence, arousal and dominance, by extracting low level features, such as fundamental frequency and MFCC. Exploiting the individual advantages of support vector regression and bidirectional LSTM-RNN, a united cascaded model is proposed in [Han 17] for continuous spontaneous emotion inference. An end-to-end speech emotion inference network comprising CNN for extracting features from raw signals, followed by an LSTM for contextual modelling is employed in [Tzirakis 18].

2.5.2 Physiological Signals

Analysing physiological signals such as skin conductance measures, EEG, etc., can provide useful insights about affective states. Early studies have attempted to infer affect from physiological signals by treating it as a classification problem, as arousal vs non-arousal, valence vs non-valence [Gu 08]. An EEG-based emotion inference system is proposed in [Khosrowabadi 10], where a self-organising map is used to identify the boundaries between separable regions of valence and arousal dimensions. Using an end-to-end network comprising of a series of convolutional and RNNs, the authors in [Keren 17] predict valence and arousal levels from ECG and EDA. A decision-level fusion of the EEG signals and facial expressions, recorded while watching movie clips with positive, negative, and neutral emotions, is performed using LSTM in [Li 19] for continuous emotion estimation.

A decision tree algorithm is employed in [Frantzidis 10] for discriminating emotional physiological signals evoked while viewing affective pictures. The recorded biosignals were initially classified along the valence dimension, and later classified along the arousal dimension with the additional gender information. One of the first works to employ DL for affect inference using physiological signals is proposed in [Martínez 13], where emotion is inferred using skin conductance and blood volume pulse individually, as well as by employing fusion. A self-supervised approach for wearable emotion inference based on physiological signals is proposed in [Wu 23]. Firstly, using a pre-trained model, unlabelled multimodal data are assigned labels through a series of transformations. Convolution-based encoders are employed for feature extraction, followed by aggregation of the features using transformer-based encoder for emotion inference task.

2.5.3 Body Gesture

Gesture is a crucial non-verbal channel for emotional expression. It can include the movement of hands, head, and other parts of the body. In [Castellano 07], the authors extracted time-domain features by calculating motion cues such as contraction index of the body, velocity, acceleration, and fluidity of the hand's *barycentre*, and employed a nearest neighbour algorithm,

decision tree, and Bayesian network for affect inference. A statistical mapping between continuous emotional states and behavioural features is presented in [Metallinou 11] using a Gaussian Mixture Model. The authors analyse the emotional content of body language cues, describing posture, relative orientation, and approach/withdrawal behaviour in affective interactions.

A computational framework for inferring affect from body movements is proposed in [Piana 16], where 3D motion data of full body movements is obtained from motion-capture systems. Histogram-based representation of movement features are computed and emotions are inferred via a linear SVM. For modelling gesture dynamics in interpersonal interactions in [Yang 16], gesture variability is represented using Gaussian Mixture Models, and gesture sequences are modelled using Hidden Markov Models. A generalised zero-shot learning framework is employed in [Wu 20] to infer affect. A bidirectional LSTM is used for feature extraction, followed by a three-branch network employing multi-task learning, where two tasks are body gesture classification, and the third branch is emotion classification.

2.5.4 Text

Textual analysis involves determining the emotional tone expressed in written communication. The choice of words, expressions, and linguistic styles can convey emotions. The authors in [Tang 14] use a supervised learning framework for learning continuous word representations as features for sentiment classification of tweets. The context information of words is modelled to distinguish words with opposite polarity, but not those with similar context. A bag-of-words linear regression model is employed in [Preoţiuc-Pietro 16] for valence and arousal inference from social media posts. Emo2vec, a multi-task learning framework for word-level representations to encode emotional semantics with fixed-size vectors, is proposed in [Xu 18] to learn a generalised emotion representation for various affect inference task. A multi-task ensemble model, leveraging learned representations of a CNN, LSTM, and a Gated Recurrent Unit (GRU), is employed in [Akhtar 19] for valence and arousal inference, emotion classification, and sentiment classification across various textual data such as tweets, Facebook posts, news headlines, blogs, letters, etc.

2.5.5 Multimodal

Emotion is naturally conveyed through a variety of *multimodal* cues, where multimodal refers to the integration of multiple sources or modalities of information to comprehensively understand and analyse affective states. The goal of multimodal affective computing is to leverage the strengths of each modality, compensating for individual modality limitations and enriching the accuracy and depth of emotion recognition. It involves (a) extracting relevant features from each modality, and (b) the fusion, or combining and integrating information from different modalities to create a unified representation of affective states. Multimodal learning compensates for the limitations of individual modalities, allowing a richer interpretation of emotions.

Two important factors to be considered while integrating the modalities are *when* to integrate (i.e., at what level of abstraction) and *how* to integrate (i.e., which criteria to use). Typically, fusion is done either at the *feature level* or at the *decision level*. Feature-level fusion involves combining features extracted from multiple modalities at a lower level, before feeding them into a machine learning model for analysis. On the other hand, decision-level fusion involves combining the decisions or outputs obtained from individual models trained on different modalities to make a final decision or prediction. A *hybrid* fusion could involve both feature-level fusion and decision-level fusion.

Combining audio-visual modalities involves integrating information from both auditory cues (for example, MFCC, Beat Histogram, etc.) and visual cues (for example, facial expressions, gestures, etc.) for affect inference [Gunes 10, Nicolaou 10, Wei 20, Praveen 21, Praveen 22, Meng 22b, Jeong 22, Praveen 23]. Studies have also explored a fusion of processed EEG signals with acoustic features, such as MFCC [Verma 14, Li 21], or a fusion of EEG signals and visual cues, such as facial expressions [Soleymani 14, Huang 16, Li 19] for multimodal emotion inference. Going beyond two modalities, emotion inference is also examined by performing a multimodal fusion of visual, audio and EEG signals [Xing 19, Nakisa 20, Choi 20a, Pan 23], text, audio and video information [Poria 16a, Poria 16b, Siriwardhana 20], face, body gesture, voice and physiological signals [Jessen 11, Ranganathan 16].

Although multiple studies establish that multimodal systems are superior over their uni-

modal counterparts [D'Mello 12, D'mello 15], multimodal affect estimation poses several challenges. The integration of diverse modalities poses difficulties in data synchronisation and preprocessing, leading to potential misinterpretations of affective cues. Translating representations between modalities is complicated by the inherent heterogeneity and subjective relationships [Baltrušaitis 19].

2.6 Summary and Research Gaps

In a nutshell, computational research on human affect is progressively rising. The various affect representation models proposed by psychologists, and studies comprehending affect from diverse cues, have played a foundational role in the development of ML/DL frameworks for automatic affect inference. Substantial research on facial affect inference can also be attributed to the development of innovative computer vision algorithms and the curation of affective databases. While early studies used classical ML algorithms with handcrafted features of the data, modern studies have leveraged state-of-the-art DL architectures for unimodal and multimodal affect inference.

The majority of the studies aiming at affect inference focus on classifying the discrete emotions. While numerous studies have attained exceptional levels of accuracy in this regard, to develop robust affect-aware machines, it is important to achieve further progress with respect to dimensional affect inference. The dimensional representation acknowledges the continuous and dynamic nature of emotions, and is more tractable than categories. However, addressing affect inference as a regression problem is no simple feat.

Data, being the foremost element for building automatic affect inference frameworks, is equally the most challenging aspect. The number of affective databases with categorical affect annotations is considerably higher than those with dimensional ones. Even among the databases with dimensional annotations, incorporating elements such as in-the-wild settings and spontaneous data, further limits the count. To the best of the author's knowledge, only two affective in-the-wild **video** databases with dimensional annotations exist currently (AFEW-VA and Aff-Wild2). As prior studies have reported superior performance with temporal data, this thesis

focuses on modelling temporal information for facial affect estimation.

As described in Section 2.1, some studies point out that universal facial expressions exist corresponding to the emotion categories. However, other studies also report that personality traits, contextual factors, etc., shape how individuals express emotions. Given the equitable rationales on the facial expressions of subjects, it is essential to consider both subject-independent and subject-dependent settings while computationally inferring affect. Since affective video databases are limited, the ML/DL models might be prone to learning identity-specific cues. Consequently, an appropriate split of the data into training and test sets contributes to affect inference. Additionally, the majority of studies have focused on the SI setting, while hardly any studies have inferred affect from a SD perspective. In this regard, Chapter 4 explores time-continuous dimensional affect inference considering both, subject-independent and subject-dependent settings.

While traditional ML algorithms offer numerous advantages, they heavily rely on hand-crafted features. DL algorithms outperform ML algorithms, as they can automatically learn hierarchical representations of data without manual feature engineering. Specifically, for time-continuous input, traditional ML methods are constrained as they cannot effectively model complex, non-linear relationships in time series data. However, DL methods are capable of modelling the intricate temporal dependencies more effectively. While DL methods perform better with large amounts of data, aggregation of large data is challenging. DL methods require large amounts of data to effectively learn and generalise without overfitting. Thus, the limited available video data poses a challenge for automatic affect inference. As a plausible solution, either massive datasets must be curated, or frameworks, which can facilitate affect inference from limited data, must be developed. While creating affective databases is equally essential for affect inference, it lies beyond the scope of this thesis. Thus, Chapters 5, 6 and 7 focus on developing frameworks based on weak-supervision for affect inference from limited video data.

One of the drawbacks of the current methods in the literature is a lack of analysis on the generalisability of the methods proposed in the studies. Generalisability refers to the ability of the models to yield reliable and valid results when applied in different settings and unseen data. Thus, an affect inference system that is capable of performing perfectly in one condition, needs

to be tested on a different condition to judge its generalisation capability. In Chapter 8, the generalisation capabilities of the proposed time-continuous affect estimation system is summarised, which elucidates the relevance of the methodology adopted in this thesis.

Chapter 3

System Design and Datasets

Contents

Designing a system for affect inference involves considering various factors, including the nature of the input data, the desired output, the available computational resources, and the deployment environment. The typical phases include data collection and preprocessing, feature extraction, handling multimodal information if multiple modalities are involved, selecting appropriate model architecture, training and optimising the model, and evaluating the model using the suitable metrics. This chapter discusses the system design adopted in this thesis, the datasets used for affect inference, and evaluation metrics.

3.1 System Design and Analysis

3.1.1 End-to-end Learning

Early studies which employed classical ML algorithms required engineers and domain experts to use their knowledge to manually design and extract relevant features from raw data. Machine

learning models were trained using these engineered features as inputs. However, manual feature engineering is time-consuming and potentially limits the performance of models based on the features selected. The evolution towards *end-to-end learning* was driven to automate the feature engineering process and enable models to learn more abstract and complex representations directly from raw data.

End-to-end learning refers to an approach where a system learns directly from raw or minimally processed data to produce a desired output, without relying on handcrafted intermediate representations or stages. This approach is advantageous as it can streamline the learning process, potentially reducing the need for feature engineering and manual intervention. Further, it can facilitate transfer learning or domain adaptation by allowing models trained on one dataset or domain to be fine-tuned or adapted to perform well on a different but related dataset. However, while end-to-end learning has its advantages, it can also pose challenges, especially when the data is limited or when interpretability of the learned representations is crucial. Additionally, in some cases, breaking down the learning process into distinct stages might provide better interpretability or control over the system's behaviour.

3.1.2 Encoder-Decoder Architecture

The encoder-decoder architecture is a powerful framework that employs end-to-end learning strategy. An *encoder* is a component or a specific type of neural network architecture that transforms input data into a different representation, often referred to as the *latent space*, through a series of computational layers. These representations or features are typically higher-level abstractions compared to the raw input. Encoders can reduce the dimensionality of the input data by transforming it into a lower-dimensional representation, which often captures the most salient information while discarding less relevant details. For instance, the latent space could be a 2D or 3D space despite the input data having much higher dimensions. The various types of encoders are, (a) *CNNs*, designed for processing grid-like data, such as images or videos, (b) *RNNs*, suited for sequential data processing to capture temporal dependencies, (c) *variational autoencoders*, which focus on learning probabilistic representations and generating new data samples, (d) *transformers*, which efficiently process sequential data by attending to different

parts of the input sequence simultaneously, (e) *pre-trained encoders*, used to learn representations from vast amounts of data and transfer these learned features to downstream tasks.

Decoders are components or architectures that complement encoders in transforming latent space representations back into an interpretable form. Decoders aim to reconstruct meaningful output from the learned representations provided by encoders. The different types of decoders are, (a) *linear layers*, which can be used to map the representation to the output space, (b) *autoencoder decoders*, which is used to reconstruct the input data from the compressed latent space representation learned by the encoder, (c) *GAN decoders*, which aim at generating new data samples, such as images or text, from learned representations, (d) *variational autoencoder decoders*, which reconstruct input data and generate new samples by sampling from the learned latent space distribution.

In the context of affect inference, the encoder processes information from various modalities, and the decoder integrates these representations for emotion inference. For affect inference using time-continuous input, the goal is to effectively capturing both spatial and temporal features. 3D CNNs naturally incorporate temporal context, which is crucial for understanding the dynamic nature of affect. Likewise, RNNs such as LSTMs, are widely used to process time-series data. However, non-continuous inputs are typically processed using 2D CNNs, which are capable of capturing local spatial dependencies, but struggle with long-range dependencies unless coupled with other techniques (for example, stacking CNNs with RNNs). The various encoders and decoders employed in this thesis are explained in the subsequent chapters.

3.1.3 Optimisation and Loss function

Weights and biases collectively determine the behaviour and predictive power of a neural network. *Weights*, which represent the strength of connections between neurons in different layers, control the flow of information through the network by scaling input values at each connection. *Biases*, the parameters added to each neuron in a layer, allow the network to fit more complex functions. Weights are initially assigned random values, and biases are often initialised to small constants or zeros. They undergo adjustments during training to optimise the network's ability to learn and generalise from the data.

Optimisation, a core component within the training phase of a neural network pipeline, determines how effectively the model learns from data. It refers to the process of adjusting the model's parameters (weights and biases) to minimise or maximise a specific *objective* or *loss function*. The *objective function* quantifies the discrepancy as a scalar values between the predicted output of the model and the actual ground truth for a given set of input data. The model modifies these hundreds of millions of internal adjustable parameters to reduce the error using backpropagation and optimisation algorithms. The weight is adjusted in the direction that minimises the loss function, scaled by the learning rate. Different optimisation algorithms offer various advantages and trade-offs in terms of convergence speed, memory usage, and suitability for different types of architectures. For instance, (a) *gradient descent (GD)* updates weights by moving in the opposite direction of the gradients with a fixed learning rate, (b) *stochastic gradient descent (SGD)* computes gradients and updates weights for each training example, and *stochastic* refers to the probabilistic nature of the algorithm. Unlike traditional GD, which computes gradients and updates weights using the entire dataset (batch) in each iteration, SGD operates on individual samples or small batches of data randomly selected from the training set, (c) *mini-batch gradient descent*, where weights are updated based on gradients computed from smaller batches of data, balancing the benefits of SGD and full-batch GD, (d) *Adam* optimiser adapts the learning rates for each parameter based on the average of past gradients, etc.

During the training process, the loss function guides the optimisation algorithm to adjust the model's parameters in a way that minimises the error. In tasks like regression and classification, the loss function measures how well the model's predictions align with the true labels in the training data. The choice of the loss function influences the model's behaviour during training, affecting its ability to learn meaningful patterns from the data. The loss functions employed in the pipelines are explained in the respective chapters.

3.2 Datasets

Publicly available datasets are considered to ensure reproducibility, transparency, and to have a standard benchmark for the research in this thesis. Further, considering the advantages of *in-*

the-wild dataset for a robust affect estimation system, as reasoned in Section 2.3.1, this section details relevant datasets used in this thesis.

3.2.1 AFEW-VA

Background and goal

AFEW-VA dataset [Kossaifi 17] is derived from AFEW dataset [Dhall 12]. In order to facilitate time-continuous dimensional affect estimation research, authors of [Kossaifi 17] curated AFEW-VA. At the time of its release, this is one of the first databases with time-continuous valence and arousal labelled in-the-wild database.

Collection and annotation procedure

Since the samples are from AFEW dataset, the collection procedure from [Dhall 12] are detailed. The construction of the database is a semi-automatic approach, where a *subtitle* is extracted from a movie DVD and then is parsed in a recommender system. Subtitles contain audio and non-audio context such as *[CHEERING], [SHOUTS]*. A recommender systems plays audio-visual part of the movie based on the subtitle keyword. Then, human labellers chose 1426 *clips* based on criteria such as visible presence of subjects, display of expressions.

For AFEW-VA dataset, authors in [Kossaifi 17] chose a subset of 600 clips. Frames are extracted from the clips, where two expert annotators (a male and a female) labelled valence and arousal values for each of the frames. The valence and arousal values are in the range of -10 to 10 integer values, resulting in a total of 21 *classes*. Ten percent of the clips were randomly selected and re-rated by the same two annotators. The inter-rater Pearson product-moment correlation coefficient is 0.87. The database is publicly available for download[1].

Data attributes

The dataset consists of 600 video clips. The length of clips varies from 10 frame to 145 frames, with ≈30,000 frames in total. The dataset comprises of 240 subjects (actors), with no infor-

[1]Web link: https://ibug.doc.ic.ac.uk/resources/afew-va-database/

mation provided on gender distribution. The age range is 1-70. The clips represents real-world conditions, which are challenging for affect estimation tasks.

Data formatting

The dataset providers released 68 landmarks including both interior and boundary points of the face for each frame of the videos. The landmarks were annotated using a semi-automatic approach, where, in the first stage a face is detected using tree-based deformable part model (DPM), and in the next stage, the bounding box of the detected face was subsequently used to initialise the Gauss-Newton generative part-based model for facial landmarks localisation. Thus, the publicly released dataset contains frame of videos, facial landmarks with valence and arousal annotation.

3.2.2 AffectNet

Background and goal

AffectNet [Mollahosseini 19] is an image-based database curated by querying emotion-related keywords from web search engines. The goal of this dataset was to create a large-scale (in the order of 500,000) *in-the-wild* image dataset for dimensional affect research.

Collection and annotation procedure

Emotion-related keyword were combined with descriptors for age, gender or ethnicity. A total of 362 English language keywords, such as “joyful girl”, “furious young lady”, “astonished senior”, were generated. The keywords were also translated to other languages from non-English proficient native speakers (well-versed in English) of the respective language. This resulted in a total of 1250 keywords, which are then used for crawling three major search engines (Google, Bing and Yahoo!) for retrieving images.

The annotations are done by 12 full-time and part-time annotators at the University of Denver. Each image was annotated by one annotator. Each image is annotated by one of the eleven discrete categories, namely Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None (“none

of the eight emotions"), Uncertain, and Non-face. For dimensional affect annotation, annotators were required to mark both valence and arousal values ($\in$ [-1, 1]) on a 2D Cartesian co-ordinate circumplex model.

To calculate the annotators' agreement, a subset of 36,000 images were annotated by two annotators. For categorical labels, annotators agreed on 60.7 percent of the total images. For dimensional labels, a Pearson's correlation coefficient (PCC) of 0.815 and 0.567 was observed for valence and arousal, respectively.

Data attributes

After crawling three search engines, $\approx 1,00,000$ images containing at least one face is kept as part of the dataset. However, a subset of 450,000 images were annotated. The average image resolution is 425 $\times$ 425, with standard deviation 349×349 pixels. The data attributes are extracted from Microsoft Cognitive Face API[2], which reported 49% of the faces as male. The average estimated age is 33.01 $\pm$ 16.96 years. Additionally, 9.63% of the faces wear glasses, 51.07% of the faces have eye make-up, and 41.4% of the faces have lip make-ups.

Data formatting

No data formatting techniques were used. The samples of the dataset are the raw RGB images from the internet which contain a face. For fair evaluation, the subset of the annotated images that are annotated by two annotators is termed as *test* set and is not available as part of the public dataset. *Validation* set comprises of five hundred samples of each category which are randomly selected[3]. For all our experiments, the original *validation* set is considered as the *test* set. The rest of the images are tagged as *training* samples.

[2]At the time of writing this thesis, Microsoft Cognitive Face API retired face attributes estimation such as gender, age, smile, as a measure to "mitigate potential misuse". More on this can be found here: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/concept-face-detection

[3]The dataset is available on request. Project page: http://mohammadmahoor.com/affectnet/

3.2.3 Aff-Wild2

Background and goal

Aff-Wild2 [Kollias 19b], an extended version of Aff-Wild [Zafeiriou 17], is a dataset comprising of videos curated from YouTube, a video sharing website. The samples in the dataset consists annotation of a dimensional and categorical affect. The goal of this dataset is to create a large-scale audiovisual dataset, with time-continuous dimensional affect labels.

Collection and annotation procedure

The videos for Aff-Wild is retrieved based on the keyword "reaction" from YouTube. the videos would then contain subjects reacting to a variety of topics (reviewing movie, products, etc.). However, for the extended Aff-Wild2 database, the keywords were from the 2D Circumplex model (see Fig 1.2). As a result, the videos are subjects who react on a surprise, express happiness, flirting or rejection, discuss political issues, etc.

For Aff-Wild, the dataset is annotated by six experts for valence and arousal values in a time-continuous manner. For each video, cross-correlations between annotators were computed. Two additional experts reviewed all videos, selecting 2 to 4 annotations per video with the highest correlations. The mean of these selected annotations is released as a "ground-truth" valence and arousal values. A similar approach is employed for the creation of the extended Aff-Wild2 dataset. The inter-annotation agreement of 0.63 and 0.60 are reported for valence and arousal, respectively. Additionally, Aff-Wild2 contains a subset of videos with "expression" label, annotated by three experts. Only the annotations where all experts agree are kept.

Data attributes

Aff-Wild2 contains a total of 558 videos (298 videos were part of Aff-Wild), with 2,786,201 frames. The videos show subtle and extreme human behaviours in real-world settings. The length of the videos vary from 0.03 to 26.22 minutes. The dataset contains 458 subjects, with 279 males and 179 females. The videos have wide range of subjects in terms of age (babies to elderly people), ethnicity (Caucasian, Asian, Black, African American, etc.), profession (politi-

cians, actors, athletes, etc.), head pose, illumination and occlusion [Kollias 19b].

For a subset of videos with categorical expression label, the total length of video duration is 3 hours and 45 minutes. This subset includes 84 videos, with 403,758 frames and 84 subjects (43 male and 43 female).

Data formatting

After extracting raw videos from YouTube, faces are extracted using [Mathias 14]. In order to match annotation time and cropped face time instance, each frame's time is the *nearest neighbour* annotation time stamp with valence and arousal annotation values. For a fair comparison across studies, pre-determined *training*, *validation*, and *test* sets are provided[4]. However, *test* set is not part of the public dataset. Thus, for all the experiments, the original *validation* set is considered as the *test* set, and the rest of the images as *training* samples.

3.3 Performance Evaluation

Evaluation metrics are important to measure the reliability of automated affective predictors and annotations. Various metrics like accuracy, F1-score, Area Under the ROC Curve (AUC), etc. are used for evaluating categorical models. As explained earlier, one of the aims of this thesis is to effectively model the temporal information to estimate continuous valence and arousal values. Given the continuous nature of affect labels, the above mentioned metrics are not considered appropriate as evaluators. Hence, in the following, several metrics that are appropriate and tested for the evaluation of the dimensional model are reviewed.

For the task of estimating valence and arousal, given the ground-truth and the prediction, performance is usually measured using the root mean squared error (RMSE), given by:

$$\text{RMSE}(\hat{\theta}, \theta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}(i) - \theta(i)\right)^2} \tag{3.1}$$

where θ and $\hat{\theta}$ are a series of n ground-truth labels, and the corresponding predicted labels,

[4]The dataset is available on request. Project page: https://ibug.doc.ic.ac.uk/resources/aff-wild2/

respectively. RMSE provides a single number that summarizes the magnitude of prediction errors. A lower RMSE indicates a better fit of the model to the data. We note that comparison of RMSE of different datasets with different scales would be inappropriate as the measure is dependent on the range of θ. However, in such scenarios, normalising the RMSE facilitates the cross-dataset comparison with different scales. Though there is no consistent means of normalisation, typically *min-max*, mean, or inter-quartile range of the data is used as a normalising factor.

Another metric typically used as a performance measure in affective databases is Pearson Correlation Coefficient. PCC measures how correlated prediction and target variables are. PCC is defined as:

$$\text{PCC}(\hat{\theta},\theta) = \frac{\mathbb{E}\left[(\hat{\theta}-\mu_{\hat{\theta}})(\theta-\mu_{\theta})\right]}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{COV(\hat{\theta},\theta)}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{3.2}$$

where μ_{θ} and σ_{θ} correspond to the mean and standard deviation of θ (correspondingly for $\hat{\theta}$ as well). By the definition, PCC ranges from -1 to +1. Although PCC reflects the strength and direction of a linear relationship, it fails to capture the slope of the relationship. This means, the PCC performance of the affect recognition can be +1, even without $\theta = \hat{\theta}$. More recently, the Concordance Correlation Coefficient (CCC) has been used for facial expression estimation [Kollias 19a, Toisoul 21]. The CCC is defined as:

$$\text{CCC}(\hat{\theta},\theta) = \frac{2\sigma_{\hat{\theta}}\sigma_{\theta}\text{PCC}(\hat{\theta},\theta)}{\sigma_{\hat{\theta}}^2+\sigma_{\theta}^2+(\mu_{\hat{\theta}}-\mu_{\theta})^2} \tag{3.3}$$

CCC is a measure of how well $\hat{\theta}$ is compared to the ground-truth θ. As seen above, CCC incorporates the PCC value but penalises correlated predictions with different means. This means, even for the predictions that are well correlated (from PCC) but with far from the ground-truth value, CCC penalises with proportion to the deviation.

Since valence and arousal values are generally annotated in the range of -1 to +1 (or can be normalised to [-1, +1] if not in the range), and the signs reveal the broader emotional state, Sign Agreement Metric (SAGR) is used to evaluate the predictions of valence and arousal. Unlike other applications, signs of valence and arousal hold significant information. For example, for a facial expression with ground-truth valence value +0.2, prediction of +0.5 is better than a

prediction of -0.1, since the former prediction corresponds to the same *class* (positive valence). Hence, SAGR evaluates whether the prediction and ground-truth matches the *class*. SAGR is defined as:

$$\text{SAGR}(\hat{\theta},\theta) = \frac{1}{n}\sum_{i=1}^{n}\delta(\text{sign}(\hat{\theta}(i)),\text{sign}(\theta(i))) \tag{3.4}$$

where δ is the Kronecker delta function, defined as

$$\delta(x,y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases} \tag{3.5}$$

Apart from the above mentioned metrics, the authors in [Kossaifi 17] use the intra-class correlation coefficient ICC(3,1). For two samples ($\hat{\theta}$ and θ for our case), the ICC is defined as:

$$\text{ICC}(\hat{\theta},\theta) = \frac{2\times\mathbb{E}\left[(\hat{\theta}-\mu_{\hat{\theta}})(\theta-\mu_{\theta})\right]}{\sigma_{\hat{\theta}}^2+\sigma_{\theta}^2} = \frac{2\times COV(\hat{\theta},\theta)}{\sigma_{\hat{\theta}}^2+\sigma_{\theta}^2} \tag{3.6}$$

Previously, ICC has been used as an evaluation metric in facial expression and pain estimation analysis [Mavadati 13, Kaltwang 16].

Chapter 4

Affect inference from limited data

Contents

In the previous two chapters, a comprehensive literature review and overall system design employed in this thesis is discussed. This chapter builds on the research gaps discussed earlier in the domain of time-continuous dimensional affect estimation. In the hindsight, this chapter serves as an exploratory analysis for subsequent chapters, specifically on designing an affect system when data is sample-limited, influence of subject information in automatic estimation of affect, and a need for loss function specifically for dimensional affect estimation task[1].

[1]The findings of this chapter is published at the International Conference on Automatic Face and Gesture Recognition (FG) 2023. Details of [Parameshwara 23b]: Ravikiran Parameshwara, Ibrahim Radwan, Ramanathan Subramanian, and Roland Goecke, *Examining Subject-Dependent and Subject-Independent Human Affect Inference from Limited Video Data*. In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG) (pp. 1-6). IEEE, 2023.

4.1 Introduction

Emotional expressions evolve dynamically [Ekman 69], making it essential to model short and long-range dependencies among emotional expression features observed over a given time interval. A time-continuous emotional space not only describes complex emotional states, but naturally enables the representation by a temporal model [Metallinou 13]. Multiple studies have employed deep neural networks to learn spatio-temporal dependencies for emotion inference [Kollias 18b, Kossaifi 20b, Tellamekala 19, Shukla 22]. Thus, the continuous model is more representative of emotions as compared to the categorical counterpart, as it can lead to an accurate assessment of the natural affective state, as evoked emotions are often mixed, complex, subtle and ambiguous in real-world scenarios [Gunes 13].

Affect recognition systems typically use affective data captured in controlled settings [Pantic 05, Abadi 13], but recent studies have focused on recognising 'in-the-wild' emotional expressions captured under naturalistic settings [Kossaifi 17, Shukla 17]. Although multiple affective video databases exist [Kossaifi 17, Ringeval 13, Kollias 18a], they are typically limited by the amounts of annotated data, unlike affective image databases that contain millions of samples [Mollahosseini 19]. The scarcity of large corpora of affective annotated video data can be attributed to (a) the difficulty in capturing emotional data under naturalistic conditions and, (b) the difficulty in assigning static and dynamic emotion labels to large amounts of data [Gunes 13]. In addition, limited video data present a challenge as emotional patterns are to be learnt spatio-temporally, unlike image data where spatial emotion representations need to be learnt.

Early studies on basic emotions state the existence of a core facial configuration reflecting the emotional state of a person [Ekman 11]. In contrast, other scientific frameworks posit that expressions of the same emotion vary substantially across individuals and situations [Barrett 19]. For example, the typical expression of anger (eyebrows furrowed, eyes wide, lips tightened) might sometimes be accompanied with additional facial movements such as a widened mouth, while in other instances, a facial movement might be missing with respect to the prototype. Such variations are considered to be a meaningful part of an emotional expression,

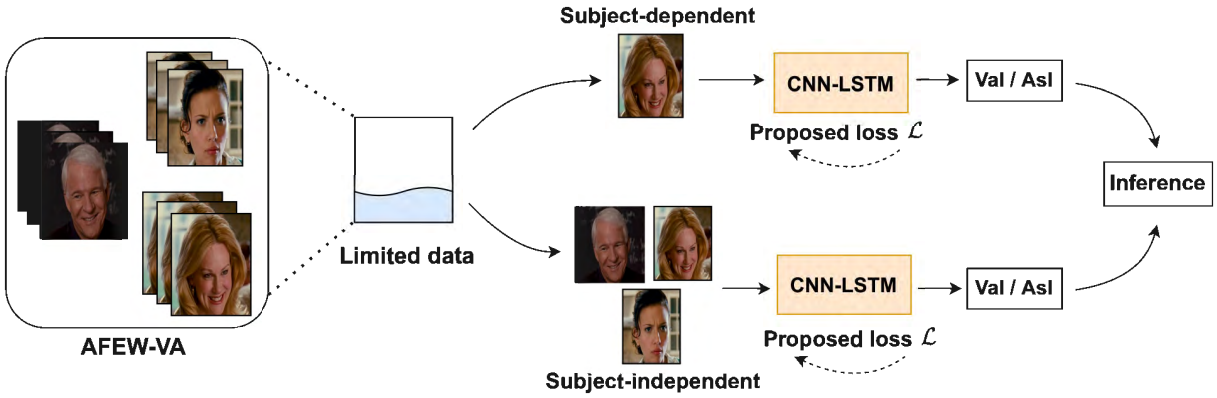


Figure 4.1: Approach overview depicting continuous/dynamic valence and arousal score prediction with limited data in the AFEW-VA [Kossaifi 17] dataset. The proposed network and loss function are evaluated in subject-dependent and subject-independent settings.

because facial movements are functionally tied to other factors such as external context and the person's internal affective state. Hence, while inferring affect computationally, it becomes essential to consider models, which are both subject-specific and subject-agnostic. Specifically, limited data is a serious impediment in learning emotion-specific facial expressions and it is, therefore, likely that machine learning and deep learning algorithms learn identity-specific characteristics for decoding observed expressions of emotions. Consequently, how training and test data are divided plays a vital role [Hajarolasvadi 21, Scheurer 20] in determining recognition performance. A comparison of subject-specific vs subject-agnostic settings is, therefore, critical in lean data settings. Both *subject-dependent* and *subject-independent* settings are employed to infer valence and arousal scores on the AFEW-VA [Kossaifi 17], an in-the-wild video dataset with limited data. The SI setting involves the use of training and test sets with mutually exclusive subjects, while training and test samples corresponding to the same subject occur in the SD setting (see Figure 4.1).

4.2 Key Contributions

The main contributions of this chapter are as follows:

1. The SI and SD settings are examined for valence and arousal inference on the sample-limited AFEW-VA dataset. Given (a) the small size of AFEW-VA dataset, and (b) the

inverse-exponential (e^{-x}) distribution observed for the number of samples (video snippets) available per subject (see Figure 4.3), vastly different emotion inference performance in the SI and SD settings are noted.

2. While both the SI and SD settings involve mutually exclusive training and test sets, these sets also involve *mutually exclusive subjects* in the SI setting as mentioned above. All performance metrics considered here substantially improve in the SD setting as compared to the SI setting. These results reveal that learning individual encoding is critical for accurate arousal and valence recognition on AFEW-VA.

3. A novel dynamically-weighted loss function is proposed to simultaneously improve the correlation as well as minimise the error between the target values and predicted *valence* and *arousal* values via the CNN-LSTM network depicted in Figure 4.4.

4.3 Experiments

In this chapter, AFEW-VA dataset (Section 3.2.1) is used for all the experiments. This section describes the pre-processing step of face extraction, our CNN-LSTM architecture, and the proposed loss function.

4.3.1 Methods

Pre-processing

As an initial step, faces are extracted from each frame in the AFEW-VA videos. Given an input video, Multitask Cascaded Convolutional Neural Networks (MTCNN) [Zhang 16b] is employed, which is a unified framework for both face detection and face alignment. Sometimes, face detection algorithms are prone to fail while detecting faces in-the-wild. In such a case, the Contrast Limited Adaptive Histogram Equalization (CLAHE) [Reza 04] technique is employed to enhance image contrast. The output of CLAHE is passed to MTCNN for face detection. If the face is still not detected, the bounding box of the neighbouring (preceding or succeeding) frame is positioned on the current frame, given that the face location differences in neighbouring

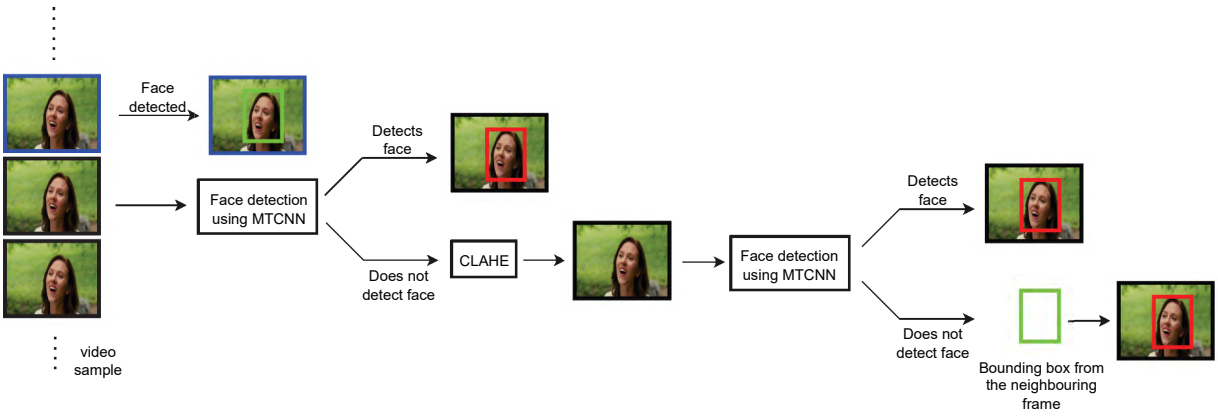


Figure 4.2: The illustration of the face detection and cropping pipeline, which acts a pre-processing step to the proposed models.

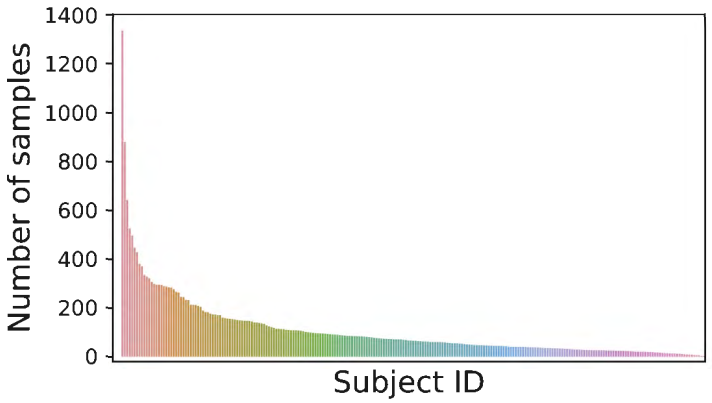


Figure 4.3: Distribution of the number of input samples per subject in the AFEW-VA dataset.

frames are negligible. Rather than discarding a frame when the face is not detected via MTCNN, the inclusion of almost all frames in the AFEW-VA is ensured, which is relatively small to begin with.

In each video, a sequence (snippet) of eight consecutive frames [Aspandi 21] with a stride of 1 is considered as an input sample. The total number of derived samples from the dataset is $25,759$, with the input dimensionality of each sample being $8 \times 3 \times 128 \times 128$, *i.e.*, each input sample (or video snippet) has eight frames of size $3 \times 128 \times 128$. Input samples and affect labels are normalised to the $[0, 1]$ and $[-1, 1]$ range, respectively, before feeding them to a CNN-LSTM network. Figure 4.3 shows the distribution of the number of video snippets per subject.

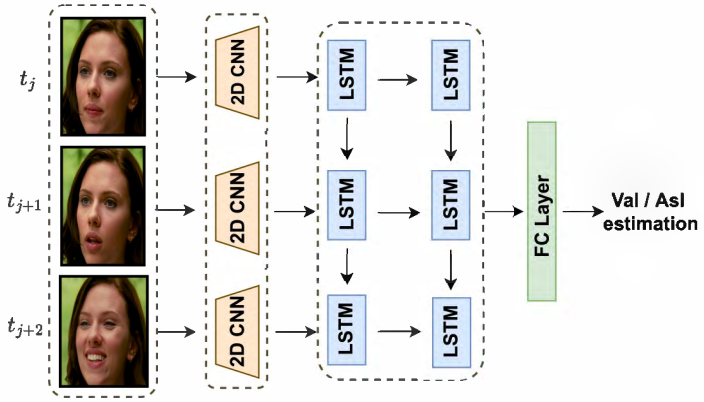


Figure 4.4: Architecture of our CNN-LSTM network for a 3-frame video snippet. t_j, t_{j+1}, and t_{j+2} denote the time steps corresponding to the three frames.

Architecture

The architecture chosen for this study is illustrated in Figure 4.4. A CNN-LSTM network is used, in which spatio-temporal patterns are learned using a 2D-CNN network for each frame followed by LSTM layers. For the CNN architecture to learn spatial patterns, ResNet-18 and ResNet-50 [He 16] architectures are chosen, however, the experimental results are reported for the ResNet-18 architecture, as obtained results were fairly similar in both cases. The final classification layer of ResNet-18 is replaced with a linear layer with 300 neurons (this number was empirically found to be optimal). The outputs of the respective CNN networks are fed as input to an LSTM layer, followed by another LSTM layer, both with 256 units. This is followed by a linear layer with 128 neurons and a final regression layer for estimating valence or arousal scores in $[-1, 1]$.

Loss function

Similar to past studies examining dimensional emotion estimation [Kossaifi 17, Kollias 19a, Toisoul 21], the performance metrics used here are RMSE, PCC, and CCC (see Section 3.3).

In dimensional affect inference, the aim is to minimise RMSE, while simultaneously maximising PCC and CCC. The most common approach employed for regression model optimisation is to use individual loss functions, namely, Mean Squared Error (MSE) or inverse-CCC [Kossaifi 19, Hu 21, Kollias 19a]. Other studies also use a combination of losses in addition to using the losses individually [Kossaifi 20b, Toisoul 21]. For example, the authors

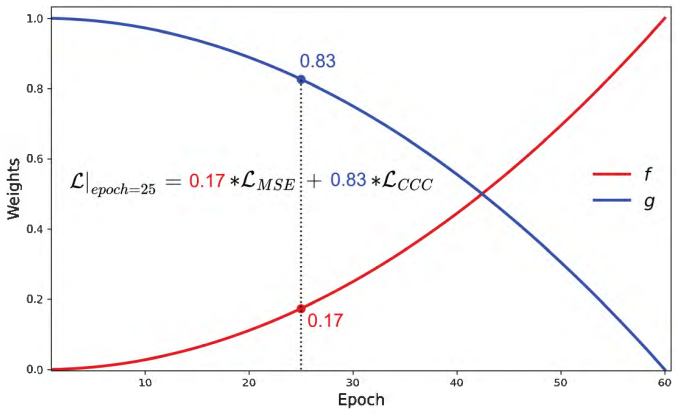


Figure 4.5: Illustration of the dynamic weight functions f and g used in Eq. 4.1 with $k = 2$, $\alpha = 1$, and $n = 60$.

of [Kossaifi 20b] use a weighted sum of the MSE, CCC and PCC losses, where weights are shake-shake regularisation coefficients [Gastaldi 17] sampled randomly and uniformly in the range $[0, 1]$. Differently, a dynamically weighted loss function $\mathcal{L}$ is proposed, and is defined as:

$$\mathcal{L} = f * \mathcal{L}_{MSE} + g * \mathcal{L}_{CCC} \tag{4.1}$$

where $\mathcal{L}_{MSE}$ is the MSE loss, $\mathcal{L}_{CCC}$ = 1 - CCC, and f and g are dynamic weight functions given by

$$f = \alpha \left(\frac{i}{n}\right)^k ; g = 1 - \left(\frac{i}{n}\right)^k, \tag{4.2}$$

where i denotes the i^{th} epoch in the training phase of a total n epochs, and $\alpha \in \mathbb{R}$ and $k \in \mathbb{Z}^+$ are hyper-parameters controlling the normalisation and non-linearity, respectively. The motivation for defining f and g in Equation 4.2 is to ensure that the network learns to maximise correlation initially and, then, minimise the error. Figure 4.5 illustrates f and g with the set parameters.

For estimating affect, generally, $\mathcal{L}_{CCC}$ is used to maximise the correlation between the ground-truth and the predicted values [Toisoul 21, Kossaifi 20b]. When a combination of loss functions with *static* coefficients is used, the model tries to simultaneously optimise error-based metric (say, MSE) and correlation-based metric (for example, CCC), which may result in a sub-optimal model. Empirically, the proposed loss function results in improved model performance as shown in Table 4.1.

Table 4.1: RMSE, PCC and CCC values of estimated valence using various loss functions in the subject-dependent setting.

Loss			*RMSE* ↓	*PCC* ↑	*CCC* ↑
MSE	PCC	CCC			
✓			0.17 ± 0.06	0.73 ± 0.26	0.68 ± 0.32
	✓		0.74 ± 0.08	0.52 ± 0.08	0.31 ± 0.05
		✓	0.25 ± 0.02	0.69 ± 0.11	0.69 ± 0.10
✓	✓	✓	0.22 ± 0.08	0.73 ± 0.08	0.73 ± 0.08
$\mathcal{L}$ (proposed)			$\mathbf{0.13 \pm 0.01}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.89 \pm 0.02}$

Table 4.2: RMSE, PCC and CCC values of estimated valence and arousal for the SI and SD settings.

Mode	Valence			Arousal		
	RMSE ↓	*PCC* ↑	*CCC* ↑	*RMSE* ↓	*PCC* ↑	*CCC* ↑
Subject-independent	0.35 ± 0.02	0.12 ± 0.11	0.10 ± 0.10	0.31 ± 0.02	0.29 ± 0.12	0.26 ± 0.12
Subject-dependent	$\mathbf{0.13 \pm 0.01}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.12 \pm 0.00}$	$\mathbf{0.93 \pm 0.00}$	$\mathbf{0.93 \pm 0.00}$

4.3.2 Implementation

The model is implemented using the open-source software library PyTorch [Paszke 19] and is trained on an NVIDIA A100 GPU with 40GB memory. Adam optimiser [Kingma 14] is used with a decrease of the learning rate by a factor of 10 for every 15 epochs, with the initial learning rate set to 10^{-3}. The models are trained for 60 epochs with a batch size of 128 and a dropout rate of 0.5. In the proposed loss function, fine-tuning is performed for hyper-parameters $k \in [1,2,3]$, and $\alpha \in [1,2,20]$. The results reported are the $\mu \pm \sigma$ values obtained via five-fold cross-validation.

4.4 Results and Discussion

Table 4.1 shows the RMSE, PCC, and CCC values obtained using the individual loss functions, a combination of the loss functions, and the proposed dynamically-weighted loss function for valence estimation within the SD setting. As mentioned earlier, a typical objective of regression models is to simultaneously minimise RMSE, while maximising PCC and CCC. When using the CCC loss function alone (row 3), the obtained RMSE is worse as compared to its MSE

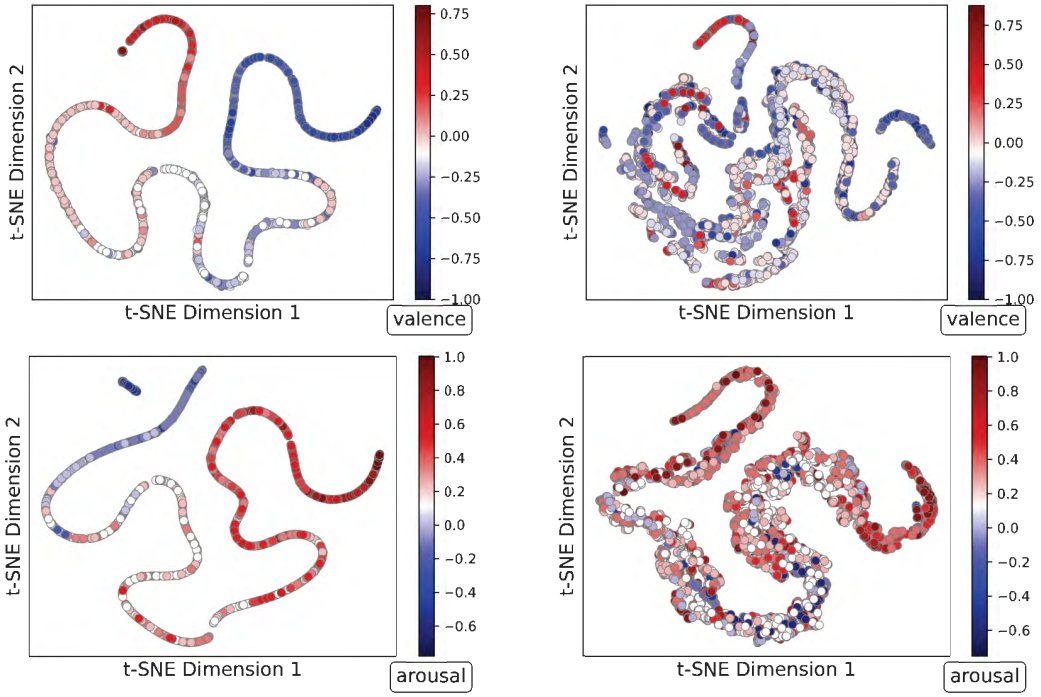


Figure 4.6: Visualisations of the feature distribution generated by t-SNE for valence (top row) and arousal (bottom row) using subject-dependent (left) and subject-independent (right) frameworks.

loss counterpart (row 1) as per t-test ($t(8) = 28.28, p < 0.0001$). When the PCC loss alone (row 2) is used, the achieved PCC and CCC values are lower than the CCC loss counterpart ($t(8) = 3.44, p < 0.0001$ and $t(8) = 8.48, p < 0.01$, respectively). This is because, as can be seen in Equation 3.3, CCC also incorporates the PCC value, but penalises correlated signals with different means [Lin 89]. That is, if the predicted feature has a trend similar to the target feature, but the predicted value is far from the target value, implying a high error, a low CCC is obtained, although the PCC is high. It is also observed that the RMSE value is optimised better when the CCC loss is used, as compared to the PCC loss as per t-test ($t(8) = 132.87, p < 0.0001$).

In comparison to the individual loss functions, when a weighted sum of the three loss functions is employed (where weights are the shake-shake regularisation coefficients [Gastaldi 17], as in [Toisoul 21]), PCC and CCC values are higher than the individual PCC loss as per t-test ($t(8) = 4.15, p < 0.005$) and no significance is observed with other individual losses. However, a trade-off is observed in terms of RMSE value as compared to the MSE loss counterpart. The proposed loss function performs the best as the RMSE value is the lowest, while the PCC and CCC values are the highest as compared to the other loss functions and are significantly different

for all appropriate pair-wise t-tests.. To account for optimising all three metrics simultaneously, the proposed loss function is designed to learn the correlations initially and, later, to minimise the mean squared error, in a continuous fashion where the (non-)linearity factor is controlled by the hyper-parameter k.

In this study, both subject-dependent and subject-independent experiments are performed for valence and arousal inference using AFEW-VA. While prior studies only performed subject-independent experiments [Kollias 18b, Kollias 19a, Kossaifi 20b, Mitenkova 19, Toisoul 21, Zhao 22], whether affect inference is required in an SI or SD setting may depend on the use-case. For example, if the affect inference system entails inferring affect from the end user, the model should be optimised for each user. Conversely, if the affect recognition is to be achieved independently of the end-user, the model should be optimised for the SI setting. The proposed loss function is employed in the experiments as it results in improved RSME, PCC and CCC values, as compared to the other loss functions.

The values in Table 4.2 are obtained using the proposed loss functions for the subject-independent and subject-dependent settings. As seen in the table, all performance metrics are improved in the SD setting as compared to the SI setting, as the SD setting results in the lowest RMSE and highest PCC and CCC values for both valence and arousal. In contrast, the model is unable to learn generalised features to discern diverse valence or arousal values in the subject-independent framework. To obtain further insights, t-distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten 08] is used to visualise the features learnt to estimate valence and arousal in the two settings (see Figure 4.6). As can be seen, the learned features for valence and arousal prediction in the SD setting are better separated as compared to the SI setting, where features corresponding to high/low valence/arousal values overlap considerably.

Overall, the results in the SD setting indicate the upper limit for arousal and valence estimation on the AFEW-VA dataset. Moreover, the results in Table 4.1 demonstrate that the proposed loss function results in superior prediction performance. Employing this loss function on the AFEW-VA dataset, considerably more precise valence and arousal estimation in the SD setting is observed as compared to the SI setting. Cumulatively, the results reveal that for the small AFEW-VA dataset with a highly imbalanced distribution of input samples per subject,

identity-specific characteristics substantially impact emotional inference. Conversely, emotion-specific representations cannot be efficiently learned across subjects as typified by the poor valence/arousal estimation in the SI setting.

4.5 Conclusion

In this study, the influence of limited video data and of an imbalanced distribution of samples per subject on continuous human affect (valence, arousal) inference is examined using the AFEW-VA dataset. While some studies in psychology state the existence of unique emotional expressions for the basic emotions, on the contrary, others hypothesise that emotional expressions of the same emotion vary substantially across individuals (and often for the same individual) due to factors such as context, social environment, *etc*. Hence, to infer affect computationally, it is essential to examine the affect inference using subject-dependent and subject-independent settings. A novel dynamically-weighted loss function is proposed, and is found to enhance correlation as well as reduce error between the target and predicted values. Empirically, it is observed that this loss function results in an improved performance than competing loss functions. Furthermore, superior performance in terms of RMSE, PCC, and CCC metrics is observed in the subject-dependent framework as compared to the subject-independent counterpart. The results indicate that the features of valence and arousal learnt by the model are not generalisable across subjects. This serves as a motivation for developing MT-CLAR and learning the affect differences, as opposed to learning the absolute valence and arousal values. Visualisations convey that the features of the subject-independent framework are not as discriminative as the subject-dependent setting.

Chapter 5

Learning Affect Differences via Weak-Supervision

Contents

The previous chapter includes the analysis of the impact of limited video and imbalanced distribution of samples per subject on continuous affect. Limited data might lack diversity in terms of emotional expressions, or scenarios depicted. This lack of diversity can hinder the model's ability to recognise a wide range of emotions accurately. While subject-independent studies focus on developing models or algorithms that generalise well across diverse populations, subject-dependent studies explore how individual traits, cultural backgrounds, or personal experiences influence emotional expressions and interpretations. Integrating insights from both

subject-independent and subject-dependent studies leads to more comprehensive affect inference systems, capable of capturing universal emotional patterns while accommodating individual variations. Training affect inference systems from scratch poses a challenge, as emotions are multifaceted and complex, making it challenging to represent them effectively with simple models using limited video data. In this chapter, transitioning from conventional techniques of using ground-truth for training, the aim is to capture affect differences for valence and arousal via weak supervision[1].

5.1 Introduction

Limited availability of precisely annotated datasets is a bottleneck, hindering the potential for comprehensive understanding and accurate inference of affect. Collecting affect data requires reporting individuals' personal experiences, raising ethical concerns regarding privacy and consent. For instance, collecting physiological signals, such as heart rate variability, or brainwave patterns, require specialised equipment and expertise to measure accurately. Some methods of measuring physiological signals can be intrusive and require physical contact with the body, leading to discomfort for the participants. While data collection presents one challenge, annotating the collected data comes as another significant hurdle. Annotating the ground truth can be done through various methods like collecting self-assessment reports, experts providing annotations, etc., as discussed in Section 2.3.2. However, obtaining ground-truth emotion labels through self-assessment reports is difficult, as people might not accurately report their emotions or might express them differently. Accurate annotations of affect data requires skilled annotators who can reliably interpret and label emotions. Consistent annotation across annotators is challenging due to the subjective nature of emotions.

AffectNet [Mollahosseini 19], a static in-the-wild database, comprises 1 million images. However, only $\approx 450,000$ are manually annotated with valence and arousal, of which only

[1]A part of this chapter is published at **ACM International Conference on Multimedia (MM) 2023**. Details of [Parameshwara 23a]: Ravikiran Parameshwara, Ibrahim Radwan, Akshay Asthana, Iman Abbasnejad, Ramanathan Subramanian, and Roland Goecke. *Efficient Labelling of Affective Video Datasets via Few-Shot & Multi-Task Contrastive Learning*. In Proceedings of the 31st ACM International Conference on Multimedia, pp. 6161-6170. 2023.

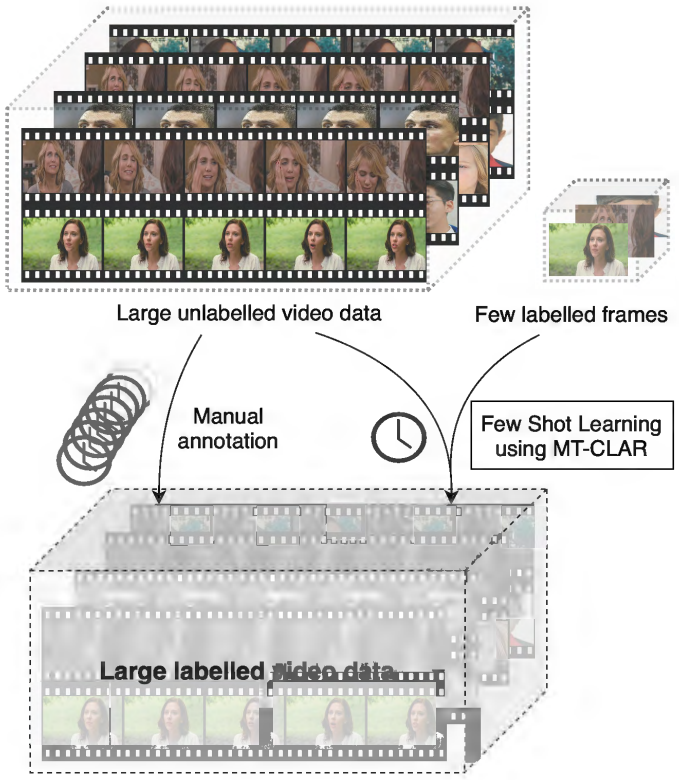


Figure 5.1: Annotating a large unlabeled video dataset is a time-consuming and tedious. With MT-CLAR few-shot learning, utilising as few as 11% labeled frames from the video dataset, excellent valence and arousal labelling is achieved for the remaining frames. While the base model MT-CLAR is discussed in this chapter, utilising MT-CLAR for few-shot affect labelling will be discussed in the next chapter.

$\approx 350,000$ are valid faces. Annotating affect videos is considerably more demanding, as it requires frame-level annotations. As a result, leveraging weak supervision methods is becoming increasingly common to harness the vast amounts of available but pseudo-labeled or unlabeled data for training affect inference models [Zhou 18, Pei 19, Narayana 23b].

Weak supervision allows the utilisation of vast amounts of data that might otherwise remain unlabeled or underutilised due to the impracticality of manual labelling. Weak supervision refers to a machine learning paradigm where training data is labeled or annotated using heuristics, or less precise sources instead of relying solely on manually labeled or ground-truth data. It offers a pragmatic solution in scenarios where obtaining high-quality manual annotations is difficult or resource-intensive. To overcome the challenges of affect inference with limited video data, this study proposes to employ weak supervision for labelling affect data. This will be done in two folds, and is presented as a detailed description across the current chapter, and the next

chapter.

As the first step towards achieving this, in the current chapter, Multi-task Contrastive Learning for Affect Representation, a metric learning-based Siamese network is employed with contrastive loss. Contrastive loss in SN captures the underlying (dis)similarity in the pair of input samples, and pulls together similar samples, while pushing apart dissimilar samples. Through metric learning, the model learns a distance metric between the samples, and is efficient for capturing discriminative features by learning the intra-class similarity and inter-class differences. Leveraging the inter-task dependency, MT-CLAR performs (a) the primary task of inferring similarity or dissimilarity of categorical emotions, (b) secondary task of predicting valence differential, Δ_v, and (c) tertiary task of predicting arousal differential, Δ_a, thus making it a weak-supervision method. Additionally, (a) landmark-driven AU attention module, and (b) background-masked facial input in MT-CLAR is proposed, to improve subject-independent affect estimation.

5.2 Key Contributions

In this chapter, the main contributions are as follows:

1. MT-CLAR, a novel metric-learning network with contrastive loss is proposed for affect representation.

2. Given a pair of images, the aim is to perform the following tasks, a) categorical classification of emotion (dis)similarity, b) prediction of difference in their valence values, and c) prediction of difference in their arousal values. To the best of our knowledge, this study is the first to predict valence differential and arousal differential.

3. For enhanced subject-independent affect estimation, this study proposes to employ (a) landmark-driven AU attention module, and (b) background-masked facial input in MT-CLAR.

4. Extensive experiments are perfomed through an ablative study to validate the design of MT-CLAR.

5. Among various configurations of the base models that are proposed, the Action Unit-guided input for MT-CLAR yielded the best performance in all three tasks, (dis)similarity classification, estimating valence and arousal differentials.

5.3 Prior Works

Examining emotions using the dimensional model with *valence* and *arousal* dimensions is a recent development [Tellamekala 19, Mitenkova 19] as discussed in Chapter 2. The various studies that have adopted contrastive learning, metric learning and multi-task learning for emotion inference are discussed below.

Contrastive learning is a self-supervised learning technique in which the model learns the *contrast* (similarity or dissimilarity) between samples. The model can generate effective representations of input data for downstream tasks with pretext contrastive learning. Contrastive loss was first proposed in [Chopra 05] to approximate the semantic distance between a pair of facial images. Facial emotional contrast on projected features in the valence-arousal space is analysed in [Kim 22a], using contrastive representation learning. Using a temporal sampling-based augmentation scheme, authors in [Roy 21] employ contrastive learning approach for facial expression recognition in videos. Further, contrastive learning has been applied for speech emotion recognition using SNs [Lian 18], cross-subject emotion recognition using EEG signal representations [Shen 22], and to learn discriminative facial Action Unit representations [Sun 21].

Associated to contrastive learning are the losses based on metric learning. In metric learning methods, the aim is to learn a metric space where the distance between points reflect similarity or dissimilarity. The model aims at learning an effective metric for generating discriminative features. In [Liu 20a], a metric learning framework is developed with an SN to investigate fine-grained distinction between facial expression. A deep SN that reflects the local structure of an embedding space, and modulates the learning to a classification space for facial expression recognition is employed in [Hayale 23]. Authors in [Wang 17] employ SN which incorporates latent facial attributes and long-term dynamics for dimensional emotion prediction.

Multi-task learning (MTL) improves the system generalisation by learning shared repre-

sentations between related tasks. It is a technique where a model is trained to perform multiple tasks simultaneously. Exploiting the inter-relatedness, these tasks can improve individual performance through a shared representation [Caruana 93]. In [Xia 15], authors integrate the secondary tasks of valence and arousal predictions to the major task of traditional categorical emotion recognition. The dependency between the tasks of valence and arousal prediction, expression classification, and Facial Action Unit detection is explored in [Zhang 23a]. Further, MTL frameworks are used to model the dependency effectively. Authors in [Chen 17] perform MTL for valence and arousal prediction on multimodal features, and observe improved performance with MTL, as compared to single-task counterparts. High performance results are observed in [Jeong 22], when multimodal input is fed to a multi-task model for performing the three tasks of valence-arousal prediction, facial expression classification, and Action Unit detection.

Observing the valence and arousal differential between two frames links to the framework of ordinal emotions proposed in [Yannakakis 18]. This framework provides evidence on the benefits of ordinal approaches for affect annotation and computational modelling of ordinal data. Authors in [Khademi 14] propose a relative facial AU classification by capturing relative changes in facial AUs, in contrast to using a single frame. This methods achieves robustness against individual differences among subjects. Considering an ordinal approach to modelling affect has shown to be superior over absolute methods [Eleftheriadis 17, Walecki 16, Narayana 23c]. Modelling the valence and arousal difference between two frames as proposed in this chapter, aligns with the idea of ordinal emotions, as it compares the positivity or negativity of the valence or arousal of one frame with respect to the other.

5.4 Proposed Framework

5.4.1 Multi-task Contrastive Learning

Contrastive learning aims at learning representations of data by contrasting between similar and dissimilar samples. Specifically, it aims at bringing similar samples into close proximity

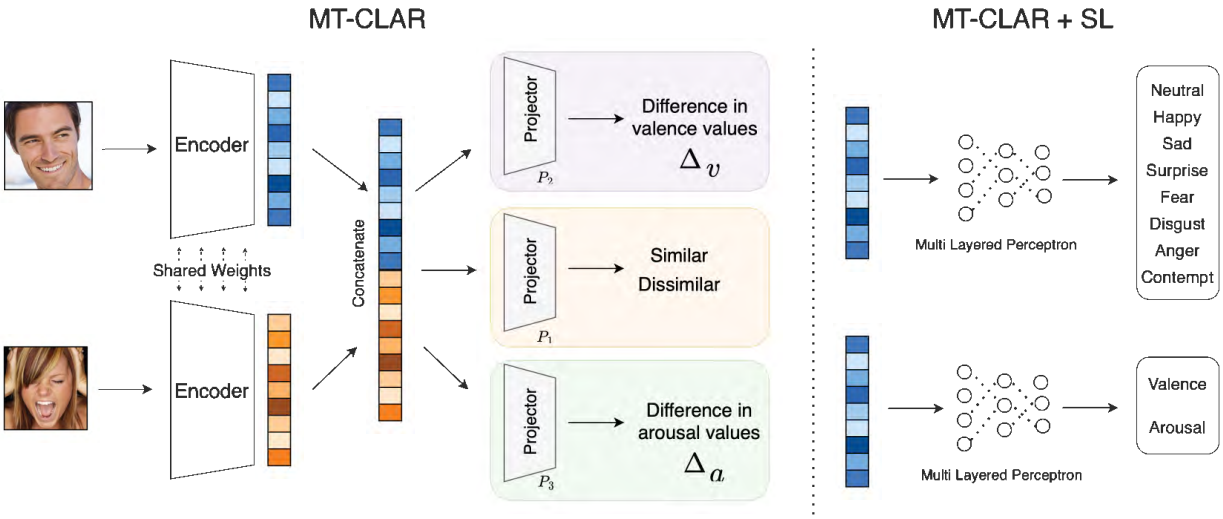


Figure 5.2: MT-CLAR overview: (Left) Differential estimation with MT-CLAR – A pair of expressive facial images is passed through a Siamese network, and their embeddings are concatenated to estimate (1) whether the expressions are similar/dissimilar, (2) the valence differential (Δ_v), and (3) the arousal differential (Δ_a) between expressions. Learned representations are utilised for supervised learning from individual images (MT-CLAR + SL). (Right) MT-CLAR + SL: Either image embedding is fed to a Multi-Layer Perceptron to infer the emotion class, and estimate the valence and arousal values.

in the representation space, while pushing dissimilar samples far apart. Several studies have shown that representational learning yields high quality representations for downstream tasks [Jing 19, Chen 20]. This study proposes to employ MT-CLAR to generate embeddings for a pair of input images. The proposed architecture of the SN has two streams with identical sub-networks for comparing the input images. Each sub-network comprises an *Encoder*, for transforming the input image to a high-level embedding. The embeddings from each branch are concatenated and a *Projector* comprising linear layers is employed for distinguishing the input images as similar or dissimilar. Additionally, to leverage the efficiency and faster learning capabilities of multi-task learning, the concatenated feature is used for predicting Δ_v and Δ_a of the input images.

Encoder

EmoFAN [Toisoul 21], which is built on top of the Face Alignment Network (FAN) [Bulat 17] is employed as the Encoder, $Enc(\cdot)$ in MT-CLAR. For an input facial image, EmoFAN jointly predicts discrete emotion classes, continuous affect dimensions (valence and arousal), and fiducial

facial landmarks. In this study, the pair of input images, x_1 and x_2 are mapped to the antepenultimate layer of EmoFAN, yielding the corresponding representation vectors, r_1 and r_2, given by $r_1 = Enc_1(x_1)$, and $r_2 = Enc_2(x_2)$, where $r_1 \in \mathbb{R}^{D_1}$ and $r_2 \in \mathbb{R}^{D_2}$ with $D_1 = D_2 = 256$. Likewise a typical SN, in the proposed model, $Enc_1(\cdot)$ and $Enc_2(\cdot)$ in the two streams share the parameters and weights to produce two features corresponding to the images.

Projector network

The vectors r_1 and r_2 obtained from $Enc_1(\cdot)$ and $Enc_2(\cdot)$ respectively are concatenated to obtain $u = r_1 \parallel r_2$, where $u \in \mathbb{R}^{D_c}$, where $D_c = 512$. To perform the three tasks of classifying the input image pairs as (dis)similar, predicting Δ_v, and Δ_a, u is fed as input to three branched projector networks, $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$, which map u to three vectors $w_1 = P_1(u), w_2 = P_2(u)$ and $w_3 = P_3(u)$. $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$ are all Multi-Layer Perceptrons (MLPs), with four identical fully connected (FC) layers, while differing in the number of neurons in the last FC layer. The four FC layers in $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$ comprise 2048, 1024, 512, and 128 neurons, however, $P_2(\cdot)$ and $P_3(\cdot)$ have a single neuron to predict Δ_v and Δ_a respectively, while $P_1(\cdot)$ has 2 neurons to classify (dis)similarity in the last FC layer. Prior to feeding to each FC layer, the input is normalised with zero mean and unit standard deviation and activated using ReLU activation.

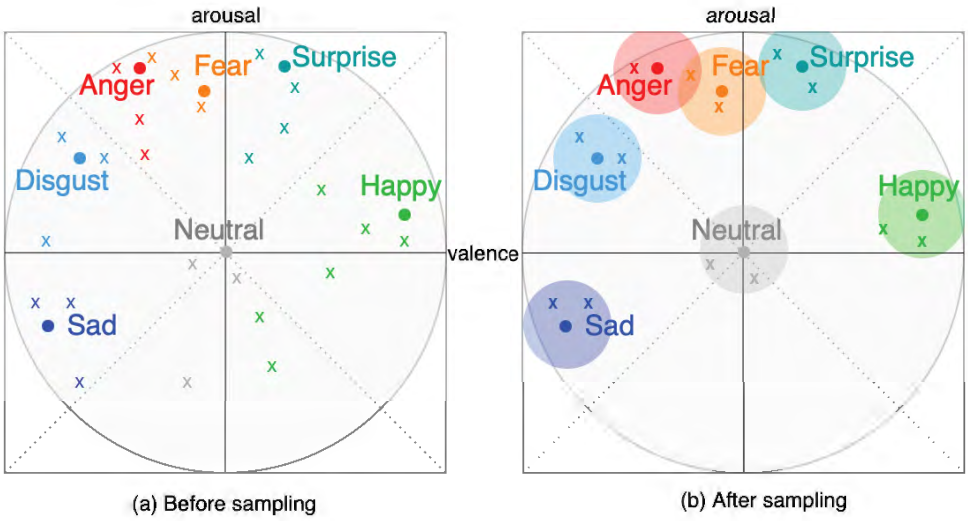
Loss function

Contrastive loss pulls embeddings of the *same class* more closer than those of *different classes* [Chen 20]. This facilitates the network to learn to distinguish between samples from different classes, while bringing samples of the same class together. To enable rich representations of the input images x_1 and x_2, contrastive loss, $\mathcal{L}_{cont}$ is applied on the vectors r_1 and r_2 as follows:

$$\mathcal{L}_{cont} = \frac{1}{N}\sum_{i=1}^{N} y_i(1-d_i) + (1-y_i)\max(0, d_i - m) \tag{5.1}$$

where, N denotes the batch size, d_i denotes the cosine distance between v_1 and v_2, m denotes the margin, and $y_i = 0$ for dissimilar samples, and $y_i = 1$ for similar samples.

Further, to classify the input images as similar or dissimilar in $P_1(\cdot)$, cross-entropy loss, $\mathcal{L}_{ce}$



(a) Before sampling (b) After sampling

Figure 5.3: Mikel's wheel [Mikels 05] illustration. Valence-arousal space visualisation pre-(a) and post-(b) sampling, where 'x' denotes data point with emotion category as its hue. *Best viewed in colour.*

is used on $z = P_1(u) = P_1(r_1 \,\|\, r_2)$.

Since predicting Δ_v and Δ_a is a regression problem, as discussed in Chapter 4, the aim is to reduce the MSE, while simultaneously maximising CCC. To predict Δ_v and Δ_a in $P_2(\cdot)$ and $P_3(\cdot)$ respectively, a dynamically weighted loss function $\mathcal{L}_\Delta$ is used, as given in Equation 4.1.

Overall, to optimise the parameters of MT-CLAR, a cumulative loss function $\mathcal{L}$ is used, given by,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cont} + \lambda_2 \mathcal{L}_{ce} + \lambda_3 \mathcal{L}_{\Delta_v} + \lambda_4 \mathcal{L}_{\Delta_a} \tag{5.2}$$

where, $\mathcal{L}_{\Delta_v}$ and $\mathcal{L}_{\Delta_a}$ are the $\mathcal{L}_\Delta$ corresponding to Δ_v and Δ_a branch respectively, and $\lambda_i, i \in \{1,2,3,4\}$, are shake-shake regularisation coefficients [Gastaldi 17], chosen randomly and uniformly in the range [0, 1], at each iteration of the training process. This ensures the network does not prioritise the reduction of any one of the losses [Toisoul 21].

Data sampling

Inspired by the success of prior studies in training SN using categorical emotions [Hayale 19], MT-CLAR takes as input a pair of images with *similar* or *dissimilar* label, based on the emotion categories.

It is crucial to identify relevant image pairs to train the SN. When the data used to train the SN is noisy or poorly labelled, the model might learn spurious information or irrelevant relationships between the samples. In the absence of clean data, the performance of SN can be severely compromised [Liu 20a]. Specifically, in our study, feeding clean data to the SN is of utmost importance, as it ensures the model can learn relevant features for identifying emotions accurately.

Mikels' Wheel of Emotions [Mikels 05] is a visual representation of categorical emotions in a valence-arousal (V,A) space, as shown in Figure 5.3 (left). For an emotion category and the corresponding $(v,a) \in (V,A)$ in the Mikel's wheel, a $d-$radius neighbourhood centered at (v,a) is created, as shown in Figure 5.3 (right). Data points within the respective emotion neighbourhood are sampled, while the anomalies are discarded. This ensures focused data points reflecting the true emotion to be considered in the study.

After sampling, a pair of images with label as *similar* or *dissimilar*, based on their emotion categories are fed as input to MT-CLAR.

5.4.2 Background-masked MT-CLAR

Background masking in emotion estimation using facial expressions helps enhance the accuracy, consistency, and focus of analysis by isolating and emphasising the critical facial features related to emotions [Mavadati 13]. A *background-masked* image has the background removed or isolated, to ensure focus remains solely on the facial features. Removing the background in a facial image helps direct attention solely to the main subject or foreground of the image, enhancing clarity and visual impact.

Active Appearance Models are statistical representation of the spatial configuration of landmarks [Cootes 01]. It encapsulates the appearance variations of the object. In [Mavadati 13], 66 facial landmark points are detected and extracted in each image using AAMs. Further, the landmarks are mapped to a canonical orientation using a similarity transformation. A binary mask is then fitted to the image to remove non-face background using the transformed landmark points. Inspired by this procedure, this study applies the background-mask for each image as depicted in Figure 5.4. The image is then fed to encoder of MT-CLAR, and subsequent steps henceforth

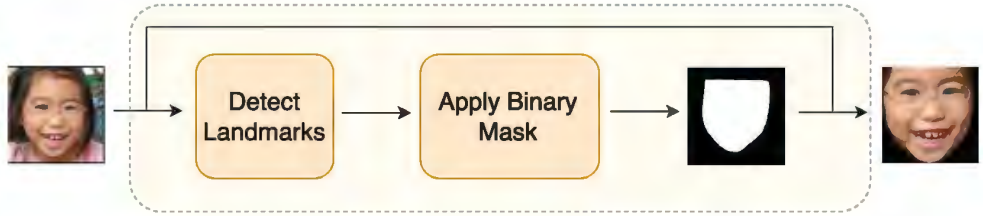


Figure 5.4: Pipeline for obtaining the background-masked input image.

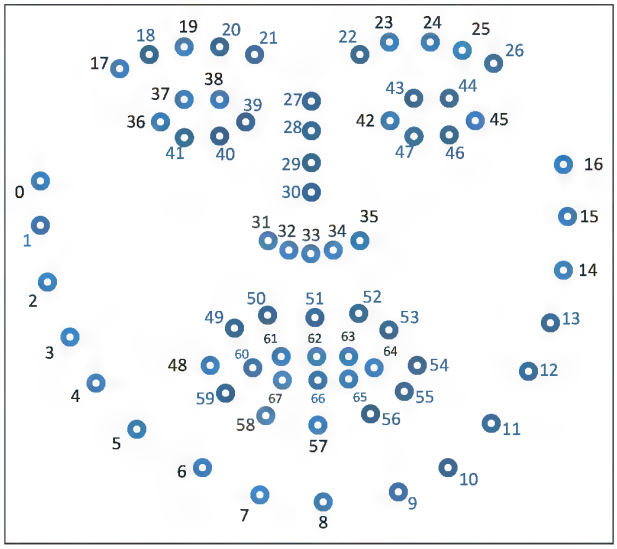


Figure 5.5: Facial landmarks of eyebrows, eyes, nose, mouth and cheeks represented in a 2-dimensional space. Image credits: [Baltrušaitis 18].

as discussed earlier in the previous section.

5.4.3 AU-guided MT-CLAR

Prior works have linked specific combinations and intensities of AUs to different emotions [Lucey 10]. Action Units, developed as part of the FACS, are used for detecting emotions because they serve as fundamental components of facial expressions, providing a standardised way to analyse and interpret facial movements related to emotions [Friesen 78]. Each AU is assigned a unique numerical code, simplifying the description and analysis of facial expressions. For instance, happiness is characterised by *AU6*, or *Cheek Raiser*, as it involves raising of cheeks. Sadness is characterised by the combination of *AU1-Inner Brow Raiser*, *AU4-Brow Lowerer*, *AU7-Lip Corner Depressor* and *AU15-Chin Raiser*. The various expressions associated with the AUs are shown in Table 5.1.

Determining the landmarks of key facial features like eye corners, mouth, etc., provide de-

Table 5.1: Description of the various Action Units and the corresponding landmarks. The landmarks ID on the 2D face region is illustrated in Figure 5.5. Images credits: [Baltrušaitis 18].

Description	**Images**	**AU** (a)	**Landmarks** (L_a)
Inner Brow Raiser		1	17, 18, 19, 20, 21, 22, 23, 24, 25, 26
Outer Brow Raiser		2	17, 18, 19, 20, 21, 22, 23, 24, 25, 26
Brow Lowerer		4	20, 21, 22, 23, 27
Upper Lid Raiser		5	36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47
Cheek Raiser		6	0, 1, 2, 3, 4, 5, 31, 36, 48, 11, 12, 13, 14, 15, 16, 35, 45, 54
Lid Tightener		7	36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47
Nose Wrinker		9	21, 22, 27, 28, 29, 30, 31, 32, 33, 34, 35
Upper Lip Raiser		10	49, 50, 51, 52, 53, 61, 62, 63
Lip Corner Puller		12	48, 54, 60, 64
Dimpler		14	48, 54, 60, 64, 4, 5, 11, 12
Lip Corner Depressor		15	48, 54, 60, 64
Chin Raiser		17	6, 7, 8, 9, 10, 48, 59, 58, 57, 56, 55, 54, 60, 64, 65, 66, 67
Lip Stretcher		20	48, 49, 50, 51, 52, 53, 54, 55, 56 ,57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Lip Tightener		23	48, 49, 50, 51, 52, 53, 54, 55, 56 ,57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Lips Part		25	48, 49, 50, 51, 52, 53, 54, 55, 56 ,57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Jaw Drop		26	6, 7, 8, 10
Lip Suck		28	48, 49, 50, 51, 52, 53, 54, 55, 56 ,57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Blink	-	45	36, 37, 38, 39, 30, 40, 41, 42, 43, 44, 45, 46, 47

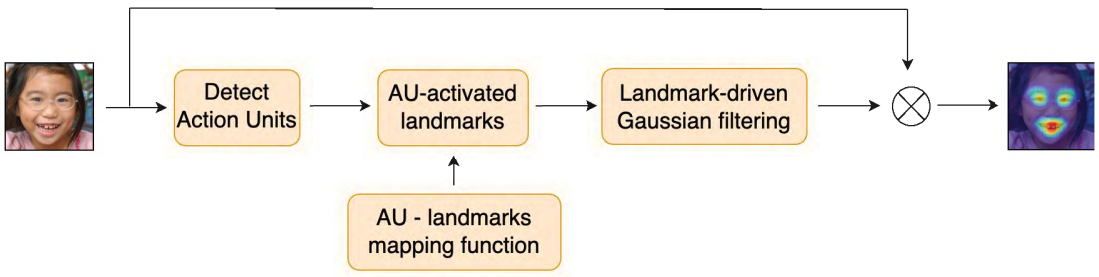


Figure 5.6: Pipeline of the integration of AUs and facial landmarks given an input image.

tailed spatial data about facial muscle movements and configurations. *Facial landmarks* help detect and track specific movements associated with emotions, such as eyebrow movements. They are often represented as coordinates (x, y) in a two-dimensional space, as shown in Figure 5.5. While they have been a long-standing part of anatomical studies and medical illustrations to understand facial structure, the initial computational approaches focused on manually defined landmarks for facial recognition, often limited to specific points on the face. Advancements in computer vision, and the adoption of deep learning architectures have significantly improved automated landmark detection.

FACS and facial landmarks offer complementary perspectives. While FACS focuses on muscle-based coding and understanding expressions through muscle movements, facial landmarks capture the spatial arrangement of facial features. With the aim of focusing on regions of the face where the emotional expressions are prominent, integrate AUs and facial landmarks are integrated in the input images fed to MT-CLAR. Essentially, this accomplishes a predefined attention strategy, specifying that the model should focus only on regions-of-interest, disregarding other parts of the image.

As depicted in Figure 5.6, the aim is to obtain a refined input with highlighted emotion-specific facial regions. To achieve this, the AUs are automatically detected, followed by obtaining the facial landmarks corresponding to the activated AUs as described in Table 5.1. This provides the geometric information of activated AUs on the face. The activated locations on the face are highlighted using Gaussian filters. Thus, an AU-highlighted facial image is obtained which is passed as input to the encoder to obtain a rich representation for downstream tasks, as illustrated in Figure 5.7.

Formally, given an input image $s \in S$, where S is the set of all images, consider a function f which maps the image to a set of AUs. That is, $\tau : S \rightarrow \mathcal{P}(A)$, such that $\tau(s) = A_s$, where $A = \{1,2,4,5,6,7,9,10,12,14,15,17,20,23,25,26,28,45\}$ as shown in Table 5.1 (AU column), $\mathcal{P}$ is the power set notation, and A_s denotes the set of activated AUs for the input s. Further, ψ maps the activated AU a to a set of facial landmarks L, as given in Table 5.1. This function provides the geometric information about the facial muscle movements based on [Liu 20b]. ψ is given by, $\psi : A \rightarrow \mathcal{P}(L)$, such that $\psi(a) = L_a$, where $a \in A$, $L = \{0,1,2,\cdots,67\}$, corresponding to the 68 landmarks in Figure 5.5, and L_a is a set of landmarks associated with a (henceforth called as *activated landmark*). As mentioned above and depicted in Table 5.1, the same landmarks can correspond to multiple AUs. For instance, *landmark 17* corresponds to AU1 and AU2. Hence, to avoid repeated occurrence of landmarks, the union of landmarks is considered when a combination of AUs are detected for an emotional expression. Therefore, for an input image s, $L_s = \bigcup \psi(A_s)$ is the set of all landmarks corresponding to the activates AUs.

For each landmark $l_i \in L_s$, where $i = \{1,2,\cdots,\mid L_s \mid\}$, the function ϕ maps the landmark l_i to the location (x,y) on s. That is, $\phi : S \times L \rightarrow X \times Y$, such that $\phi(s,l) = (x,y)$. A zero matrix, z_s of dimension identical to the size of the input image s is initialised. A 2-dimensional Gaussian filter is applied on each activated landmark, given by,

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}},$$

where σ is the standard deviation of the Gaussian distribution. This results in the regions corresponding to the activated landmarks being highlighted on z_s, while the unactivated regions are retained as zeros. Finally, an element-wise multiplication of z_s and s is performed to obtain the AU-driven attended image $\tilde{s}$, which is fed as input to MT-CLAR.

5.5 Experimental Setup

AffectNet dataset (see Section 3.2.2) is used to train all the MT-CLAR base models. The radius d is set to 0.2 for generating the neighborhoods in Mikel's Wheel for the data sampling proce-

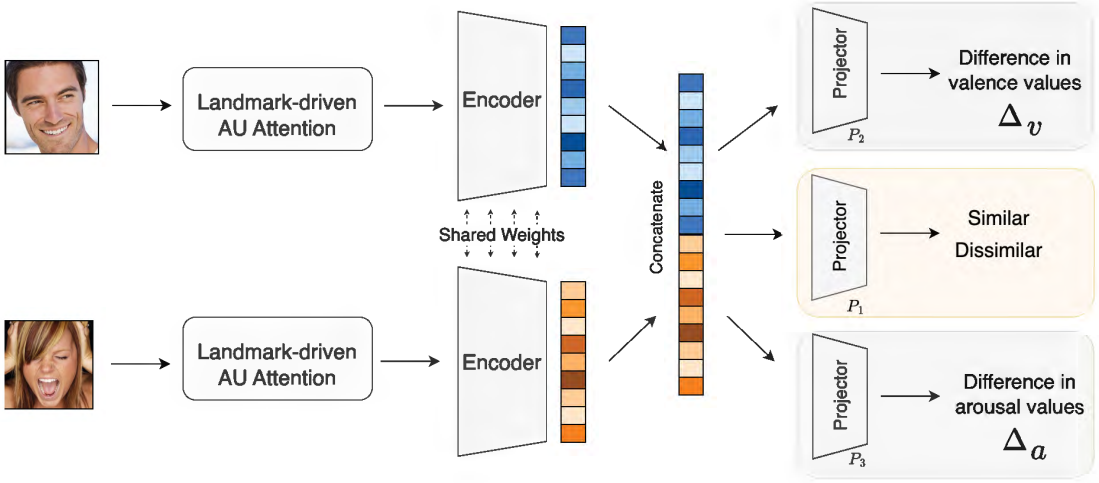


Figure 5.7: Overview of AU-guided MT-CLAR: The refined input guided by Action Units and landmarks are fed to the encoders of MT-CLAR.

dure. As a preliminary approach, d is constant for all the emotions. All the implementations are done using PyTorch [Paszke 19] software. The models are trained using four NVIDIA GeForce GTX 2080 Ti GPUs, each with 12 GB memory. Since the test set of AffectNet is not released, the validation set is used to evaluate the framework.

Following the implementations from [Toisoul 21, Bulat 22], the input images for MT-CLAR are first resized to 288×288, then randomly cropped to 256×256. Random affine transformation is applied on the training images with a rotation of up to 20 deg, translations up to 20% on both directions, scaling up to 20% up or down, and shearing up to 10 degrees. A horizontal flip is performed with a chance of 50%. MT-CLAR is trained for 40 epochs with a batch size of 256 using Adam [Kingma 14] optimiser. Learning rate is scheduled based on plateau detection, with the base learning rate set to 0.0001. It is decreased by a factor of 0.1 whenever a plateau is detected with *patience* value set to 5. The margin m used in contrastive loss (see Equation 5.1) is empirically set to 0.25. In the dynamic weight functions f and g (see Equation 4.2), used to compute $\mathcal{L}_{\Delta_v}$ and $\mathcal{L}_{\Delta_a}$ (see Equation 5.2), the fine-tuned hyper-parameters are $k \in \{1,2,3\}$ and $\alpha \in \{1,2,20\}$.

Histograms of Oriented Gradients (HOG)-based and geometry features (shape parameters and landmark locations)-based model are employed for automatic facial AU detection as proposed in [Baltrušaitis 15, Baltrušaitis 18]. The σ for Gaussian filtering is initially set as $min(h,w)/12$, where h and w denote the height and width of the input image, respectively. For

Table 5.2: Evaluating MT-CLAR design aspects via *AffectNet*. CE, Reg refer to cross entropy and regression loss, respectively. ↑ indicate higher the better.

Data sampler	Loss function	Task	Similarity accuracy ↑
No	CE	Single (similarity)	0.53
No	Contrastive	Single (similarity)	0.60
No	Contrastive + CE	Single (similarity)	0.67
Yes	Contrastive + CE	Single (similarity)	0.69
Yes	Contrastive + CE + Reg	Multi (similarity + Δ valence)	0.70
Yes	Contrastive + CE + Reg	Multi (similarity + Δ arousal)	0.70
Yes	Contrastive + CE + Reg	Multi (similarity + Δ valence + Δ arousal)	**0.72**

a fixed σ, if the Gaussian kernel does not fit the input image, then σ is iteratively reduced by half till a kernel fits the image. Fixing the value of σ is not possible, as the activated landmark points vary across images. OpenFace [Baltrušaitis 18] software is used to detect landmarks, apply binary mask, and get the background-masked output image.

5.6 Results and Discussion

5.6.1 Design of MT-CLAR

One of the primary aims of MT-CLAR is to learn robust representation of images. To this end, classification of a pair of input images is performed as similar or dissimilar in terms of affect. Since the task is to perform binary classification of a pair of input images, cross-entropy (CE) loss is employed. In this task, an accuracy of 0.53 is obtained as shown in row 1 of Table 5.2.

However, since it is observed that CE loss lacks robustness to noisy labels [Zhang 18] and the possibility of poor margins [Liu 16], it leads to reduced generalisation performance. Hence, to implement the idea of pulling together embeddings of the *same class*, while pushing apart embeddings of *different classes*, contrastive loss is used in place of CE loss. Contrastive loss facilitates in the distinction of samples belonging to the same class and those belonging to different classes. This plays a vital role in our study, as the aim is to obtain larger range of similarity scores, rather than predicting one of the binary classes. As shown in row 2 of Table 5.2, an improvement in the accuracy of MT-CLAR is observed, as compared to the case when CE loss is employed.

To optimise the similarity metric and the classification performance of the model, both CE loss and contrastive loss is employed in the model. CE loss is applied on the prediction obtained from the projector network, $P_1(\cdot)$, while contrastive loss is applied on the embeddings obtained as part of the SN. This enables the network to accurately classify the samples, besides recognising similarity. A further improvement in the accuracy is observed as shown in row 3 of Table 5.2.

Since the similarity measure is heavily reliant on the quality of annotations of the samples, an additional data sampling technique is induced on the AffectNet dataset (described in Section 5.4.1), as the training set has noisy labels. This technique narrows the region of samples of a particular class, and enables training on more realistic and reliable samples. With the inclusion of data sampler and both CE loss and contrastive loss, an increase in the accuracy is noted, compared to the previous cases, as shown in row 4 of Table 5.2.

In the aforementioned cases, MT-CLAR is assessed with a single task alone, namely binary classification of similar and dissimilar images from the AffectNet dataset. An additional task of predicting Δ_v, which refers to the difference in valence of the input image pairs is introduced. In order to improve the model's ability to distinguish dissimilar samples, feeding the ground-truth valence values of the image pairs enhances the discrimination power. For example, two image pairs such as *happy-sad*, and *happy-surprise*, have the same label (dissimilar), but the difference in valence values of *happy-sad* is greater than the difference in valence values of *happy-surprise*. In this regard, a slight increase in the accuracy of MT-CLAR is observed with the inclusion of this task, as shown in row 5 of Table 5.2. Furthermore, the prediction of Δ_a, referring to the difference in arousal of input image pairs, is employed as an additional task with the classification of image pairs as similar or dissimilar, identical result is obtained as can be seen in the row 6 of Table 5.2. Since row 5 and row 6 correspond to employing multiple tasks, the total loss is given by the sum of contrastive loss $\mathcal{L}_{cont}$, CE loss $\mathcal{L}_{CE}$, and Regression loss $\mathcal{L}_{\Delta_v}$ or $\mathcal{L}_{\Delta_a}$ (described in Section 5.4.1). In these cases, MT-CLAR leverages the shared information between the related tasks, and hence can learn more generalisable representations that are useful for the discriminating the samples.

Finally, both the tasks, namely (a) prediction of Δ_v, and (b) prediction of Δ_a are considered in

Table 5.3: A comparison of the results of MT-CLAR, AU-guided MT-CLAR, and Background-masked MT-CLAR on the validation set of *AffectNet*. Arrows indicate lower (↓) or higher (↑) the better.

Input	Dimensional labels	Δ valence			Δ arousal			Similarity accuracy ↑
		RMSE ↓	PCC ↑	CCC ↑	RMSE ↓	PCC ↑	CCC ↑	
Raw	Raw ∈ [-1, 1]	0.50	-0.17	-0.17	0.43	0.00	-0.01	0.72
Raw	Normalised ∈ [0,1]	0.26	0.61	0.60	0.27	0.46	0.44	0.71
Masked	Normalised ∈ [0,1]	0.20	0.67	0.64	0.19	0.52	0.49	0.71
AU-guided	Normalised ∈ [0,1]	**0.19**	**0.69**	**0.66**	**0.18**	**0.55**	**0.53**	**0.75**

addition to the binary classification. Similar to the previous two cases, this multi-task approach also considers the sum of contrastive loss $\mathcal{L}_{cont}$, CE loss $\mathcal{L}_{CE}$, and Regression loss $\mathcal{L}_{\Delta_v}$ and $\mathcal{L}_{\Delta_a}$ (described in Section 5.4.1) as the overall loss function. This addition is incorporated with the aim of further increasing the robustness of MT-CLAR. In this scenario, MT-CLAR achieves the maximum accuracy of 0.72, compared to the previous cases, as can be seen in the last row of Table 5.2. Hence, this model is referred to as MT-CLAR, and is considered for obtaining the affect representation of image for downstream tasks.

5.6.2 Performance Comparison of Base Models

Table 5.3 presents a comparison of the results of Raw-MT-CLAR (trained with original images), AU-guided MT-CLAR and background-masked MT-CLAR. Since the best performance is obtained by combining the contrastive loss, cross-entropy loss, and regression loss, the same is employed in all of the models. When a pair of original images are fed as input, and trained using the ground-truth dimensional labels for Δ valence and Δ arousal estimation, poor performance is observed on the latter tasks. This setup yields the least PCC and CCC, and the highest RMSE. However, with respect to (dis)similarity classification, the model yields an accuracy of 72%, as shown in row 1 of Table 5.3.

Normalising the true labels can stabilise the learning process by preventing extreme fluctuations in loss or gradient descent during model training. Normalisation can act as a form of regularisation, aiding in preventing overfitting, and allowing models to converge faster and more smoothly [Goodfellow 16]. Considering this, the dimensional labels in AffectNet are nor-

malised for training the MT-CLAR base models. An improved PCC and CCC, and a lower RMSE as compared to the Raw-MT-CLAR is observed for Δ_v and Δ_a estimation, as shown in row 2 of Table 5.3. However, a similar classification accuracy is obtained, as the (dis)similar labels remain the same as in the previous setup.

Using the background-masked image as input and normalised labels, a further superior performance is achieved in background-masked MT-CLAR, as compared to Raw-MT-CLAR. An increase in PCC and CCC, and decrease in RMSE for Δ_v and Δ_a estimation, and a comparable classification accuracy is observed as shown in row 3. Since the training hyperparameters are identical for all the models, the performance efficacy can be attributed to the proposed model design as the model focuses only on the facial regions where emotions are expressed.

Row 4 of Table 5.3 corresponds to the results obtained with AU-guided MT-CLAR. This version of the base model yields the best result, in terms of least RMSE, and highest PCC and CCC for Δ_v and Δ_a estimation, and highest accuracy in the classification task. This could be attributed to the activated landmark-highlighted inputs, in addition to masking of background pixels.

5.7 Conclusion

Different from the previous chapter which dealt with subject-specific and subject-agnostic emotion inference using limited data, this chapter aims to develop robust models that can learn similarity of affect, and the valence and arousal differential between a pair of images. This results in the model yielding effective and general representations, which can be further used for downstream applications. The proposed MT-CLAR base models are trained using the (dis)similarity information alone, without knowing the ground-truth affect labels, thereby making MT-CLAR robust weakly-supervised models. Among various configurations of the base models that are proposed, the Action Unit-guided input for MT-CLAR yielded the best performance in all three tasks, (dis)similarity classification, estimating valence and arousal differentials.

In the next chapter, the aim is to utilise the affect representations obtained from these base models for performing various downstream tasks.

Chapter 6

Few-Shot Labelling

Contents

The previous chapter highlights the challenges of data collection and annotation for affect inference. The availability of limited affect data inspired the development of MT-CLAR base models for inferring affect similarity, valence differential and arousal differential employing weak-supervision approach. The primary aim of these base models is learning rich affect representations, which can additionally be used for downstream tasks. This chapter describes two applications of the base models, (a) a fully supervised learning setup for affect estimation, and (b) a few-shot learning-based approach for affect labelling. Few-shot learning can generalise well with only a few examples, making it more practical in scenarios where labeled data is

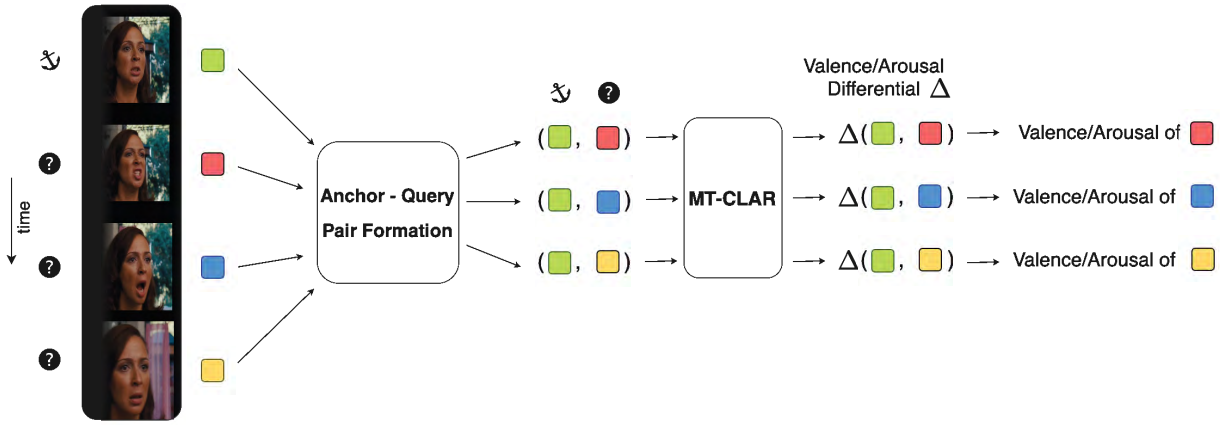


Figure 6.1: FSL overview: Given a video with known valence and arousal values for a few frames (*anchors*), our approach aims at inferring valence and arousal values for the remaining frames (*query frames*) using few-shot learning. MT-CLAR, a Multi-Task Contrastive Learning framework for Affect Representation is employed to infer similarity or dissimilarity of a pair of input images, valence differential (difference in valence of the input image pair), and arousal differential (difference in arousal of the input image pair). Valence (Arousal) of the query frame is inferred using the anchor and the valence (arousal) differential value. *Best viewed in colour.*

scarce or expensive to obtain[1].

6.1 Introduction

Affective data annotation is a time-consuming, costly, and an intensive task. It requires skilled annotators to carefully scrutinise each sample. Furthermore, neither the emotional manifestations, nor the perception are universal; facial expressions which are often considered to be universal, can vary in meaning and interpretation depending on the culture of the subject expressing them, and the culture of the annotator [Gendron 18]. Hence, annotators providing a subjective judgement of the emotion creates plausible bias. *Evaluator reaction lag* is a common problem while annotating continuous emotions in videos [Huang 15]. Such annotation problems are seen in multiple dimensional affect datasets, where the valence-arousal values are misaligned, and shifted either forward or backward [Tellamekala 19, Kollias 19a]. These misaligned annotations could result in a deceptive supervised representation learning model. Consequently, this

[1]A part of this chapter is published at ACM International Conference on Multimedia (MM) 2023. Details of [Parameshwara 23a]: Ravikiran Parameshwara, Ibrahim Radwan, Akshay Asthana, Iman Abbasnejad, Ramanathan Subramanian, and Roland Goecke. *Efficient Labelling of Affective Video Datasets via Few-Shot & Multi-Task Contrastive Learning*. In Proceedings of the 31st ACM International Conference on Multimedia, pp. 6161-6170. 2023.

ambiguity might hinder the model's ability to learn generalisable representations.

To address these challenges, this study proposes to use ***Few-Shot Learning*** (FSL) as an alternative, which compensates for the shortage of annotated samples in the target domain [Wang 20]. FSL algorithms learn from a few labelled examples and can generalise to new tasks with limited or no additional data. A *support-set* comprising a few labelled samples per class is used to train the model to label *query* (test) samples.

This study performs FSL via MT-CLAR, a novel approach to infer affect differences proposed in the previous chapter. MT-CLAR involves a Siamese network trained via contrastive loss, which captures the underlying (dis)similarity in a pair of expressive facial images. Using metric learning, MT-CLAR effectively learns intra-class similarities and inter-class differences. Leveraging multi-task learning, MT-CLAR primarily infers expressive facial pair similarity/dissimilarity in terms of categorical emotions, and secondarily predicts differentials in valence (Δ_v) and arousal (Δ_a). Utilising a few labelled *anchor* video frames, and the estimated Δ_v and Δ_a from MT-CLAR, the remainder of a video can be automatically labelled for valence and arousal, respectively (Figure 6.1).

Additionally, MT-CLAR + SL is implemented, where supervised learning (SL) is performed on the representations obtained from MT-CLAR, to classify categorical emotions and predict valence and arousal values of images.

6.2 Key Contributions

The main contributions of this chapter are as follows:

1. To the best of our knowledge, this study is the first to employ FSL-based approach to dynamic facial valence and arousal labelling in videos. Experiments on AffectNet [Mollahosseini 19] and AFEW-VA [Kossaifi 17] confirm that MT-CLAR generalises well, and can outperform the state-of-the-art with a support-set of only 6% the size of AFEW-VA, a video dataset.

2. AU-guided MT-CLAR results in superior performance when employed as a base model

for FSL-based affect labelling, as compared to background-masked MT-CLAR and Raw MT-CLAR, confirming that AU-activated attended input is effective for dimensional affect labelling with limited labelled samples.

3. MT-CLAR is further extended via supervised learning (MT-CLAR + SL) to deduce categorical and dimensional emotion labels for singleton images as in [Toisoul 21, Kossaifi 20b]. Extensive experiments confirm that MT-CLAR + SL achieves state-of-the-art results on multiple metrics for the AFEW-VA dataset [Kossaifi 17] and highly competitive results on AffectNet [Mollahosseini 19].

6.3 Prior Works

In this section, a brief survey of works on FSL in the context of emotion inference is provided. FSL relieves the burden of collecting large-scale annotated data [Wang 20]. It aims at classifying samples from a target domain, using a small number of labelled examples. While FSL has been widely applied for gesture recognition [Pfister 14], person identification [Wu 18], video action recognition [Careaga 19] etc., much attention is also given for FSL-based emotion inference tasks recently. *Meta learning* is a technique in FSL, where the model learns generic information across tasks, in order to adapt to new tasks based on few samples [Hochreiter 01]. Authors in [Ciubotaru 19] demonstrate the efficacy of low-shot learning for facial expression recognition via meta learning. An effective cross-domain FSL method is proposed in [Zou 22], where a two-stage learning framework is employed to infer compound facial expressions. *Metric learning* is another promising FSL technique which deals with learning a distance function that can compare the distance between samples [Hilliard 18]. Metric-based FSL is used to infer categorical emotions in scripted speech data [Feng 23, Ahn 21]. FSL is also applied to infer fine-grained valence and arousal values using physiological signals [Zhang 22].

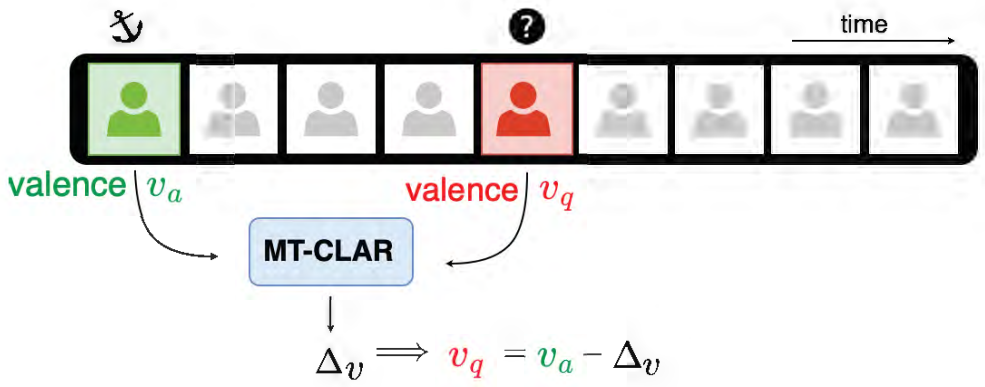


Figure 6.2: Few-shot learning: Given an *anchor* video frame (green) whose valence/arousal rating is known, and a *query* frame (red), MT-CLAR predicts valence/arousal rating of the query frame.

6.4 Proposed Framework

6.4.1 Few-Shot Learning

Besides requiring large amount of annotated data, the standard affect inference systems assume that a model learned from the training videos can generalise well on the test videos. This assumption increases the difficulty for applying in real-world scenarios, as it is unfeasible to obtain the training data from practical applications where the system would be deployed. Generalisability emerges as a major problem, as the affect inference system can overfit on certain training data, but will not generalise on the test data. Few-shot learning is a promising approach for affect inference, as they can learn to generalise from the limited samples of the support set, additionally eliminating the need for large amount of labelled data for training [Tyukin 21].

Typically, in a metric-based FSL model, leveraging a *support set* $S = \{(x_1, y_1), (x_2, y_2), ..., (x_k, y_k)\}$, which is a small set of k samples, the goal is to predict the label of the query sample x_q. A metric function $f(x; \theta)$ is defined to map the sample x to an embedding space, parameterised by θ. The *distance* between x_q and each support sample $x_i \in S$ is computed as, $d(f(x_q), f(x_i))$. The predicted label for x_q is the one that corresponds to the closest support sample given by,

$$\arg\min_i d(f(x_q), f(x_i)) \tag{6.1}$$

The analogous version of metric-based FSL employed in this study is illustrated in Fig 6.2 and is described as follows. Similar to the typical scenario, a support set S is considered which has frames with known valence and arousal values, with the aim of predicting the valence and arousal value for a query frame. However, instead of computing the distance between a query frame x_q and each $(x_i, y_i) \in S$, differently, for a given query frame x_q, an *Anchor set* A_S is formed, which is a subset of S to predict the valence and arousal value. For a pair of frames, x_i (from A_S) and x_q, using MT-CLAR, Δ_v is obtained, which is,

$$\Delta_v = y_{i_{val}} - y_{q_{val}} \tag{6.2}$$

where $y_{i_{val}}$ is the valence of x_i. Hence, $y_{q_{val}}$ is given by,

$$y_{q_{val}} = y_{i_{val}} - \Delta_v \tag{6.3}$$

Similarly, using Δ_a obtained from MT-CLAR, and $y_{i_{asl}}$ (arousal of x_i), the arousal of x_q is given by,

$$y_{q_{asl}} = y_{i_{asl}} - \Delta_a \tag{6.4}$$

Configurations of support set

For dynamic emotion annotation, only an anchor set $A_S \subset S$ is used for precision and efficiency. Multiple A_S configurations are shown in Figure 6.3, and described below:

1. **First frame of a video:** As depicted in Figure 6.3 (a), for a query image x_q, the first frame of the parent video to which x_q belongs is considered as the anchor x_a. In this case, for each x_q, the anchor set A_S has a single frame, given by $A_S = \{x_a\}$. Hence, the support set S constitutes the first frame of each video.

2. **Random frame of a video:** As depicted in Figure 6.3 (b), for a query image x_q, a random frame of the parent video to which x_q belongs is considered as the anchor x_a. Similar to

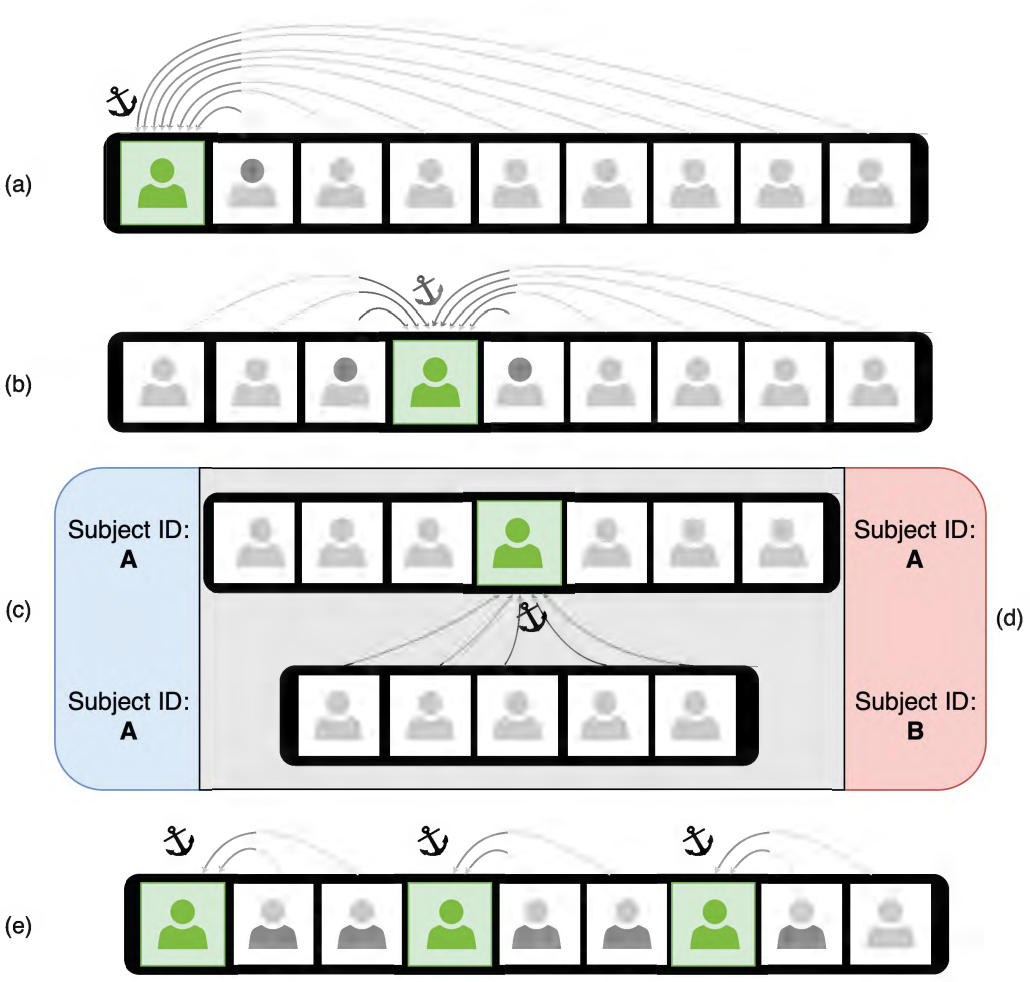


Figure 6.3: Support-set (S) configurations involve one or more anchor video frame(s) (in green) such as (a) first frame, (b) random frame, random frame involving (c) same, and (d) different subject, and (e) recurring n^{th} frame of a video.

case (1), for each x_q, the anchor set A_S has a single frame, given by $A_S = \{x_a\}$. However, the support set S constitutes a random frame of each video.

3. **Random frame of a subject-specific video:** As shown in Figure 6.3 (c), given a query image x_q, a random frame from a video of the same subject as the parent video of x_q is considered as the anchor x_a. The constraint followed here is that, the parent videos of x_q and x_a are different, but of the same subject. For each x_q, the anchor set A_S has a single frame, given by $A_S = \{x_a\}$. Here, the support set S comprises random frames of subject-specific videos, and query frames are from other videos of the same subject.

4. **Random frame of a subject-agnostic video:** As shown in Figure 6.3 (d), given a query image x_q, a random frame from a video of a different subject as the parent video of x_q is considered as the anchor x_a. Different to case (3), videos with the same subject as parent videos of x_q and x_a are not considered, and are restricted to videos with different subjects. For each x_q, the anchor set A_S has a single frame, given by $A_S = \{x_a\}$. In this case, the support set S comprises random frames of videos in a subject-agnostic manner.

5. **Recurring n^{th} frame of a video:** As depicted in Figure 6.3 (e), every n^{th} frame of a video is considered as an anchor. Hence, S comprises every n^{th} frame of each video. The following two sub-cases are considered in this setting:

   - For a query image x_q, the latest preceding anchor x_a of the parent video is considered for predicting the valence and arousal values, and for each x_q, the anchor set A_S has a single frame, given by $A_S = \{x_a\}$.
   - For a query image x_q, all the anchors of the parent video of x_q constitute A_S. Hence, A_S is given by,

   $$A_S = \{x_{a_1}, x_{a_2}, \cdots, x_{a_t}\} \tag{6.5}$$

   where $t = \lceil v/n \rceil$, where v is the total number of frames in the parent video of x_q. In this case, the mean of the valence (arousal) value obtained with each anchor A_S is considered as the predicted valence (arousal) of x_q. Different from all the aforementioned cases, here A_S has $\lceil v/n \rceil$ frames.

6.4.2 MT-CLAR with Supervision

Additionally MT-CLAR + SL is employed, which is a supervised learning framework to evaluate the robustness of the embeddings obtained from the SN of MT-CLAR. To this end, MT-CLAR + SL is used to perform (a) classification with respect to eight emotion categories, and (b) prediction of continuous valence and arousal values.

Architecture

For an image, the embedding obtained from the SN is used as input for MT-CLAR + SL. In the case of classification, MT-CLAR + SL uses a Multi-Layer Perceptron, $M_C(\cdot)$ to map an input representation vector x to a vector y_c, given by, $y_c = M_C(x)$, where $y_c \in \mathcal{R}^{D_C}$ and $D_C = 8$, corresponding to the eight emotion categories. In the case of prediction of valence and arousal values (a regression problem), MT-CLAR + SL uses a Multi-Layer Perceptron, $M_R(\cdot)$ to map an input representation vector x to a vector y_r, given by, $y_r = M_R(x)$, where $y_r \in \mathcal{R}^{D_R}$ and $D_R = 2$, corresponding to valence and arousal values. Both $M_C(x)$ and $M_R(x)$ have four FC layers, with 1024, 512, 256, and 128 neurons, respectively. The input is normalised with zero mean and unit standard deviation, and activated with ReLU activation function before passing to each fully-connected layer.

Loss function

Cross-entropy loss, $\mathcal{L}_{CE}$ is applied to classify the embeddings to eight emotion classes. The loss function given in Equation 4.1 is used to predict valence and arousal values. Following [Toisoul 21], as valence and arousal are jointly predicted, $\mathcal{L}_{mse}$ and $\mathcal{L}_{ccc}$ are given by,

$$\mathcal{L}_{mse} = MSE_v + MSE_a \tag{6.6}$$

$$\mathcal{L}_{ccc} = 1 - \frac{CCC_v + CCC_a}{2} \tag{6.7}$$

where, MSE_v (MSE_a) and CCC_v (CCC_a) denote the mean square error and CCC obtained with valence (arousal) prediction, respectively.

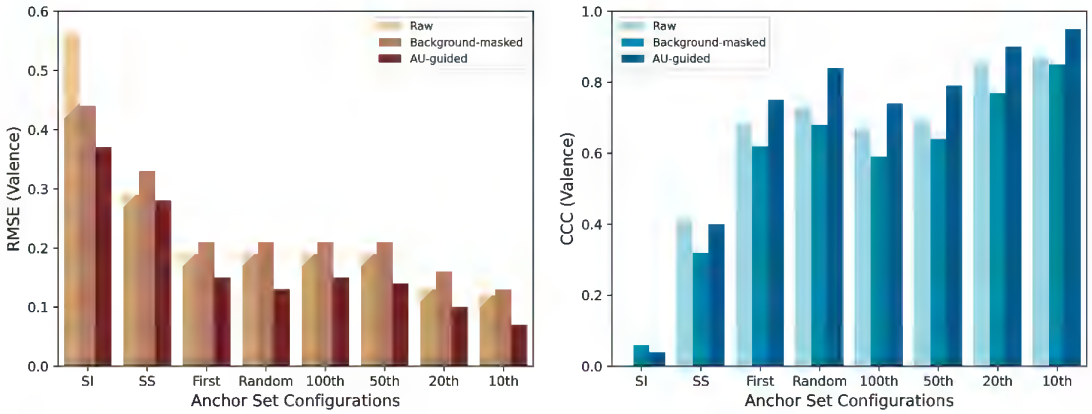
6.5 Experimental Setup

In this study, AFEW-VA is used to perform FSL and infer the valence and arousal values for each frame in the videos (described in Section 6.4.1). Additionally, it is used to train the MT-CLAR + SL model for continuous emotion inference, using both subject-independent and subject-dependent data-split strategy.

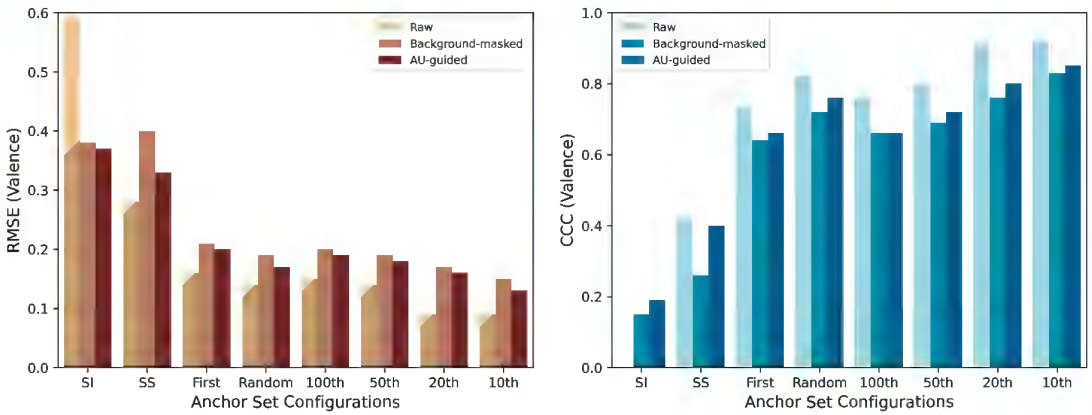
Similar to the previous chapter, RMSE, PCC, CCC, and SAGR are used as performance metrics. The details about the evaluation metrics can be found in Section 3.3. Additionally, 'accuracy' is used as the measure to evaluate categorical emotion classification performance of MT-CLAR + SL.

All the implementations are done using PyTorch [Paszke 19] software. The models are trained using four NVIDIA GeForce GTX 2080 Ti GPUs, each with 12 GB memory. Since the test set of AffectNet is not released, the validation set is used to evaluate the framework. All the results on AFEW-VA are fine-tuned using subject-independent and subject-dependent 5-fold cross validation strategy (5FCV).

The input images for MT-CLAR are center-cropped to 256 $\times$ 256, and no augmentations are applied since MT-CLAR is used in the *inference* mode. The input vector dimension for MT-CLAR + SL is 256. MT-CLAR + SL is trained for 60 epochs with a batch size of 512 using Adam optimiser. The base learning rate is set 0.001, and is decreased by a factor of 10 every 15 epochs. The hyper-parameters k and α in the dynamic weight functions f and g are fine-tuned identical to MT-CLAR, as mentioned above.



(a) FSL Labelling **without fine-tuned** base models.



(b) FSL Labelling **with fine-tuned** base models.

Figure 6.4: Comparison of RMSE (left) and comparison of CCC (right) across *Raw, background-masked*, and *AU-guided* MT-CLAR as base models for **valence** estimation using various anchor set configurations. SI and SS denote the anchor set configuration of a random frame from a subject-independent and subject-specific video, respectively.

6.6 Results and Discussion

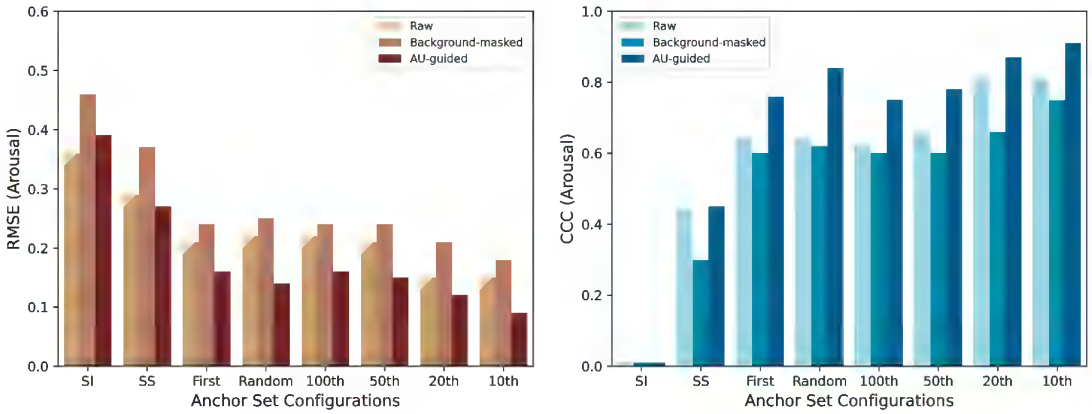
6.6.1 Effectiveness of AU-guided MT-CLAR as Base Model

This study is the first work to generate valence and arousal labels for videos via FSL utilising a labelled *anchor set* (Section 6.4.1). All results in Table 6.1, Table 6.2, and Table 6.3 are obtained by training *Raw MT-CLAR*, *background-masked MT-CLAR*, and *AU-guided MT-CLAR*, respectively on AffectNet, and evaluating the model on the AFEW-VA test set (mean values obtained over 5FCV). For each A_S configuration, results are reported without and with fine-tuning on the AFEW-VA train set. Notably, the 'No' rows correspond to conditions where only a specified number of AFEW-VA anchor frames (equal to $|S|$) are available. In this section, a comparison of the results of FSL-based labelling using the three versions of MT-CLAR is presented.

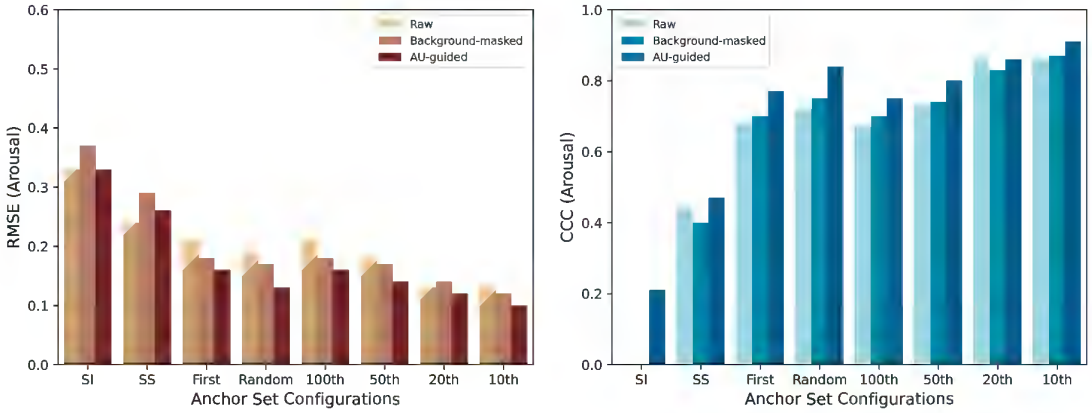
Figure 6.4 and Figure 6.5 presents the RMSE and CCC values corresponding to valence and arousal estimation, respectively. For brevity, the comparison is performed using a correlation-based metric and error-based metric, similar to [Kossaifi 17] and [Mitenkova 19].

- **Random frame of a subject-agnostic video:** Across various configurations, worst RMSE and CCC is obtained when a random frame of a subject-agnostic video is considered. Further, AU-guided MT-CLAR as base model yields significantly less RMSE, as compared to Raw and Background-masked MT-CLAR as base model as per one-way ANOVA ($F(2,12) = 42.21, p < 0.00001$) and pairwise Tukey's Honestly Significant Difference (HSD) test ($p < 0.05$). However, no significant differences were observed for CCC at $p < 0.05$. Similarly, in the fine-tuned case, worst RMSE and CCC is observed in all three models across various configurations. Further, Raw MT-CLAR results in the worst RMSE and CCC, as shown by a one-way ANOVA, and confirmed by a Tukey's HSD test at $p < 0.05$.

  In the case of arousal estimation, across various configurations, a similar trend of worst RMSE and CCC is observed when a random frame of a subject-agnostic video is considered. A one-way ANOVA shows significance of MT-CLAR as base models ($F(2,12) =$



(a) FSL Labelling **without fine-tuned** base models.



(b) FSL Labelling **with fine-tuned** base models.

Figure 6.5: Comparison of RMSE (left) and comparison of CCC (right) across *Raw, background-masked*, and *AU-guided* MT-CLAR as base models for **arousal** estimation using various anchor set configurations. SI and SS denote the anchor set configuration of a random frame from a subject-independent and subject-specific video, respectively.

$10.4, p < 0.05$)) for RMSE. However, no significance is observed in Raw vs AU-guided input, as per Tukey's HSD test. In the fine-tuned case, this configuration results in the worst RMSE and CCC values. Further, a one-way ANOVA did not reveal any significant difference between the RMSE values of the three MT-CLAR base models. However, AU-guided MT-CLAR as base model results in the best CCC, as shown by a one-way ANOVA and further confirmed by a Tukey's HSD test.

- **Random frame of a subject-specific video:** Considering valence estimation without fine-tuning and a random frame from a subject-specific video as the anchor set, there is no significant difference of RMSE values across the three MT-CLAR models, as per a one-way ANOVA test at $p < 0.05$. Background-masked MT-CLAR as base model yields the least CCC as per a one-way ANOVA ($F(2,12) = 9.96, p < 0.05$) and confirmation by a Tukey's HSD test. In the fine-tuned case, a one-way ANOVA shows significant differences in the RMSE and CCC values, with the Tukey's HSD test confirming that Background-masked MT-CLAR yields the worst RMSE and CCC.

  With respect to arousal estimation without fine-tuning, a similar trend of worst RMSE and CCC is observed in the background-masked MT-CLAR as base model, as shown by a one-way ANOVA and further confirmed by a Tukey's HSD test. With fine-tuning, the RMSE values are not significantly different as shown by a one-way ANOVA, while the CCC values of Background-masked MT-CLAR and AU-guided MT-CLAR differ significantly ($p < 0.05$).

- **First frame of the video:** In the valence estimation task without fine-tuning and using the first frame of the video as the anchor set, a significant difference is observed between the RMSE values of Raw and AU-guided MT-CLAR, and Raw and background-masked MT-CLAR ($F(2,12) = 17.26$, $p < 0.05$). A Tukey's HSD test further confirms that AU-guided MT-CLAR yields the best RMSE. Further, the CCC values of all three models are significantly different as revealed by a one-way ANOVA ($F(2,12) = 67.87, p < 0.05$), with AU-guided MT-CLAR yielding the best CCC. With fine-tuning, Raw MT-CLAR as base model yields the best RMSE and CCC, as shown by one-way ANOVA and confirmed

by Tukey's HSD test.

With respect to arousal estimation without fine-tuning, a one-way ANOVA shows a significant difference in the RMSE values of AU-guided and background-masked MT-CLAR as base model ($F(2,12) = 9.54, p < 0.05$). However, the CCC values of all three models are significantly different as shown by a one-way ANOVA, with AU-guided MT-CLAR as base model yielding best results as confirmed by Tukey's HSD test. With fine-tuning, a significant difference of RMSE values is observed only in the case of Raw and AU-guided MT-CLAR, while AU-guided MT-CLAR results in the best CCC values as shown by a one-way ANOVA and confirmed by Tukey's HSD test.

- **Random frame of a video:** Considering valence estimation without fine-tuning, and using random frame of a video as anchor set, AU-guided MT-CLAR as base model results in the best RMSE, as revealed by a one-way ANOVA ($F(2,12) = 12.81, p < 0.05$), and confirmed by a Tukey's HSD test. The CCC values obtained using all three MT-CLAR versions as base models are significantly different as shown by a one-way ANOVA ($F(2,12) = 40.31, p < 0.05$), with the best CCC obtained using AU-guided MT-CLAR as base model as confirmed by Tukey's HSD test. With fine-tuning, both RMSE and CCC values of Raw and background-masked MT-CLAR differ significantly as revealed by a one-way ANOVA.

  In the case of arousal estimation without fine-tuning, AU-guided MT-CLAR as base model yields the best RMSE and CCC values, as shown by a one-way ANOVA and further confirmed by Tukey's HSD test. In the fine-tuned case, the RMSE values of Raw MT-CLAR and AU-guided MT-CLAR as base models differ significantly as shown by a one-way ANOVA ($F(2,12) = 5.73, p < 0.05$). AU-guided MT-CLAR as base model results in the best CCC values as shown by a one-way ANOVA ($F(2,12) = 21.6, p < 0.05$), and further confirmed by a Tukey's HSD test.

- **Recurring 100^{th} frame of a video:** Considering valence estimation without fine-tuning, using recurring 100^{th} frame as the anchor set, RMSE values of background-masked MT-CLAR and AU-guided MT-CLAR differ significantly, as shown by a one-way ANOVA

($F(2,12) = 4.56, p < 0.05$). However, the CCC values of all three models differ significantly as shown by a one-way ANOVA ($F(2,12) = 32.41, P < 0.05$), and a Tukey's HSD test confirms that AU-guided MT-CLAR as base model yields the best result. Considering the fine-tuned case, the RMSE values are not significantly different, while Raw MT-CLAR as base model results in the best CCC, as shown by a one-way ANOVA ($F(2,12) = 12.07, p < 0.05$) and confirmed by a Tukey's HSD test.

In the case of arousal estimation without fine-tuning, AU-guided MT-CLAR as base model yields the best RMSE as revealed by a one-way ANOVA ($F(2,12) = 8.6, p < 0.05$), and confirmed by Tukey's HSD test. However, no difference is observed in the CCC values using the three base models. With fine-tuning, a similar trend as valence estimation is observed. The RMSE values are not significantly different, while the CCC values using Raw and AU-guided MT-CLAR as base models are significantly different as revealed by a one-way ANOVA ($F(2,12) = 4.2, p < 0.05$).

- **Recurring 50^{th} frame of a video:** In the valence estimation task without fine-tuning, and using the recurring 50^{th} frame in the anchor set, the RMSE values of background-masked and AU-guided MT-CLAR as base models are significantly different, as shown by a one-way ANOVA ($f(2,12) = 5.8, p < 0.05$). The best CCC value is obtained using AU-guided MT-CLAR as base model, as shown by a one-way ANOVA ($F(2,12) = 27.8, p < 0.05$), and confirmed by a Tukey's HSD test. With fine-tuning, the RMSE values obtained using the three base models are not significantly different, while the Raw MT-CLAR base model yields the best CCC, as shown by a one-way ANOVA ($F(2,12) = 18.93, p < 0.05$), and confirmed by a Tukey's HSD test.

  With respect to arousal estimation without fine-tuning, AU-guided MT-CLAR as base model yields the best RMSE as revealed by a one-way ANOVA ($F(2,12) = 40.67, P < 0.05$), and confirmed by a Tukey's HSD test. The CCC values obtained using the three base models differ significantly, as shown by a one-way ANOVA and confirmed by a Tukey's HSD test. With fine-tuning, the RMSE values obtained using the three base models are not significantly different, as shown by a one-way ANOVA. However, AU-

guided MT-CLAR as base model results in the best CCC, as shown by a one-way ANOVA ($F(2,12) = 13.61, p < 0.05$), and confirmed by a Tukey's HSD test.

- **Recurring 20^{th} frame of a video:** In the valence estimation task without fine-tuning, using recurring 20^{th} frame in the anchor set, the RMSE values obtained using the three MT-CLAR base models are not different as shown by a one-way ANOVA. The CCC values using background-masked MT-CLAR significantly differs from Raw and AU-guided MT-CLAR as base model, as revealed by a one way ANOVA ($F(2,12) = 17.21, p < 0.05$). With fine-tuning, the RMSE and CCC value of Raw MT-CLAR is significantly different from background-masked and AU-guided MT-CLAR as base model.

  In arousal estimation without fine-tuning, the RMSE and CCC value obtained using background-masked MT-CLAR as base model significantly differs from the other two values, as revealed by a one-way ANOVA. However, with fine-tuning, the RMSE and CCC values obtained using the base models are not significantly different, as revealed by a one-way ANOVA.

- **Recurring 10^{th} frame of a video:** Considering valence estimation without fine-tuning, and using the recurring 10^{th} frame as the anchor set, the RMSE value obtained using AU-guided MT-CLAR as base model is different from the other two values, as shown by a one-way ANOVA. AU-guided MT-CLAR results in the best CCC value, as revealed by a one-way ANOVA ($f(2,12) = 13.4, p < 0.05$) and confirmed using a Tukey's HSD test. With fine-tuning, Raw MT-CLAR as base model yields an RMSE and CCC value that is significantly different from the other two values, as revealed by a one-way ANOVA.

  Considering arousal estimation without fine-tuning, the RMSE value obtained using AU-guided MT-CLAR as base model is significantly different from the other two models, as shown by a one-way ANOVA. However, the CCC values of the three base models are not different. With fine-tuning, both RMSE and CCC values of the three models do not have a significant difference, as revealed by a one-way ANOVA.

Overall, using a random frame from a subject-independent video resulting in the worst RMSE and CCC values, and recurring 10^{th} frame resulting in the best RMSE and CCC, is a

Table 6.1: Few-shot affect inference on AFEW-VA with varying S configurations with *Raw MT-CLAR* as the base model. Best results in bold, while '*' denotes results significantly better than SOTA ($p < 0.05$) as per Kolmogorov-Smirnov (KS) test. † indicate % of total frames.

Row	A_S configuration	$\|S\|$ (†)	Shot	Fine-tuned	Valence				Arousal			
					RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Toisoul 21]	-	-	-	*0.23*	*0.70*	*0.69*	*0.65*	*0.22*	*0.67*	*0.66*	*0.81*
1	First frame of corresponding video	91 (2.02%)	One	*No*	*0.19* *	*0.68*	*0.68*	*0.59*	*0.21*	*0.66*	*0.64*	*0.78*
2				Yes	0.16	0.73	0.73	0.61	0.21	0.72	0.68	0.78
3	Random frame of corresponding video	91 (2.02%)	One	*No*	*0.19* *	*0.75* *	*0.73* *	*0.63*	*0.22*	*0.64*	*0.64*	*0.83*
4				Yes	0.14	0.83	0.82	0.63	0.19	0.74	0.72	0.84
5	Random frame from a subject-specific video	34 (0.76%)	One	*No*	*0.29*	*0.44*	*0.41*	*0.47*	*0.29*	*0.47*	*0.44*	*0.65*
6				Yes	0.28	0.46	0.42	0.46	0.24	0.44	0.44	0.80
7	Random frame from a video of different subject	34 (0.76%)	One	*No*	*0.56*	*0.01*	*0.00*	*0.11*	*0.36*	*0.01*	*0.01*	*0.82*
8				Yes	0.59	-0.14	-0.03	0.14	0.33	-0.01	-0.01	0.82
9	Recurring 100^{th} frame of corresponding video	96 (2.13%)	Few	*No*	*0.19* *	*0.67*	*0.67*	*0.6*	*0.22*	*0.65*	*0.62*	*0.78*
10				Yes	0.15	0.77	0.76	0.6	0.21	0.72	0.67	0.78
11	Recurring 50^{th} frame of corresponding video	132 (2.93%)	Few	*No*	*0.19* *	*0.69*	*0.69*	*0.59*	*0.21*	*0.67*	*0.66*	*0.79*
12				Yes	0.14	0.80	0.80	0.61	0.18	0.75	0.73	0.80
13	Recurring 20^{th} frame of corresponding video	268 (5.96%)	Few	*No*	*0.13* *	*0.85* *	*0.85* *	*0.66*	*0.15* *	*0.81* *	*0.81* *	*0.86* *
14				Yes	**0.09**	**0.92**	0.91	**0.66**	**0.13**	0.86	**0.86**	0.88
15	Recurring 10^{th} frame of corresponding video	494 (10.98%)	Few	*No*	*0.12* *	*0.88* *	*0.87* *	*0.64*	*0.15* *	*0.82* *	*0.81* *	*0.86* *
16				Yes	**0.09**	**0.92**	**0.92**	0.64	**0.13**	**0.87**	**0.86**	0.87
17	Recurring 10^{th} frame of corresponding video (mean)	494 (10.98%)	Few	*No*	*0.20* *	*0.73* *	*0.71*	*0.62*	*0.16* *	*0.78* *	*0.77* *	*0.89* *
18				Yes	0.18	0.76	0.75	0.63	0.15	0.81	0.81	**0.91**

general trend in both valence and arousal estimation across configurations. As increasing n increases the number of anchors of a video, it yields better results, as the valence and arousal values are predicted with the latest preceding anchor. Thus, the time-dependency of an anchor frame with respect to query frame is an important factor of consideration when building FSL-based automatic affect system. Additionally, in both valence and arousal estimation task, AU-guided MT-CLAR as base model resulting in the best RMSE and CCC is a noteworthy trend. This aligns with the prior findings that specific combinations and intensities of AUs are linked to different emotions [Lucey 10]. Thus, the proposed AU-guided base model proposed, focusing on specific parts of the input image based on the activated AUs is effective compared to its raw-image counterpart.

6.6.2 Comparison with State-of-the-art

Whilst there are no competing methods to this end, Table 6.1, Table 6.2, and Table 6.3 nevertheless presents interesting insights regarding the impact of the anchor set configuration on the precision of the valence and arousal estimates. The state-of-the-art method by Toisoul *et al.* [Toisoul 21] is compared with the proposed FSL-based method, with AU-guided MT-CLAR as the base model (Table 6.3). Results significantly better than SOTA, as per Kolmogorov-Smirnov test, are denoted via a '*'.

Table 6.2: Few-shot affect inference on AFEW-VA with varying S configurations with *Background-masked MT-CLAR* as the base model. Best results in bold, while '*' denotes results significantly better than SOTA ($p < 0.05$) as per Kolmogorov-Smirnov test.

Row	A_S configuration	Fine-tuned	Valence				Arousal			
			RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Toisoul 21]	-	*0.23*	*0.70*	*0.69*	*0.65*	*0.22*	*0.67*	*0.66*	*0.81*
1	First frame of corresponding video	*No*	*0.21*	*0.63*	*0.62*	*0.59*	*0.24*	*0.62*	*0.60*	*0.77*
2		Yes	0.21	0.64	0.64	0.59	0.18	0.72	0.70	0.82
3	Random frame from corresponding video	*No*	*0.21*	*0.70*	*0.68*	*0.62*	*0.25*	*0.66*	*0.62*	*0.81*
4		Yes	0.19	0.73	0.72	0.60	0.17	0.76	0.75	0.85
5	Random frame from a subject-specific video	*No*	*0.33*	*0.34*	*0.32*	*0.51*	*0.37*	*0.34*	*0.30*	*0.71*
6		Yes	0.40	0.31	0.26	0.47	0.29	0.42	0.40	0.73
7	Random frame from a video of different subject	*No*	*0.44*	*0.07*	*0.06*	*0.37*	*0.46*	*0.01*	*0.01*	*0.67*
8		Yes	0.38	0.16	0.15	0.47	0.37	0.00	0.00	0.67
9	Recurring 100^{th} frame of corresponding video	*No*	*0.21*	*0.60*	*0.59*	*0.59*	*0.24*	*0.63*	*0.60*	*0.76*
10		Yes	0.20	0.66	0.66	0.58	0.18	0.72	0.70	0.82
11	Recurring 50^{th} frame of corresponding video	*No*	*0.21*	*0.64*	*0.64*	*0.60*	*0.24*	*0.63*	*0.60*	*0.77*
12		Yes	0.19	0.70	0.69	0.59	0.17	0.75	0.74	0.83
13	Recurring 20^{th} frame of corresponding video	*No*	*0.16* *	*0.78* *	*0.77* *	*0.62*	*0.21*	*0.68*	*0.66*	*0.81*
14		Yes	0.17	0.77	0.76	0.61	0.14	0.83	0.83	0.88
15	Recurring 10^{th} frame of corresponding video	*No*	*0.13* *	*0.86* *	*0.85* *	*0.63*	*0.18* *	*0.77* *	*0.75* *	*0.84* *
16		Yes	0.15	0.83	0.83	0.62	0.12	0.87	0.87	0.88

The following remarks are made from Table 6.3. Focusing on the 'No' rows, measures very comparable to SOTA are obtained including a significantly better RMSE (valence) when only the first video frame is employed as anchor. Using any random video frame as anchor further improves measures compared to SOTA with significantly better RMSE, PCC and CCC metrics generated; note that only 2% of AFEW-VA are labelled in either case. Expectedly, poorer measures are noted in the limiting case when an anchor corresponding to the same subject ID is utilised as anchor, in which case $< 1\%$ of labelled AFEW-VA frames are used. The lowest measures are observed in the extremely challenging case where an anchor corresponding to a different ID is employed, with arousal-related metrics faring better than valence metrics. Overall, these results are testimony to the assertion that FSL-based emotion inference *in-the-wild* is highly difficult. Measures better than SOTA are obtained when multiple video frames are employed as anchors, significantly out-competing SOTA for all-but-one measure with as few as 5.96% labelled anchor frames.

The 'Yes' rows in Table 6.1 correspond to the condition where MT-CLAR is fine-tuned using AFEW-VA (4 out of 5 folds constituting the training set). It is noted that fine-tuned MT-CLAR models perform no better than MT-CLAR without fine-tuning for all-but-one A_S configurations. The notable exception is when FSL is attempted with an anchor frame corresponding to a different subject ID (row 8). This reveals that for a challenging task of FSL-labelling when different

Table 6.3: Few-shot affect inference on AFEW-VA with varying S configurations with *AU-guided MT-CLAR* as the base model. Best results in bold, while '*' denotes results significantly better than SOTA ($p < 0.05$) as per Kolmogorov-Smirnov test.

Row	A_S configuration	Fine-tuned	Valence				Arousal			
			RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Toisoul 21]	-	*0.23*	*0.70*	*0.69*	*0.65*	*0.22*	*0.67*	*0.66*	*0.81*
1	First frame of corresponding video	*No*	*0.15* *	*0.76* *	*0.75* *	*0.62*	*0.16* *	*0.77* *	*0.76* *	*0.82*
2		Yes	0.20	0.66	0.66	0.56	0.16	0.79	0.77	0.82
3	Random frame from corresponding video	*No*	*0.13* *	*0.85* *	*0.84* *	*0.67*	*0.14* *	*0.85* *	*0.84* *	*0.88* *
4		Yes	0.17	0.77	0.76	0.62	0.13	0.85	0.84	0.87
5	Random frame from a subject-specific video	*No*	*0.28*	*0.41*	*0.40*	*0.53*	*0.27*	*0.46*	*0.45*	*0.79*
6		Yes	0.33	0.45	0.40	0.46	0.26	0.48	0.47	0.75
7	Random frame from a video of different subject	*No*	*0.37*	*0.04*	*0.04*	*0.38*	*0.39*	*0.01*	*0.01*	*0.71*
8		Yes	0.37	0.21	0.19	0.45	0.33	0.22	0.21	0.68
9	Recurring 100^{th} frame of corresponding video	*No*	*0.15* *	*0.75* *	*0.74* *	*0.62*	*0.16* *	*0.76* *	*0.75* *	*0.81*
10		Yes	0.19	0.67	0.66	0.56	0.16	0.76	0.75	0.85
11	Recurring 50^{th} frame of corresponding video	*No*	*0.14* *	*0.79* *	*0.79* *	*0.62*	*0.15* *	*0.79* *	*0.78* *	*0.83*
12		Yes	0.18	0.73	0.72	0.58	0.14	0.81	0.80	0.84
13	Recurring 20^{th} frame of corresponding video	*No*	*0.10* *	*0.90**	*0.90* *	*0.66*	*0.12* *	*0.87* *	*0.87* *	*0.87* *
14		Yes	0.16	0.80	0.80	0.61	0.12	0.87	0.86	0.87
15	Recurring 10^{th} frame of corresponding video	*No*	***0.07*** *	***0.95*** *	***0.95*** *	*0.66*	***0.09*** *	***0.91*** *	***0.91*** *	***0.89*** *
16		Yes	0.13	0.86	0.85	0.63	0.10	**0.91**	**0.91**	**0.89**

subject is employed as an anchor frame, additional training (fine-tuning) is required for the specific dataset. Furthermore, inadequate MT-CLAR labelling performance when the anchor frame corresponds to a different subject ID conveys that pairwise expressive face comparisons become easier when identity-related facial variations are accounted for, and identity-related facial representations, capturing global facial structure, are utilised by the MT-CLAR network to make predictions relating to emotions, which are characterised by the local facial structure. Thus, the choice of anchor(s) for few-shot learning is nevertheless critical, and can considerably impact prediction results.

The best measures with MT-CLAR are obtained when the anchor set comprises recurring 20^{th} or 10^{th} frames in the video to be labelled. Human-level measures for valence and arousal are obtained in either configuration, implying that close-to-peak affective labelling performance is achieved with MT-CLAR even as only ≈6% frames in a video dataset are labelled.

6.6.3 MT-CLAR + SL Prediction on Images

For MT-CLAR to be used with images as in [Toisoul 21, Kossaifi 20b] and to validate the observation that contrastive learning generates robust, high-quality representations [Jing 19, Chen 20], MT-CLAR is combined with with supervised learning in the MT-CLAR + SL architecture (Figure 5.2). When applied on images, it estimates their (a) emotion category labels, (b) valence,

Table 6.4: Comparison with state-of-the-art studies on AffectNet. Best results are in bold, while the second best are underlined.

Methods	Accuracy	Valence				Arousal			
		RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
[Mollahosseini 19]	0.58	0.37	0.66	0.60	0.74	0.41	0.54	0.34	0.65
[Jang 19]	-	0.44	0.58	0.57	0.73	0.39	0.50	0.47	0.71
[Kollias 18b]	0.60	0.37	0.66	0.62	0.78	0.39	0.55	0.54	0.75
[Toisoul 21]	**0.62**	**0.33**	**0.73**	**0.73**	**0.81**	**0.30**	**0.65**	**0.65**	**0.81**
MT-CLAR + SL	0.56	0.36	0.67	0.67	0.78	0.32	0.60	0.60	**0.81**

and (c) arousal labels. Results for MT-CLAR + SL and prior methods on the AffectNet dataset are presented in Table 6.4.

For categorical emotion labelling, MT-CLAR's performance is comparable to other models. For valence estimation, MT-CLAR achieves the second-best performance with respect to four valence metrics, including an equal second-best SAGR as [Toisoul 21]. For arousal, the second-best performance is achieved with respect to three metrics, and an equal-best SAGR as [Toisoul 21]. These results confirm that MT-CLAR + SL predictions are comparable to the state-of-the-art.

Table 6.5 presents continuous valence and arousal labelling results for MT-CLAR + SL on the AFEW-VA video dataset and comparisons with SOTA. Two 5FCV data-split strategies are considered for the AFEW-VA dataset: subject-dependent and subject-independent. While both involve mutually exclusive training and test sets, the subject-independent setting also involves mutually exclusive subject IDs so as to preclude a *data leak* from the training sets to the test set.

Observing Table 6.5, the following remarks are made. Consistent with Table 5.2, predictions in the subject-dependent setting are much better than those in the subject-independent setting. In comparison to other models, MT-CLAR + SL achieves the lowest valence RMSE and the second-best PCC. For arousal, the lowest RMSE, and the second-best PCC and SAGR are obtained. Considering subject-dependent splits, the CNN-RNN-based framework proposed in Section 4.3.1 is outperformed with respect to valence RMSE and PCC, and obtain an identical CCC.

Tables 6.4 and 6.5 reveal that MT-CLAR + SL, designed to enable the MT-CLAR model to predict both categorical *and* continuous emotion labels for singleton images, is highly compet-

Table 6.5: Comparison with SOTA on AFEW-VA with 5FCV. Best results for each metric in the SI condition are denoted in bold, and second-best underlined. Best results in the SD condition are in bold.

Data-split strategy	Method	Valence				Arousal			
		RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC↑	SAGR ↑
Subject-independent	[Kossaifi 17]	0.27	0.41	-	-	0.23	0.45	-	-
	[Mitenkova 19]	0.40	0.33	-	-	0.41	0.42	-	-
	[Handrich 20]	0.28	0.58	-	-	0.26	0.46	-	-
	[Kollias 18b]	0.48	0.56	-	-	0.27	0.61	-	-
	[Kossaifi 20b]	0.24	0.55	0.55	0.64	0.24	0.57	0.52	0.77
	[Toisoul 21]	0.23	**0.70**	**0.69**	**0.65**	0.22	**0.67**	**0.66**	**0.81**
	MT-CLAR + SL	**0.21**	0.69	0.46	0.58	**0.19**	0.62	0.42	0.78
Subject-dependent	CNN-LSTM (Section 4.4)	0.13	0.89	0.89	-	0.12	**0.93**	**0.93**	-
	MT-CLAR + SL	**0.12**	**0.90**	0.89	0.67	0.12	0.88	0.88	0.87

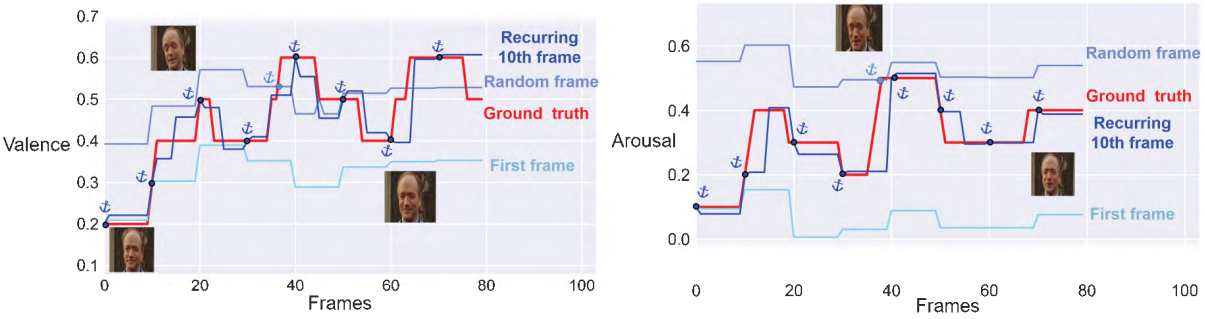


Figure 6.6: Continuous affect prediction in videos: Visualisation of the predicted valence (left) and arousal (right) values with multiple A_S configurations, namely first frame, random frame, and recurring 10^{th} frame in an exemplar AFEW-VA video.

itive compared to other models exclusively designed to this end.

6.6.4 Qualitative Analysis

The empirical results confirm that MT-CLAR (a) enables accurate annotation of continuous valence and arousal values in videos when a labelled support-set is available (Table 6.1), and (b) achieves competitive valence and arousal level estimation for singleton images (Table 6.4) and video frames (Table 6.5). Labelling images for emotion category, valence, and arousal enables automated emoji generation [Ali 17], while dynamic affect labelling in videos greatly eliminates human effort and bias, and enables applications such as highlights detection [Qi 21].

Still, precisely estimating continuous valence, arousal levels from *in-the-wild* videos presents a significant challenge, even if excellent RMSE, PCC, CCC and SAGR metrics are achieved over the test set (see Figure 6.6). The figure presents true and MT-CLAR-predicted valence

(left) and arousal (right) values with multiple A_S configurations. As per Table 6.1, valence and arousal estimates employing the first video frame as anchor already compare well with SOTA [Toisoul 21]; however, a considerable gap between true valence and arousal levels, and first frame-based MT-CLAR predictions can be noted for a majority of the considered video.

Consistent with Table 6.1 results, it is noted that the true valence and arousal trends are captured better with a random frame anchor. MT-CLAR predictions employing recurrent 10^{th} frames as anchor are far better than [Toisoul 21] from Table 6.1; whilst correspondingly, the dark blue curve best aligns with the (true) red curve for both valence and arousal prediction, the estimates are still far from precise. Overall, Figure 6.6 clearly reveals that the RMSE, PCC, CCC and SAGR metrics are rather coarse-grained for the arduous problem of dynamic emotion inference, and even excellent results achieved with respect to these measures does not imply generation of precise estimates. Thus, worthy objectives for future work in this direction would be to (1) attempt precise affect predictions with few annotations, and (2) explore alternate performance metrics to better validate the precision of estimates.

6.7 Conclusion

Addressing the challenges of annotating video datasets with affect, in this chapter, a Few-Shot Learning-based approach is proposed as an alternative for efficient affect labelling. This is performed using the MT-CLAR models developed in Chapter 5, which captures affect (dis)similarity and affect differences between a pair of input images. The three versions of MT-CLAR (Raw, background-masked, AU-guided) are used as base models for the proposed approach. The proposed FSL-based approach considers an *Anchor Set* which comprises frames with ground-truth valence/arousal annotations, and computes the affect difference between the query frame (for which valence has to be estimated) and the frames in the anchor set using MT-CLAR base models. Extensive experiments are performed using various configurations of the anchor set and various MT-CLAR base models. The results reveal that best performance in terms of RMSE and CCC is achieved using AU-guided MT-CLAR as the base model.

As an additional downstream task, MT-CLAR + SL, a supervised learning framework is pro-

posed to assess the robustness of the affect representations obtained from the Siamese Network of MT-CLAR. Using MT-CLAR + SL, the classification of discrete emotions, and prediction of valence and arousal values on the AFEW-VA dataset is performed. A limitation of the proposed MT-CLAR approach is that it considers static input for training MT-CLAR models. Although considering the anchor set configurations in the proposed FSL-based approach provides a temporal perspective, as MT-CLAR is devoid of temporal input, the affect representations obtained from MT-CLAR are not temporally-aware, which will be addressed in the next chapter.

Chapter 7

Time-continuous Affect Representation

Contents

Chapter 5 demonstrated the significance of modelling affect differences, rather than using the affect annotations directly, and Chapter 6 indicated the importance of developing robust affect representations, while additionally using affect difference for efficient labelling task. These chapters considered static data for training the affect inference base models. This chapter aims to draw attention towards the prominence of temporal data for a more comprehensive analysis of affect. With temporal data, affect estimation models can grasp the evolution of emotions, understand context shifts, and discern intricacies that might not be evident in static data.

7.1 Introduction

While the duration of emotion is a long-standing debate, the leading emotion theories such as the appraisal theories [Moors 14], and constructionist theories [Barrett 14] consent the inherent *dynamic* nature of emotions, and their gradual progression. Multiple studies argue that a complete understanding of emotions can be achieved only when the dynamic nature of emotions is taken into account [Eaton 02, Hemenover 03]. This dynamic nature fundamentally serves as alerts to significant changes and events, compelling us to respond to these shifts [Scherer 09].

The temporal nature of emotions is crucial for accurate emotion recognition as it captures the dynamic evolution of emotional experiences over time. Early studies on automatic affect inference primarily focused on using images for analysing facial affect using simpler techniques and feature extraction methods. The universality of certain facial expressions across cultures made facial images a convenient means to study emotions [Ekman 71]. Additionally, studies often relied on manually designed algorithms to detect facial landmarks, identify basic expressions, and recognise patterns associated with emotions. The design of ML algorithms and computational capabilities favored the analysis of static images over complex data types such as video or real-time streaming data. While studies employing static data have laid a strong foundation for research in automatic affect inference, it is essential to consider temporal data, such as videos, as emotions manifest gradually over time.

With an attempt to efficiently label videos, Chapter 5 first proposed MT-CLAR, a metric learning-based framework to learn affect differences, and later performed FSL-based labelling using MT-CLAR as base model. The FSL-based approach is performed using the valence (arousal) values of the anchor frame and valence (arousal) differential obtained through MT-CLAR. This approach is temporally-aware as the valence (arousal) values of the anchor frame encodes the temporal information (for instance, first frame of the video being considered as the anchor frame translates to time $t = 0$, using recurring 100^{th} frame as anchor frame corresponds to considering anchors periodic with respect to time, as depicted in Figure 6.3). However, the affect representations learnt through MT-CLAR are temporally unaware, as MT-CLAR takes static data as input without considering temporal information. This chapter proposes *Temporal*

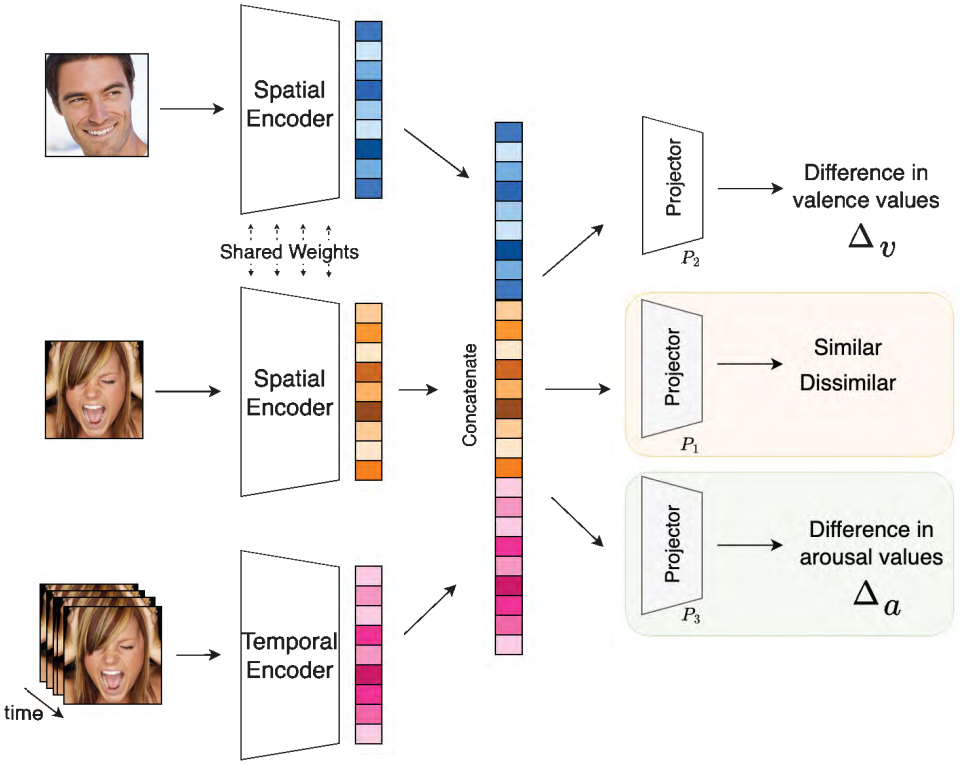


Figure 7.1: Temporal MT-CLAR overview: Similar to MT-CLARs proposed in Chapter 5, a pair of expressive facial images is passed through a Siamese network, along with the temporal information (*clip*) of one of the images in the pair to a non-local neural network as its temporal encoder. All the embeddings obtained are then concatenated to estimate (1) whether the expressions are similar/dissimilar, (2) the valence differential (Δ_v), and (3) the arousal differential (Δ_a) between expressions.

MT-CLAR, which comprises a metric-learning based Siamese network to capture spatial information, and non-local operations for capturing long-range temporal dependencies. An overview of the Temporal MT-CLAR model is shown in Figure 7.1. Different from the base models proposed previously, Temporal MT-CLAR learns affect representations which are aware of the temporal context. These representations are further employed for downstream tasks.

7.2 Key Contributions

The key contributions presented in this chapter are as follows:

- *Temporal MT-CLAR*, a Multi-Task non-local neural network with Contrastive Learning is proposed for Affect Representation. Temporal MT-CLAR aims to learn affect differences between a pair of images, by considering additional temporal information of one of the images.
- To model temporal dynamics effectively, a non-local neural block is employed within the ResNet-50 architecture to capture long-range dependencies and model global interactions in the input clip.
- An FSL-based approach is proposed for time-continuous frame-level affect labelling in videos, using the affect differences obtained from Temporal MT-CLAR.
- With $\approx 1\%$ frames of the dataset, results exceeding SOTA are achieved. While with $\approx 10\%$ of total frames, near-ceiling[1] performance is achieved in both valence and arousal estimation.

7.3 Prior Works

Temporal data provides additional cues as compared to static data, thereby contributing to a richer understanding of emotions compared to analysing isolated frames. Studies have em-

[1] "Ceiling" refers to the upper bound of CCC value, which is 1. For the anchor set configuration of recurring 10^{th} frame of corresponding video (Table 7.3), a CCC of 0.96 is obtained, therefore referred as "near-ceiling" value.

ployed fully-connected DNNs to achieve dimensionality reduction on high-dimensional hand-crafted frame-level features. [Zhang 16a] and [Ranganathan 16] use Deep Belief Network to learn from aggregated facial feature point trajectories. RNNs are well-suited for modelling frame-level spatial features, where high-level facial features are extracted for each frame, and are then fed as sequential input to RNN. Modern studies combine RNNs with deep features from the last layer of a CNN trained for affect inference [Brady 16, Kim 17]. In studies incorporating CNN-LSTM architecture, for instance [Tzirakis 17, Khorrami 16], CNNs capture local spatial features within frames, while LSTMs handle temporal relationships between these features over time. Authors in [Chen 17] compare the performance of a non-temporal SVR model and temporal LSTM-RNN, and observe that the temporal model alleviates the feature engineering efforts and improves affect inference performance. An ensemble methodology applied in [Kollias 21], where the features extracted from CNNs are fed to multiple RNNs, outperforms the traditional CNN-RNN where the features are fused fed to a single RNN.

In [Mao 19], continuous emotion is inferred from video sequences using bi-directional LSTM model by learning the hierarchical context information. An encoder-decoder network is employed in [Lee 18], where the spatial features of each frame are extracted using 2D-CNNs, and the spatio–temporal attention is estimated using convolutional LSTM. To perform temporal modelling from both audio and visual modalities, authors in [Chao 14] use temporal pooling functions in the deep neutral network, and fuse the results from various modalities at the decision-level. In [Meng 22b], a transformer-based encoder is used to capture the temporal context information, with fully connected layers as prediction heads for affect inference. Extending beyond self-attention, authors in [Praveen 22] propose a joint cross-attention fusion model to exploit inter-modal relationships, and encode the visual temporal information with 3D-CNN. A constrained representation learning is proposed in [Tellamekala 19] to learn the 'temporally coherent' latent features, through a regularised contrastive loss based on the temporal coherency principle [Hurri 02]. The principle states that when processing temporal input, the representation changes as little as possible over time. In [Kossaifi 20b], the temporal information is learned through transduction of spatial information learnt from tensor factorisation technique.

7.4 Proposed Framework

7.4.1 Temporal MT-CLAR

Spatial encoder

To encode spatial information, a variant of Face Alignment Network, called EmoFAN [Toisoul 21] is used as a spatial encoder, $SpEnc(\cdot)$. EmoFAN [Toisoul 21] is trained to jointly predict categorical emotion, dimension affect and fiducial landmarks from an input face image. Similar to MT-CLAR proposed in Chapter 5, a pair of input images, x_1 and x_2 are given as an input to $SpEnc(\cdot)$, and are mapped to its penultimate layer's output, yielding the corresponding representation vectors, r_1 and r_2, given by $r_1 = SpEnc_1(x_1)$, and $r_2 = SpEnc_2(x_2)$, where $r_1 \in \mathbb{R}^{D_1}$ and $r_2 \in \mathbb{R}^{D_2}$ with $D_1 = D_2 = 256$. In the proposed model, $SpEnc_1(\cdot)$ and $SpEnc_2(\cdot)$ in the two streams share the parameters and weights to produce two features corresponding to the images.

Temporal encoder

Since videos are inherently dynamic and contextually rich, temporal information is modelled concurrently with spatial information. Spatio-temporal concurrent modelling is a widely studied topic in the field of computer vision for various tasks such as action recognition and face recognition [Ji 13, Hassani 17]. Prior works have established that the fusion of spatial and temporal features in video encoding yields superior results compared to methods that treat these dimensions in isolation [Tellamekala 19, Handrich 20]. To this end, both convolutional and recurrent neural networks encode local neighbourhood information either in space or time dimension. 3D-CNNs exhibit computational intensity, with increased cost and memory requirements compared to their 2D counterparts [Tran 18]. This is due to the additional temporal dimension, making training and inference resource-intensive, and posing challenges in scalability, especially in resource-constrained environments. Additionally, training large networks with limited data increases the risk of overfitting and hinders generalisation, especially for a challenging task like dimensional affect estimation. On the other hand, RNNs capturing long-range dependencies due to the vanishing gradient problem, where gradients diminish exponentially during

backpropagation through time [Hochreiter 98]. This impedes the effective learning of temporal dependencies over extended sequences, limiting the capacity of traditional RNNs to capture complex patterns.

Amidst the challenges posed by 3D-CNNs and RNNs in effectively capturing temporal dependencies for video processing, Non-local Neural Networks (NLNN) have emerged as a promising alternative [Wang 18]. Unlike traditional architectures that rely on repeated local operations to propagate information across temporal dimensions, NLNN introduces a non-local operation that enables the model to capture long-range dependencies in a more holistic manner. This non-local operation, inspired by the non-local means algorithm in computer vision [Buades 05], allows each element in the sequence to directly interact with all other elements, facilitating the identification of intricate temporal relationships.

The non-local operation in NLNN [Wang 18] is defined as:

$$v_i = \frac{1}{\mathcal{C}(u)} \sum_{\forall j} f(u_i, u_j) g(u_j), \tag{7.1}$$

where u is the input signal and v is the output signal (same size as u). i is the index of an output position, and j is the index that enumerates all possible i. $f(u_i, u_j) \in \mathbb{R}$ is a pair-wise scalar function which computes relationship between u_i and u_j, and g is a representation of the input u at j. The summation is normalised by a factor $\mathcal{C}(u)$.

Equation 7.1 is said to be non-local due to its computation over all j to obtain the output signal. The unary function g is obtained through a linear embedding, given by

$$g(u_j) = W_g u_j, \tag{7.2}$$

where W_g is a learnable weight matrix. For f, following [Shokri 20], the Gaussian function in an embedded space is considered, given by

$$f(u_i, u_j) = e^{\theta(u_i)^T \phi(u_j)}, \tag{7.3}$$

where $\theta(u_i) = W_\theta u_i$ and $\phi(u_j) = W_\phi u_j$ are two embeddings. $\mathcal{C}(u)$ is set as

$$\mathcal{C}(u) = \sum_{\forall j} f(u_i, u_j). \tag{7.4}$$

Finally, the non-local operation in Equation 7.1 is defined as a neural network block as:

$$z_i = W_z v_i + u_i, \tag{7.5}$$

where W_z is a learnable weight matrix, v_i is from Equation 7.1, and "$+u_i$" denotes a residual connection. As this non-local block can be plugged to any major computer vision architectures, ResNet-50 [He 16] is chosen, depicted in Table 7.1. As done in [Wang 18, Shokri 20], non-local block is added into ResNet-50 architecture by making it a *non-local neural network*. For this study, NLNN is considered as a temporal encoder, $TeEnc(\cdot)$. The input video clip $\hat{x}_2$, with the last frame of the clip being x_2, is given as an input to $TeEnc()$, which are mapped to the penultimate layer's output to yield a representation vector, $\hat{r}_2$, given by $\hat{r}_2 = TeEnc(\hat{x}_2) \in \mathbb{R}^{D_3}$. For the current study, D_3 is set as 512 to match the cumulative dimension of the representation vector from spatial encoders. The implementation details are described in the next section.

Projector network

The vectors r_1, r_2, and $\hat{r}_2$ obtained from $SpEnc_1(\cdot)$, $SpEnc_2(\cdot)$, and $TeEnc$ respectively are concatenated to obtain $w = r_1 \parallel r_2 \parallel \hat{r}_2$, where $w \in \mathbb{R}^{1024}$. As done in Chapter 5, to perform the three tasks of classifying the input image pairs as (dis)similar, predicting Δ_v, and Δ_a, w is fed as input to three branched projector networks, $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$, which map w to three vectors $w_1 = P_1(w), w_2 = P_2(w)$ and $w_3 = P_3(w)$. $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$ are all MLPs, with four identical FC layers, while differing in the number of neurons in the last FC layer. The four FC layers in $P_1(\cdot), P_2(\cdot)$, and $P_3(\cdot)$ comprise 2048, 1024, 512, and 128 neurons, however, $P_2(\cdot)$ and $P_3(\cdot)$ have a single neuron to predict Δ_v and Δ_a respectively, while $P_1(\cdot)$ has 2 neurons to classify (dis)similarity in the last FC layer. Prior to feeding to each FC layer, the input is normalised with zero mean and unit standard deviation and activated using ReLU activation. The overall

Table 7.1: Architectural details of the temporal encoder used for video encoding, without the non-local block. Residual blocks are shown within the square brackets.

Layer name	Layer
conv_1	7×7, 64, stride 2, 2, 2
pool_1	$3 \times 3 \times 3$ max, stride 2, 2, 2
res_2	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$
pool_2	$3 \times 1 \times 1$ max, stride 2, 2, 2
res_3	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$
res_4	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$
res_5	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$
global average pool, FC	

architecture is henceforth referred as *Temporal MT-CLAR*, as the base architecture is based on the ideas proposed in Chapter 5.

7.4.2 Few-shot Labelling

To efficiently label unlabelled frames in a video, the few shot learning-based method proposed in Chapter 6 is extended. The Temporal MT-CLAR base model learnt with rich temporal characteristics is utilised to estimate valence and arousal values for a frame in the video, as shown in Figure 7.2. Similar to the proposed method in the previous chapter, a support set S which has frames with known valence and arousal values is considered, with the aim of predicting the valence and arousal value for a query frame. For a given query frame x_q, an *Anchor set* A_S, is formed as a subset of S. For a pair of frames, x_i (from A_S) and x_q, and a history of frames for the anchor frame x_q, denoted as $\hat{x}_q$, Δ_v is obtained using Temporal MT-CLAR as:

$$\Delta_v = y_{i_{val}} - y_{q_{val}}, \tag{7.6}$$

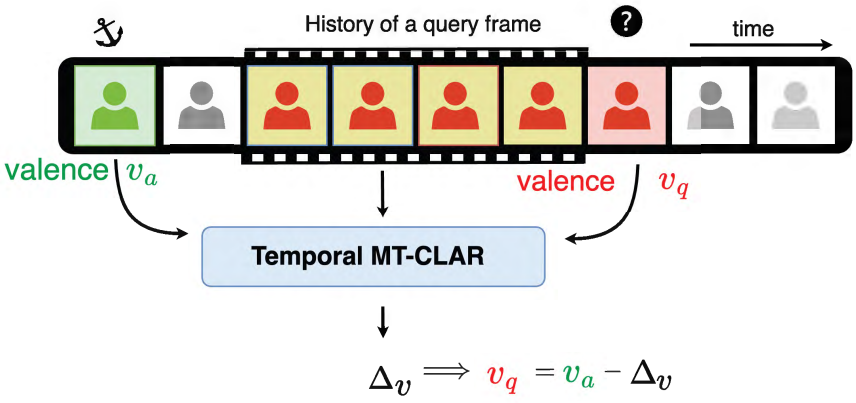


Figure 7.2: Few-shot learning through Temporal MT-CLAR: Given an *anchor* video frame (green) whose valence/arousal rating is known, a *query* frame (red), and a clip comprising antecedent frames of the query frame (yellow), Temporal MT-CLAR predicts valence/arousal values of the query frame.

where $y_{i_{val}}$ is the valence of x_i. Hence, $y_{q_{val}}$ is given by,

$$y_{q_{val}} = y_{i_{val}} - \Delta_v. \tag{7.7}$$

Similarly, $y_{q_{asl}}$ can be derived using Δ_a.

For frame-level labelling of valence and arousal values in a video, an anchor set $A_S \subset S$ is used. Various configurations of A_S are considered, similar to those described in Section 6.4.1, with an additional clip considered for the query frame.

It is worth noting that the availability of ground truth valence/arousal values is not assumed for any of the *historical* frames of x_q. The availability of annotated valence/arousal values is only assumed for the anchor frame, identical to the FSL-based method proposed earlier in the thesis. In contrast to the approaches detailed in the preceding chapter, the variations in facial movements that contribute to the information in the query frame is leveraged, without the need for additional annotations for the supplementary frames considered. Thus, the proposed FSL-based method leverages the representations learnt through metric-learning based Siamese network for spatial encoding and non-local neural network for temporal encoding for better estimating valence and arousal values based on its dynamics in the video.

7.5 Experimental Setup

In this chapter, unless otherwise specified, AffectNet is used to pre-train the spatial encoder, while Aff-Wild2 is used to train Temporal MT-CLAR end-to-end and is used to perform FSL to infer valence and arousal values for each frame in the test videos. To comprehensively analyse the temporal dynamics of emotions in videos and investigate the impact of temporal modelling, this chapter leverages long-range video datasets, particularly focusing on Aff-Wild2, as opposed to AFEW-VA, which primarily comprises shorter video sequences.

All the implementations are done using PyTorch [Paszke 19] software. The models are trained using an NVIDIA A100 with 40GB memory. Since the test set of Aff-Wild2 is not released, the validation set is used to evaluate the framework. The face cropped and aligned images provided by the dataset creators are used as input to our models. The input images for Temporal MT-CLAR are center-cropped to 256×256. As an augmentation, following the implementations from [Toisoul 21, Bulat 22], random affine transformation is applied on the training images with a rotation of up to 20 deg, translations up to 20% on both directions, scaling up to 20% up or down, and shearing up to 10 degrees. A horizontal flip is performed with a chance of 50%. To implement $TeEnc(\cdot)$, one non-local block is used after res_3 (see Table 7.1) in ResNet-50. The input to the temporal encoder, $\hat{x}_2$, is a *clip* of T frames. In order to optimise the utilisation of computational resources, instead of considering all the preceding T frames of x_2, dilated sampling is used for selecting 1 in 4 frames. This results in providing more temporal context to the model, with no additional computation overhead. For all the experiments, $T = 16$ frames is considered. Each frames are resized to 224×224 thus making the input size to the temporal encoder as $16 \times 224 \times 224 \times 3$. As the input videos provided in the dataset are recorded at 30 frames per second (fps), the following setup yields temporal context of 2.13 seconds. Temporal MT-CLAR is trained for 20 epochs with a batch size of 64 using Adam [Kingma 14] optimiser. Learning rate is initialised to 0.001 and is decreased by a factor of 10 for every 15 epochs. A cumulative loss function $\mathcal{L}$ given in Equation 5.2 is used. Following the same setup as in Chapter 5, the margin m used in contrastive loss (see Equation 5.1) is empirically set to 0.25. In the dynamic weight functions f and g (see Equation 4.2), used to

Table 7.2: Comparison of results of different configurations of base models on Aff-Wild2 dataset. 'Transfer learning' indicates the model trained on AffectNet (Chapter 5) and tested on Aff-Wild2. Models where spatial encoder's weights are initialised from Chapter 5 (trained on AffectNet) are denoted as 'fine-tuned'.

Input	Method	Δ Valence				Δ Arousal				Similarity accuracy ↑
		RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	
Image pair	Transfer learning	0.34	0.23	0.18	0.58	0.24	0.23	0.23	0.58	0.57
Image pair	Scratch MT-CLAR	0.37	0.47	0.47	0.68	0.25	0.20	0.20	0.65	0.77
Image pair	Fine-tuned MT-CLAR	0.26	0.64	0.61	0.72	0.19	0.56	0.54	0.54	0.71
Image pair + Clip	MT-CLAR + ResNet	0.35	0.41	0.39	0.67	0.24	0.31	0.30	0.67	0.75
Image pair + Clip	Fine-tuned MT-CLAR + ResNet	0.26	0.66	0.65	0.72	0.24	0.34	0.34	0.70	0.82
Image pair + Clip	MT-CLAR + Non-local ResNet	0.27	0.67	0.67	0.70	0.19	0.45	0.39	0.67	0.75
Image pair + Clip	Fine-tuned MT-CLAR + Non-local ResNet	**0.22**	**0.76**	**0.73**	**0.75**	**0.16**	**0.69**	**0.61**	**0.73**	**0.85**

compute $\mathcal{L}_{\Delta_v}$ and $\mathcal{L}_{\Delta_a}$ (see Equation 5.2), the fine-tuned hyper-parameters are $k \in \{1,2,3\}$ and $\alpha \in \{1,2,20\}$.

7.6 Results and Discussion

Temporal MT-CLAR takes an image pair and a few *history* frames of one of the images in the image pair as an input, and predicts (dis)similarity between the image pair, Δ_v, and Δ_a values. The model comprises Siamese network as a spatial encoder, a NLNN as a temporal encoder, and a multi-task architecture for predicting similarity and affect differentials. The model is trained end-to-end optimising the cumulative sum of contrastive loss, a dynamically weighted regression loss (defined in Chapter 4), and cross-entropy loss. Chapter 5 presented the analysis of the various components of the model mentioned above. This chapter specifically looks at the efficacy of the *temporally-aware* learned representations. Table 7.2 highlights the impact of temporal modelling using a non-local neural network encoder, in comparison to its static counterpart, on similarity and valence, arousal differentials. All experiments are performed on Aff-Wild2, a large-scale video dataset.

As shown in row 1 of Table 7.2, a baseline is obtained by employing a pre-trained network from Chapter 5, where the MT-CLAR base model is trained on AffectNet. The similarity accuracy achieved is 0.57, which is slightly better than chance (0.50). Due to domain mismatch

phenomenon [Yosinski 14], this outcome is sub-optimal, as the model is trained on AffectNet, an image-based dataset, and subsequently tested on Aff-Wild2, a video-based dataset. The deficiency arises from the absence of temporal processing in the model. To establish an equitable benchmark, it is trained (from *scratch* or initialised using the weights of the trained network from Chapter 5) on the Aff-Wild2 dataset. All-but-one metrics are better for Δ_v and Δ_a estimation for the fine-tuned model, with similarity accuracy as 0.71, an increase of 42% from chance.

Further, include a clip is included in addition to the input image pair, with ResNet-50 as the temporal encoder. A drop in the performance is observed when the model is trained from scratch as compared to using a fine-tuned MT-CLAR (for spatial encoding), as can be seen in rows 4 and 5 of Table 7.2. For a challenging task of affect estimation, it is posited that emotion dynamics are related non-locally in spacetime. To this end, non-local neural network as a temporal encoder resulted in the best performance with an accuracy of 0.85 for (dis)similarity classification, and an impressive CCC of 0.73 and 0.61 for valence and arousal differential estimation, as shown in rows 6 and 7. This is attributed to non-local neural network's ability to capture long-range dependencies and model global interactions within temporal sequences [Wang 18]. A robust base model, like Temporal MT-CLAR, can be used for a variety of related downstream tasks such as categorical emotion inference, dimensional affect estimation, etc. Following are the results when Temporal MT-CLAR is applied for the estimation of dimensional affect in a setting where limited labelled data is available.

Table 7.3 presents the results on valence and arousal estimation using the proposed few-shot learning-based method described in Section 7.4.2. A consistent trend observed with various anchor set configurations is the superior performance of the model when temporal branch is included, as compared to the counterpart without temporal branch. The worst RMSE and CCC is observed when the anchor set comprises the first frame of the clip. Since the duration of the videos in Aff-Wild2 is long (up to 26.22 minutes), there is a subsequent increase in the temporal distance between the first frame and the query frame. Although the sample is of the same subject, this increasing distance results in glaring difference between the pair of images. However, considering a recurring 1000^{th} frame in the anchor set, which is a mere increase of $\approx$300 anchor

Table 7.3: Few-shot affect inference on the validation set of Aff-Wild2 with varying S configurations with *Temporal MT-CLAR* as the base model. † indicate % of total frames.

Row	A_S configuration	$\|S\|$ (†)	Temporal	Valence				Arousal			
				RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA ([Tellamekala 22])	-	Yes	-	-	0.56	-	-	-	0.65	-
1	First frame of corresponding video	71 (0.02%)	No	0.43	0.22	0.18	0.52	0.39	0.21	0.15	0.64
2			Yes	0.39	0.25	0.24	0.57	0.35	0.36	0.30	0.68
3	Random frame of corresponding video	71 (0.02%)	No	0.38	0.57	0.56	0.79	0.31	0.40	0.40	0.84
4			Yes	0.36	0.59	0.58	0.80	0.29	0.46	0.46	0.89
5	Random frame from a video of different subject	71 (0.02%)	No	0.55	0.30	0.17	0.32	0.29	0.24	0.19	0.97
6			Yes	0.41	0.35	0.33	0.74	0.24	0.33	0.30	0.97
7	Recurring 1000^{th} frame of corresponding video	375 (0.11%)	No	0.35	0.60	0.60	0.78	0.31	0.41	0.40	0.83
8			Yes	0.33	0.62	0.61	0.79	0.28	0.45	0.44	0.88
9	Recurring 500^{th} frame of corresponding video	713 (0.21%)	*No*	0.32	0.66	0.66	0.81	0.29	0.46	0.45	0.86
10			Yes	0.30	0.68	0.68	0.80	0.25	0.53	0.51	0.91
11	Recurring 250^{th} frame of corresponding video	1390 (0.41%)	No	0.30	0.71	0.71	0.84	0.26	0.55	0.55	0.88
12			Yes	0.29	0.76	0.74	0.85	0.24	0.60	0.60	0.93
13	Recurring 100^{th} frame of corresponding video	3424 (1.01%)	No	0.25	0.80	0.80	0.88	0.22	0.68	0.67	0.93
14			Yes	0.22	0.83	0.81	0.84	0.20	0.72	0.71	0.94
15	Recurring 10^{th} frame of corresponding video	33908 (10.01%)	*No*	0.12	0.95	0.95	0.93	0.11	0.92	0.92	0.94
16			Yes	**0.11**	**0.96**	**0.96**	**0.94**	**0.10**	**0.93**	**0.93**	**0.96**

frames, improves the affect estimation performance considerably. This improvement can be attributed to the vicinity of the anchor frame with the query frame, as compared to considering only the first frame as an anchor. Further, as the frequency of the recurring frame increases, the performance improves, which aligns with the results obtained using the non-temporal MT-CLAR base models, as shown in Table 6.1, Table 6.2, and Table 6.3. With recurring 10^{th} frame as the anchor, a near-ceiling performance is achieved. With only $\approx 10\%$ size of the validation set, excellent results are achieved across all metrics for both valence and arousal. This holds a profound significance in labelling video datasets, as maximum affect labelling can be performed automatically with minimal human intervention. Considering the validation set of Aff-Wild2, obtaining human annotations for $\approx 10\%$ of the frames yields affect labels for the remaining $\approx 90\%$ that is on par with human performance (PCC = 0.96).

To the best of our knowledge, there are no FSL-based approaches for facial affect estimation for a direct comparison. Nevertheless, the performance of the proposed approach is compared with [Tellamekala 22], as they use temporal context for affect modelling from faces. As can be seen in Table 7.3, with respect to valence estimation, when random frame of the video, and recurring frames are considered as anchor set configurations, measures very comparable or higher than SOTA are achieved. For arousal estimation, the anchor set configuration of recurring 100^{th} frame and 10^{th} frame yields measures higher than SOTA.

7.7 Conclusion

The temporal dynamics of affect is well-established in the psychology literature. Most of the computational studies aiming at facial affect estimation, use static data for modelling affect. The previous chapters presented MT-CLAR, a base model which determines (dis)similarities, and affect differences between image pairs. Affect representations are learnt in the process, which are used for downstream tasks. This chapter focuses on developing the base MT-CLAR model, by further incorporating a non-local neural network as the temporal encoder, which takes a clip as input for modelling the temporal dynamics of affect. This study presents a *Temporal MT-CLAR*, which learns temporally-aware affect representations, and affect differences. Additionally, an FSL-based approach is presented, which considers an anchor set comprising a anchor frame and a clip for determining the affect of the query frame. The results obtained reveal that temporal modelling enhances affect estimation, and with $\approx 1\%$ of the total frames, measures higher than SOTA are achieved for both valence and arousal estimation. Further, a near-ceiling performance is achieved when $\approx 10\%$ frames from the dataset are used employing the proposed Temporal MT-CLAR base model and FSL-based approach.

Chapter 8

Generalisation

Contents

The influence of limited data and imbalanced distribution of samples per subject on subject-independent affect inference was demonstrated in Chapter 4. With limited data, it was observed that the features of valence and arousal learnt by the model are not generalisable across subjects. Further, to address the challenges of limited affect data, various MT-CLAR base models were proposed in Chapter 5, with an aim of achieving effective and general representations useful for downstream applications. These base models were employed to perform an FSL-based affect labelling task in Chapter 6, to avoid the copious task of manual annotation. Extensive experiments were performed on various affect datasets to assess the robustness of the proposed approaches, and the explicit results were presented, while the implicit results regarding the generalisability of the frameworks remain unexplored. In this chapter, our focus is on discussing generalisability, a pivotal aspect in affective computing. A comprehensive understanding of what generalisation means, and significance that it holds in affect inference is presented.

8.1 Introduction

The question of generalisability has been a historic concern in emotion research. A long-standing debate is regarding the variability of facial expressions from culture-to-culture. Studies have demonstrated that facial expressions are interpreted similarly across cultures, while the context and consequences of facial expressions might vary [Ekman 73]. These studies have the foundation for the development of standardised tools and methods for inferring emotions. From a psychological perspective, *generalisation* refers to the ability to understand affective states from one situation or context to another that share similarities [Ekman 92]. It involves the transfer of learned emotional responses from specific instances to new stimuli.

From an affective computing perspective, *generalisation* refers to the ability of models to interpret and predict emotions effectively not only from the data they were trained on, but also from new, unseen data or different contexts [Hastie 09]. For instance, while inferring emotions from facial expressions, a well-generalised model can accurately identify emotions from different people, varying facial expressions, lighting conditions, or camera angles. It ensures that the system's performance is robust and reliable when applied to diverse, real-world situations. In affect inference from speech, models that generalise effectively can accurately infer emotions in noisy environments, unseen speakers, spontaneous conversations, and considering various voice characteristics.

One of the factors that contribute to the generalisation ability of affect inference models is training data. For facial affect inference, the dataset should contain wide range of facial expressions from subjects, facial images from individuals with diverse demographics, such as different age groups, genders, ethnicities, and cultural backgrounds, data collected from different contexts, environments, and conditions where facial expressions might vary, including variations in lighting, poses, angles, occlusions, and facial accessories [Picard 00]. A balanced representation of different emotional states prevents biases and ensures that the model learns all expressions equally. Imbalanced datasets might skew the model's learning towards over-represented *classes*, leading to poor generalisation. Generalisable models perform consistently and reliably in different situations, making them more dependable for applications in HCI, mental health

monitoring, or personalised experiences. They help mitigate biases and ensure fairness across diverse populations. By inferring emotions accurately regardless of individual characteristics, they contribute to more ethical and inclusive applications.

Besides sufficient data, splitting the data appropriately, applying regularisation methods, expanding the training dataset by applying augmentation techniques like rotation, translation, etc., play a crucial role in enhancing model generalisability. *Transfer learning* is a technique where knowledge gained from solving one problem is applied to a different but related problem. In this approach, a model trained on a large dataset or a specific task (source domain) is adapted for a different but related dataset (target domain). Leveraging knowledge obtained from pre-trained models, transfer learning mitigates the need for large amounts of labeled data in the target domain. This is particularly beneficial when the target dataset is limited, expensive, or challenging to acquire. Pre-trained models also yield rich, generalised representations which can capture essential patterns.

In the context of this thesis, generalisability refers to (a) *cross-dataset generalisation*, which refers to efficient affect inference from FSL frameworks, where the MT-CLAR base model is trained on a dataset, but tested on a different dataset, (b) *subject-independent generalisation*, which refers to the model's ability to accurately estimate valence and arousal when the subjects in the frames of anchor set and query set are different (see Chapter 6).

8.2 Cross-dataset Generalisation

Cross-dataset generalisation refers to the model's ability to extend its learning from one dataset to perform effectively on a different, unseen dataset [Pan 10]. A common challenge in affective computing arises from variations in data distribution, collection setups, and demographic factors of subjects across different datasets [Hastie 09]. Addressing this challenge involves developing models that can generalise well beyond the training data's specific characteristics.

Achieving cross-dataset generalisation is crucial for enhancing the practical utility and reliability of affect estimation models [Baltrušaitis 15]. The significance lies in the capacity to develop robust models that are not confined to specific datasets, thereby ensuring the broader

applicability of affective computing technologies. The ability to generalise across datasets is particularly vital for real-world deployment, where diverse data sources are encountered. Furthermore, cross-dataset generalisation contributes to the advancement of the field by addressing the challenge of domain shift, enabling models to capture the underlying patterns of human emotion expressions that transcend dataset-specific characteristics [Baltrušaitis 15]. This section presents cross-dataset evaluation results of the proposed FSL-based method for valence and arousal estimation, where the base model is trained on a specific dataset (for example, *in-the-wild* image dataset, such as AffectNet), and tested on a different dataset with diverse characteristics (for example, long-range videos, such as Aff-Wild2).

The aim of this section is to present cross-dataset generalisability of the proposed FSL-based labelling of video frames described in Chapter 6. For all the results presented henceforth, an identical experimental setup mentioned in Section 6.5 is followed, except for the dataset used for training and evaluating models. First, an MT-CLAR base model is trained on a dataset X (as described in Chapter 5). Next, the trained base model is utilised to estimate per-frame valence and arousal values of a video from a dataset Y, employing an FSL-based approach (described in Chapter 6) using various support set configurations. It is emphasised that there is **no fine-tuning or re-training of our base model** on Y; the reported results are solely a consequence of employing transfer learning. In this setup, $X \in \{\text{AffectNet, Aff-Wild2}\}$ and $Y \in \{\text{AFEW-VA, Aff-Wild2}\}$.

8.2.1 Train on AffectNet, Test on AFEW-VA

Table 8.1 presents results of valence and arousal estimation using the proposed FSL-based method for AFEW-VA dataset, where the MT-CLAR base model is trained on AffectNet. For detailed description on the implementation details, refer Section 6.5 and Table 6.3. When only the first video frame is employed as anchor, measures very comparable to SOTA are obtained including a significantly better RMSE for valence estimation. It is crucial to note that the model in SOTA is trained directly on AFEW-VA, while our method employs transfer learning from AffectNet dataset to AFEW-VA dataset; from multi-task contrastive learning-based method to FSL-based method. Moreover, utilising any arbitrary video frame as an anchor enhances per-

Table 8.1: Few-shot affect inference on AFEW-VA (using 5FCV) with varying S configurations with *AU-guided MT-CLAR* as the base model, trained on AffectNet. Best results in bold, while '*' denotes results significantly better than SOTA ($p < 0.05$) as per Kolmogorov-Smirnov test. ‡ While the model in SOTA is trained directly on AFEW-VA, our method employs transfer learning, with results derived from adapting to AFEW-VA after pre-training the base model on AffectNet. † indicate % of total frames.

Row	A_S configuration	$\|S\|$ (†)	Valence				Arousal			
			RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Toisoul 21]†	-	0.23	0.70	0.69	0.65	0.22	0.67	0.66	0.81
1	First frame of corresponding video	91 (2.02%)	0.15 *	0.76 *	0.75 *	0.62	0.16 *	0.77 *	0.76 *	0.82
2	Random frame of corresponding video	91 (2.02%)	0.13 *	0.85 *	0.84 *	0.67	0.14 *	0.85 *	0.84 *	0.88 *
3	Random frame from a subject-specific video	34 (0.76%)	0.28	0.41	0.40	0.53	0.27	0.46	0.45	0.79
4	Random frame from a video of different subject	34 (0.76%)	0.37	0.04	0.04	0.38	0.39	0.01	0.01	0.71
5	Recurring 100^{th} frame of corresponding video	96 (2.13%)	0.15 *	0.75 *	0.74 *	0.62	0.16 *	0.76 *	0.75 *	0.81
6	Recurring 50^{th} frame of corresponding video	132 (2.93%)	0.14 *	0.79 *	0.79 *	0.62	0.15 *	0.79 *	0.78 *	0.83
7	Recurring 20^{th} frame of corresponding video	268 (5.96%)	0.10 *	0.90*	0.90 *	0.66	0.12 *	0.87 *	0.87 *	0.87 *
8	Recurring 10^{th} frame of corresponding video	494 (10.98%)	**0.07** *	**0.95** *	**0.95** *	0.66	**0.09** *	**0.91** *	**0.91** *	**0.89** *

formance metrics beyond SOTA [Toisoul 21], resulting in significantly improved RMSE, PCC, and CCC metrics. It is noteworthy that only 2% of AFEW-VA is labeled and used as support set. However, inferior metrics are observed in the constrained scenario where an anchor corresponding to the same subject ID is employed, with less than 1% of labeled AFEW-VA frames being utilised. As discussed in Section 6.6.2, the configuration of an anchor frame corresponding to a different subject ID of a query video is an extremely challenging case. As expected, lowest measures are observed in this case, with arousal-related metrics faring better than valence metrics. Surpassing SOTA, superior metrics are achieved by utilising multiple video frames as anchors, outperforming SOTA in all-but-one measure, even with as few as 5.96% labeled anchor frames.

8.2.2 Train on AffectNet, Test on Aff-Wild2

Extending the application of transfer learning, Table 8.2 presents the results of valence and arousal estimation for Aff-Wild2, an in-the-wild long-range video dataset, with the base model trained on AffectNet dataset. As the duration of the videos in Aff-Wild2 is long (up to 26.22

Table 8.2: Few-shot affect inference on the validation set of Aff-Wild2 with varying S configurations with base model trained on AffectNet. † indicate % of total frames. * denote values higher than SOTA on the validation set of Aff-Wild2.

Row	A_S configuration	$\|S\|$ (†)	Valence				Arousal			
			RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Tellamekala 22]	-	-	-	0.56	-	-	-	0.65	-
1	First frame of corresponding video	71 (0.02%)	0.43	0.22	0.18	0.52	0.39	0.21	0.15	0.64
2	Random frame of corresponding video	71 (0.02%)	0.38	0.57	0.56	0.79	0.31	0.40	0.40	0.84
3	Random frame from a video of different subject	71 (0.02%)	0.55	0.30	0.17	0.32	0.29	0.24	0.19	0.97
4	Recurring 1000^{th} frame of corresponding video	375 (0.11%)	0.35	0.60	0.60 *	0.78	0.31	0.41	0.40	0.83
5	Recurring 500^{th} frame of corresponding video	713 (0.21%)	0.32	0.66	0.66 *	0.81	0.29	0.46	0.45	0.86
6	Recurring 250^{th} frame of corresponding video	1390 (0.41%)	0.30	0.71	0.71 *	0.84	0.26	0.55	0.55	0.88
7	Recurring 100^{th} frame of corresponding video	3424 (1.01%)	0.25	0.80	0.80 *	0.88	0.22	0.68	0.67*	0.93
8	Recurring 10^{th} frame of corresponding video	33908 (10.01%)	**0.12**	**0.95**	**0.95** *	**0.93**	**0.11**	**0.92**	**0.92** *	**0.94**

minutes), there is a subsequent increase in the temporal distance between the first frame and the query frame. This leads to a reduced measure when compared to 'random frame' configuration. Amongst all the configurations, as expected, worst measures are observed when subjects in the anchor frames are different than query frame. However, measures better than SOTA [Tellamekala 22] are obtained for valence estimation on the validation set of Aff-Wild2 when recurring 1000^{th} frame is employed as an anchor frame, with ground-truth values available for just 0.11% (375 frames) of the validation dataset. Consistent with other FSL-based results (for example, see Table 8.1), the trend follows where measures are better when more anchor frames are employed for a video. For all the metrics, best values are obtained when recurring 10^{th} frame is considered as anchor frames.

8.2.3 Train on Aff-Wild2, Test on AFEW-VA

Different to the previous settings, where MT-CLAR base models were trained on image dataset, and the FSL-based approach was performed on a video dataset, in this setting, a temporal MT-CLAR is trained using Aff-Wild2, a video dataset, and the model is employed for estimating valence and arousal using the FSL-based approach on AFEW-VA, another video dataset. In

Table 8.3: Few-shot affect inference on AFEW-VA (using 5FCV) with varying S configurations with *Temporal MT-CLAR* as the base model, trained on Aff-Wild2. Best results in bold, while '*' denotes results significantly better than SOTA ($p < 0.05$) as per Kolmogorov-Smirnov test. † indicate % of total frames.

Row	A_S configuration	$\|S\|$ (†)	Valence				Arousal			
			RMSE ↓	PCC ↑	CCC ↑	SAGR ↑	RMSE ↓	PCC ↑	CCC ↑	SAGR ↑
0	SOTA [Toisoul 21]	-	0.23	0.70	0.69	0.65	0.22	0.67	0.66	0.81
1	First frame of corresponding video	91 (2.02%)	0.13 *	0.83 *	0.80 *	0.62	0.17 *	0.77 *	0.74 *	0.78
2	Random frame of corresponding video	91 (2.02%)	0.12 *	0.88 *	0.87 *	0.66	0.14 *	0.84 *	0.84 *	0.87 *
3	Random frame from a subject-specific video	34 (0.76%)	0.27	0.44	0.43	0.53	0.25	0.49	0.49	0.77
4	Random frame from a video of different subject	34 (0.76%)	0.37	-0.03	-0.03	0.38	0.34	0.09	0.09	0.68
5	Recurring 100^{th} frame of corresponding video	96 (2.13%)	0.14 *	0.82 *	0.78 *	0.62	0.17 *	0.76 *	0.73 *	0.78
6	Recurring 50^{th} frame of corresponding video	132 (2.93%)	0.14 *	0.83 *	0.80 *	0.62	0.16 *	0.79 *	0.77 *	0.80
7	Recurring 20^{th} frame of corresponding video	268 (5.96%)	0.10 *	0.92 *	0.90 *	0.65	0.12 *	0.88 *	0.87 *	0.87 *
8	Recurring 10^{th} frame of corresponding video	494 (10.98%)	**0.07** *	**0.96** *	**0.95** *	0.66	**0.09** *	**0.93** *	**0.92** *	**0.89** *

this setting, transfer learning is performed from a video dataset to another video dataset, and the results are presented in Table 8.3. As discussed in Chapter 7, the efficacy of non-local neural networks is evident when first frame of a video is considered in the anchor set configuration. With $\approx 2.02\%$ frames, maximum CCC is achieved in this setting, as compared to the counterparts where knowledge is transferred from non-temporal MT-CLAR base models (see Table 8.1). Further, superior performance is observed when a random frame of a video is considered as the anchor set, as compared to using first frame. Identical trends are observed when recurring n^{th} frame is considered, with recurring 10^{th} frame yielding a near-ceiling performance. However, even with temporal MT-CLAR, subject independent frame as anchor set remains a challenge, and worst performance is observed in this configuration.

8.3 Subject-independent Generalisation

In Chapter 6, an FSL-based method is defined to label valence and arousal values for frames in videos. This is achieved by employing the valence and arousal differentials learnt from MT-CLAR base model and through the support of anchor frames. Specifically, given a video

Table 8.4: CCC (↑) values of valence and arousal estimation from FSL-based method (described in Chapter 6) for anchor frames from '*different subjects* of a corresponding video'. The results are on AFEW-VA dataset using various MT-CLAR base models. The base models trained on AffectNet and fine-tuned on AFEW-VA dataset. BG (background) masked and AU-guided MT-CLARs are as defined in Chapter 5.

Base model	Valence	Arousal
Raw MT-CLAR	-0.03	-0.01
BG masked MT-CLAR	0.15	0.00
AU-guided MT-CLAR	**0.19**	**0.21**

and ground-truth valence (arousal) value for one of the frames of the video (anchor frame), subsequent frames are labelled by passing a pair of images (an anchor frame and a query frame) to the MT-CLAR base model, which is trained to output the difference between the valence (arousal) values of the input pair. Since true valence (arousal) value is known for one of the images in the pair, the value for the other image (query) is estimated. In order to examine the effect of anchor frames on valence (arousal) estimation, five configurations (see Figure 6.3) are selected. Out of the configurations proposed, "random frame from a video of different subject"–corresponding to anchor-query image pair with different subject IDs–is of particular interest in this section. The reason for focusing on subject-independent anchor frames is to avoid dependence on the data specific to a subject, thereby creating a generalised configuration to estimate affect without relying on subject information in the input image pair.

Table 8.4 presents valence and arousal estimation results for the configuration in focus, using various MT-CLAR base models. The experimental setups are as described in Section 6.5. Best results are obtained for both valence and arousal estimation using AU-guided MT-CLAR as the base model. This is attributed to the ability of AU-guided MT-CLAR to focus on specific facial regions associated with emotional expressions. AUs represent distinct facial muscle actions [Friesen 78], and by guiding attention based on these units, the model can selectively attend to crucial regions of the face that contribute most to emotion representation [Lucey 10]. This approach allows for a more fine-grained analysis of facial features, capturing subtle nuances and individual differences in emotional expressions. This attention mechanism enhances the model's sensitivity to relevant facial dynamics while mitigating the impact of subject-specific information. The tailored focus on specific AUs enables the model to extract discrimi-

native features for affect estimation, contributing to its superior performance.

The sub-optimal performance of the subject-independent configuration, as compared to subject-dependent setting (for example, see row of Table 6.3) is acknowledged. This reflects the inherent difficulty of considering a configuration where affect is estimated based on a different subject's affect. This issue has been a focal point in the field, where the inherent variations in facial expressions, demographics, and contextual factors pose significant hurdles [Mollahosseini 19]. However, solving this challenge could open up new directions for robust models, indicating a promising avenue for future research. Overcoming subject-independent limitations has the potential to advance the capabilities of affective computing models, enhancing their adaptability and generalisation across a wide range of real-world scenarios, contributing to the development of more resilient and accurate affect estimation systems.

8.4 Discussion and Conclusion

This chapter aims to provide a broad view of generalisation with respect to (a) datasets used for affect estimation, and (b) subject dependency in anchor set configurations. It is observed that the proposed MT-CLAR base model can be used for efficient affect labelling in videos using the FSL-based approach. Two comparisons are presented, (a) training non-temporal MT-CLAR base models using AffectNet, and using it for affect labelling in two datasets namely AFEW-VA, and Aff-Wild2, (b) training temporal MT-CLAR base model using Aff-Wild2 base model, and using it for labelling AFEW-VA. This setup of heterogeneous transfer learning yields measures comparable to SOTA with certain configurations, and near-ceiling performance with recurring 10^{th} frame as anchor set. A cross-dataset generalisation is important in the context of affect estimation, as it shows the robustness of the proposed methods especially in affect estimation tasks, where there is a significant possibility of datasets having varied distributions and diverse characteristics. Further, an excellent CCC on a new dataset with mere $\approx 2.02\%$ frames indicates that employing the proposed FSL-based approach, efficient affect annotation can be performed with minimal human contribution, yet achieving a quality equivalent to human standards.

With respect to subject-independent generalisation, a comparison of the performance when

anchor set configurations have random frames from a subject-specific video vs subject-independent video are presented. The inferior performance achieved with the subject-independent setting can be attributed to the additional challenge of discerning the emotional cues from identity-specific information. Understanding the distinctive features unique to each identity adds to the prevailing complexity of identifying the emotional differences. Therefore, the proposed AU-guided MT-CLAR alleviates the need to explicitly learn identity information, as it highlights the activated AUs. This strategic emphasis on AUs, allows the model to focus on the facial muscle movements expressed through AUs, thereby facilitating the model to capture the subtleties of emotional differences without explicitly delving into the intricacies of individual identities.

Overall, the proposed methods can be adapted to diverse datasets from real-world scenarios where conditions may vary. Affect estimation in a subject-independent setting is a challenging task. Nevertheless, time-continuous affect estimation in videos using limited labelled samples employing the proposed MT-CLAR and FSL-based framework is a promising direction. FSL, with its ability to generalise and learn from limited-labeled samples is a valuable approach to enhance the efficiency and accuracy of the affect labelling process.

Chapter 9

Conclusion

Contents

9.1 Summary of Thesis

The aim of this thesis is to introduce methodologies centered around the temporal modelling of affect, specifically addressing the challenge of limited labelled data. In Chapter 4, the research delves into subject-specific idiosyncrasies by employing the AFEW-VA dataset in subject-independent and subject-dependent settings. The proposed dynamically-weighted loss function optimised in the CNN-LSTM architecture reveals that valence and arousal features are not universally generalisable across subjects when dealing with limited information. Chapter 5 presents Multi-Task Contrastive Learning for Affect Representation, a novel architecture for estimating affect differences between a pair of images. MT-CLAR's effectiveness is demonstrated through extensive experiments on architecture design, and the additional integration

of a landmark-driven Action Units attention module, which enhances subject-independent affect estimation. Chapter 6 introduces a novel few-shot learning framework that leverages MT-CLAR for affect estimation from sparsely labelled data, surpassing SOTA on AFEW-VA with a support-set size $< 6\%$.

Subsequently, the thesis progresses into temporal modelling with Chapter 7 introducing *Temporal MT-CLAR*. This novel time-continuous affect inference system builds upon MT-CLAR and incorporates a non-local neural network to encode temporal information, capturing long-range dependencies spatially and temporally. Chapter 8 scrutinises the generalisability of proposed methods across diverse datasets, demonstrating the reusability of MT-CLAR and its applicability in different affect-related tasks. Collectively, these contributions advance the field by analysing affect inference in subject-specific and subject-agnostic settings, developing robust models for limited labelled data scenarios, introducing innovative temporal modelling techniques, and validating the generalisability of proposed methodologies across varied datasets and conditions. Figure 9.1 gives an overview of individual chapters and their interconnectedness[1].

9.2 Answering Research Questions

Based on the proposed methodologies and discussions so far, the following research questions posed in Section 1.5 are answered.

- ***Do subject-specific idiosyncrasies play a role in time-continuous automatic affect estimation?***

  In the literature (Chapter 2), it is noted that studies have established the existence of *signature* facial expressions corresponding to the basic categorical emotions, individual differences in emoting facial expressions nevertheless exist; factoring out these idiosyncrasies is critical for effective emotion inference [Ekman 11, Barrett 19]. Chapter 4 explores continuous human affect recognition using AFEW-VA [Kossaifi 17], an 'in-the-wild' video

[1]The source codes supporting the findings of this thesis can be obtained at the following public profile page: https://github.com/ravikiranrao.

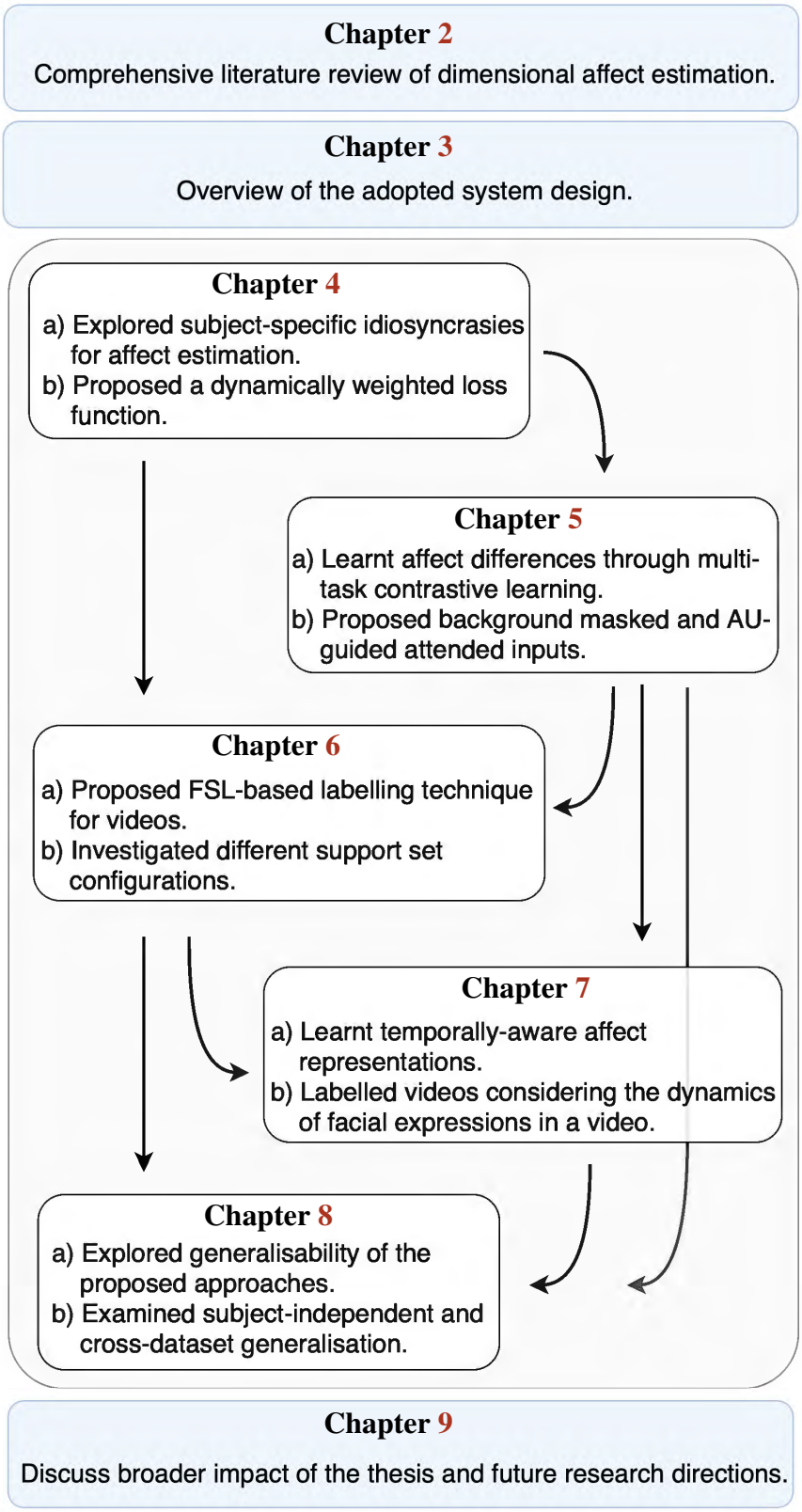


Figure 9.1: A summary of the interconnected themes across chapters of this thesis. Arrows signify how the content of one chapter relates to, influences, or builds upon the content of another chapter.

dataset with limited data, employing *subject-independent* and *subject-dependent* settings. The SI setting involves the use of training and test sets with mutually exclusive subjects, while training and test samples corresponding to the same subject can occur in the SD setting. A novel, dynamically-weighted loss function is employed with a CNN-LSTM architecture to optimise dynamic affect prediction. The results indicate that the features of valence and arousal learnt by the model are not generalisable across subjects. Visualisations convey that the features of the subject-independent framework are not as discriminative as the subject-dependent setting.

- ***Can learning affect differences be useful for learning affect representations?***

  To learn affect differences, Multi-Task Contrastive Learning for Affect Representation is proposed in Chapter 5. Learning affect differences holds significance due to two primary reasons: Firstly, it is posited that affect differences forms the foundational basis, focusing on the essential task of discerning similarities and dissimilarities or variations in affect, as opposed to specific patterns characteristic of individual emotions. Secondly, employing a paired learning approach facilitates the generation of additional data, a critical advantage particularly in scenarios characterised by limited available data. Additionally, resorting to (dis)similarity information reduces the information required, as compared to using the data with ground truth absolute emotion labels. Likewise, MT-CLAR combines multi-task learning with a Siamese network trained via contrastive learning to infer from a pair of expressive facial images (a) the (dis)similarity between the facial expressions, and (b) the difference in valence and arousal levels of the two faces. Results show that MT-CLAR base model can be fine-tuned and effectively used for other downstream affect related tasks. To improve subject-independent affect estimation, landmark-driven Action Units attention module is further introduced.

- ***How to estimate affect from limited labelled data?***

  Addressing the challenge of annotating dimensional affect, a Few-Shot Learning-based approach is proposed in Chapter 6. Traditionally, FSL algorithms learn from a few la-

belled samples and generalise to new tasks with limited or no additional data. The proposed FSL framework leverages affect differences learnt from MT-CLAR. Given one or a few labelled video frames (termed *support-set*), the framework labels the remainder of the video for valence and arousal. Experiments are performed on the AFEW-VA dataset with multiple support-set configurations; moreover, supervised learning on representations learnt via MT-CLAR are used for valence, arousal and categorical emotion prediction on the AffectNet [Mollahosseini 19] and AFEW-VA datasets. The results show that valence and arousal predictions via MT-CLAR are very comparable to the SOTA, and significantly outperform SOTA with a support-set ≈6% the size of the video dataset.

- ***How can the temporal context be used for better affect estimation?***

  The temporal dynamics of affect is well-established in the psychology literature. Most of the computational studies aiming at facial affect estimation, use static data for modelling affect. In Chapter 7, the base MT-CLAR model is extended by further incorporating a non-local neural network as the temporal encoder. *Temporal MT-CLAR* learns temporally-aware affect representations, and affect differences. Additionally, an FSL-based approach employing Temporal MT-CLAR is proposed. The results obtained reveal that temporal modelling enhances affect estimation, and with ≈ 1% of the total frames, measures higher than SOTA are achieved for both valence and arousal estimation on Aff-Wild2 dataset. Further, a near-ceiling performance is achieved when ≈ 10% frames from the dataset are used employing the proposed Temporal MT-CLAR base model and FSL-based approach.

- ***Are the proposed methodologies generalisable across various datasets and subjects?***

  In order to validate the reusability and generalisability of proposed methods in Chapter 4, Chapter 5, Chapter 6, and Chapter 7, investigation with respect to (a) datasets used for affect estimation, and (b) subject dependency in anchor set configurations are performed in Chapter 8. It is observed that the proposed MT-CLAR base model can be used for efficient affect labelling in videos using the FSL-based approach using *transfer learning*. With respect to subject-independent generalisation, a comparison of the per-

formance when anchor set configurations have random frames from a subject-specific video vs subject-independent video is presented. The proposed AU-guided MT-CLAR alleviates the need to explicitly learn identity information, as it highlights the activated AUs.

Overall, time-continuous dimensional affect estimation is a challenging task. Employing the proposed MT-CLAR and FSL-based framework is a promising direction for time-continuous affect estimation when the labelled data is limited. The proposed methods can be adapted to diverse datasets from real-world scenarios where conditions may vary.

9.3 Broader Impact

The broader implications are listed, but not limited to, the following.

Cost and time efficient annotation: The proposed methodologies significantly reduce cost and time associated with data annotation. By employing a few-shot learning framework–proposed in Chapter 6–and leveraging expert annotations on a sparse subset of the dataset, the resource-intensive task of annotating the entire dataset is alleviated. For instance, Chapter 8 demonstrates achieving annotation at par with humans on AFEW-VA with less than 6% of the data as a support set, and with approximately 1% on the Aff-Wild2 dataset, showcasing the remarkable efficiency and effectiveness of the proposed few-shot learning framework in real-world scenarios. Besides saving annotation costs, it accelerates the overall annotation process.

Enhancing data quality through expert annotation: The research promotes the move towards high-quality data by enabling the hiring of domain experts for annotation. Since only a small portion of the dataset requires expert annotation, the cost-effectiveness allows for the iterative improvement of data quality over time. This iterative process aligns with the ongoing advancements in understanding affective states, contributing to richer and more accurate datasets [Mollahosseini 19, Kollias 23].

Human-computer interaction: The research facilitates a true paradigm shift in human-computer interaction. By incorporating sparse annotations from human experts, the machine iteratively learns and refines its annotations. This collaborative approach not only streamlines the

annotation process but also creates a dynamic interaction between human expertise and machine learning. This concept aligns with research trends that emphasise the importance of collaborative human-machine interaction for improved affective computing outcomes. For example, the work of [Picard 01] discusses the significance of creating emotionally intelligent interfaces that can recognise and respond to users' affective states. The idea is further supported by studies on affect-aware HCI, emphasising the need for systems to adapt to users' emotional states for enhanced user experiences [Sekhavat 21].

Applicability to challenging Affective Computing problems: The proposed methodologies have broader applicability beyond the specific context of time-continuous affect estimation. They can be extended to other problems in affective computing where collecting and annotating data is difficult, for instance, in automatic depression detection, inferring micro-expressions, etc.

In the context of depression data, the proposed FSL-based framework's ability to learn from a minimal set of annotated samples allows for the creation of robust models even when dealing with sensitive and scarce datasets. This aligns with the challenges of collecting and annotating depression data [Ware 18], as the framework demonstrates effectiveness in scenarios where traditional approaches may be constrained by privacy concerns and ethical considerations. Moreover, in the field of micro-expression recognition, where subtle and transient facial expressions are crucial, the MT-CLAR-based models may excel in identifying subtle changes between a pair of images depicting micro expressions. The adaptability of the framework aligns with the challenges of diversity in micro-expressions [Xie 22], enabling accurate recognition even with limited labelled examples. This innovative approach contributes to the broader landscape of affective computing research, promoting advancements in understanding and addressing mental health issues.

Commercial products: Recently, research in Affective computing is not limited to developing computational models, but adopting a commercial perspective by further deploying the models for the benefit of end-users. For instance, advertisements that dynamically adapt based on users' emotional responses to gauge reactions and tailor content for a more engaging experience [Narayana 23a, Shukla 22]. The methodologies introduced in this thesis offer practical

solutions for enhancing the emotional intelligence of virtual assistants. In the consumer market, this translates to virtual assistants that can dynamically adapt their responses based on users' emotional states, creating a more empathetic and personalised interaction. This research's emphasis on subject-specific idiosyncrasies (Chapter 4) and temporal modelling (Chapter 7) directly contributes to the creation of mental health applications that can provide personalised support by understanding and responding to users' emotional well-being [Picard 00]. Furthermore, the robust models developed in this research find application in gaming interfaces, ensuring that games can respond to players' emotions, creating immersive and emotionally engaging experiences [Hemenover 18]. The commercial impact extends beyond specific sectors, influencing the design and functionality of a diverse range of consumer products from smart devices to educational applications.

9.4 Ethical Concerns

While some of the positive impacts of the research in the direction of building emotionally intelligent systems that enhance human-computer interaction and contribute to various applications are detailed, it is crucial to acknowledge and address ethical concerns associated with these advancements. The models developed in the thesis, designed to infer human emotions, could be exploited in ways that violate privacy, consent, or even lead to unintended social consequences. Misuse might manifest in scenarios where emotional data is collected without individuals' knowledge or used for manipulative purposes, such as targeted advertising or sentiment analysis without informed consent.

To foster transparency and reproducibility, publicly available datasets, such as AffectNet and Aff-Wild2, are utilised. Given the dynamic and complex nature of emotions, employing in-the-wild datasets is a conscious choice as a step towards avoiding biases related to specific races, religions, or other demographic factors (see Section 2.3). However, despite efforts to ensure a diverse representation, it is acknowledged that biases may exist in these datasets, and further investigation into potential biases is deemed necessary. Further, it is crucial to note that while efforts are made to mitigate biases, the models developed in this thesis may still be

intrinsically biased towards facial expressions of individuals from a particular region/culture. This thesis does not encompass an exhaustive exploration of bias mitigation strategies, as this topic extends beyond the defined scope. However, a comprehensive investigation into bias mitigation is recognised as a crucial area for future research.

Additionally, the environmental impact of model training is considered [Strubell 19], emphasising the conscientious use of resources. To minimize carbon footprint, small-sized models, particularly the MT-CLAR base models (approximately 25 million trainable parameters), are employed. The key advantage of the base model is its ability to be trained as a generic model once, and subsequently reused across various affect-related tasks. In this thesis, MT-CLAR base models are re-used for two kinds of tasks—FSL-based labelling (Chapter 6), and categorical and dimensional emotion inference (through MT-CLAR + SL as discussed in Section 6.4.2). Thus, the proposed methodologies is a step towards a more sustainable approach in model development.

Lastly, in terms of data protection, a responsible approach is taken by hosting and running models internally at the University of Canberra. This ensures that control over sensitive data is maintained, aligning with principles of privacy and data security. These ethical considerations collectively underscore the commitment to responsible and mindful practices in the development of affective computing models [Iren 23].

9.5 Limitations and Future Work

Affective computing systems operate in uncertain and dynamic environments. Modelling the variability in individual emotional expression is a challenge, as there can be individual differences in expressing the same emotion. Emotions are context sensitive, as the same emotional expression can have different meanings depending on the context. To reach one of the ultimate aims of affective computing systems, which is modelling and responding to emotions in real time, efficient algorithms and powerful processing capabilities are required. These systems must adapt continuously to changes in the user behaviour, requiring robust adaptive algorithms.

The exploration of methodologies addressing the challenge of limited labelled data for time-

continuous dimensional affect estimation has been addressed in this thesis. While the thesis detailed on the existing literature in the dimensional affect estimation, proposed MT-CLAR base models, and FSL-based methods for affect labelling using limited labelled data, there are avenues for further research. In this section, future research directions are proposed.

Model architecture: Transformer [Vaswani 17] models, known for their ability to capture long-range dependencies and contextual information, can be highly beneficial in modeling the temporal dynamics of affective states. Although Chapter 7 employs non-local neural network [Wang 18], a generic family of neural networks for capturing long-range dependencies for which *self-attention* module of the transformer model [Vaswani 17] is a special case [Wang 18], recent advancements in the variations of transformers models, such as Mamba [Gu 23], TimeSformer [Bertasius 21], Reformer [Kitaev 20], offer specific advantages that could enhance the performance of affect inference. Reformer, for instance, introduces locality-sensitive hashing to manage memory and speed constraints, making it suitable for real-time affect estimation.

Anchor frame selection: In Section 6.4.1, a finite set of anchor frames configurations are proposed. Consideration of finite anchor set configuration is viewed as a limitation when no such prior is available in the real-world settings. In future, a key focus is on automatically selecting anchor frames based on the input video. While the proposed method selects an anchor frame from a predefined support set (for example, first frame of videos, recurring 10^{th} frame of videos), dynamic adaptation of anchor frames may generate support set based on the characteristics of input videos. Integrating contextual information and implementing feedback mechanisms will contribute to more robust, scalable, and adaptable affect modelling systems.

Annotation tool: Based on the success of generalisability of the proposed FSL method (see Chapter 8), it is imperative to develop a tool for time-continuous affect annotation. The current time-continuous annotation tools [Cowie 00, Cowie 12] allow subjects to watch an audiovisual recording and move a cursor simultaneously to provide affect annotations. One of the limitations of this approach is the imposition of cognitive load in the subjects, which results in *time-delay* ground-truth values [Huang 15]. Instead, the proposed FSL-based annotation can interpolate based on the support set sampled from the ground-truth annotations at regular intervals, at reduced costs and time.

Multimodal MT-CLAR: Another intriguing area of research is an extension of MT-CLAR to process multimodal data. The integration of diverse modalities, such as incorporating facial expressions, speech, and physiological signals, can provide a more comprehensive understanding of affective states. Recently, numerous works details the significance of combining multiple modalities for improved affective and behavioural analysis [Huang 15, Ma 19, Chen 17, Malik 24]. By adapting MT-CLAR to a multimodal context, the model can learn intricate patterns and correlations across various data sources, leading to a more robust representation of affect.

Specifically, recent trends in large vision-language models (VLMs) have demonstrated significant potential in affective computing [Bustos 23]. Integrating VLMs into dimensional affect estimation frameworks can significantly improve the interpretability and precision of affective state analysis. Future research should focus on leveraging these models for real-time emotion inference and interactive applications, advancing affective computing systems' ability to understand and respond to human emotions in a multimodal context.

MT-CLAR Applications: Beyond the applications mentioned in the previous section, the proposed MT-CLAR methodology can be applied to domains other than affective computing, such as anomaly detection in videos and person re-identification. In video anomaly detection, MT-CLAR's representation learning, akin to the effectiveness shown in contrastive learning approaches [Lu 20], can be used in discerning subtle patterns indicative of abnormal behaviour. Similarly, in person re-identification task [Song 18], contrastive learning aids in learning discriminative features for accurate identification across varied scenes.

Unified model: The development of a unified model capable of jointly addressing valence, arousal, expression, mood, and other affective states is an exciting research direction. Integrating multiple dimensions of affect into a cohesive framework allows for a more comprehensive understanding of human emotions. While the current implementation of MT-CLAR integrates valence, arousal, and facial expressions in a weakly supervised manner, the incorporation of mood and other affective states remain unexplored. Along with the proposed MT-CLAR, works such as [Toisoul 21] emphasise the importance of unified models in capturing the intricacies of affective states. By leveraging such a unified approach, the model can inherently capture the interdependencies between different affective dimensions, contributing to a more accurate and

contextually rich representation of affect.

Defence against adversarial attacks: Affective computing systems are vulnerable to adversarial attacks, leading to failure in detecting emotions accurately. Small, imperceptible changes to input data, introducing malicious data into the training set, etc., can cause the system to predict incorrect emotions. Adversarial attacks can expose sensitive information about users' emotional states, leading to potential misuse of data. As part of the future work, adversarial examples will be included in the training set of MT-CLAR to make the model more robust to such attacks. This involves augmenting the training data with perturbed examples. Additionally, mechanisms to detect adversarial examples will be deployed into MT-CLAR, before they are processed by the model. This can involve integrating secondary models to recognise adversarial inputs.

Theoretical and explainable framework: While the proposed approaches demonstrated empirical effectiveness on various datasets, in future, a more robust theoretical and explainable framework has to be provided. Integrating theoretical constructs and explaining the decision-making processes within the model can enhance its interpretability. As a positive recent development in affective computing [Cortiñas-Lorenzo 23], there's an emphasis towards developing transparent and interpretable affect systems, as the systems are increasingly deployed in real-world contexts like education and healthcare sectors.

Bibliography

[Abadi 13] Mojtaba Khomami Abadi, Seyed Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. *User-centric Affective Video Tagging from MEG and Peripheral Physiological Responses*. In Affective Computing and Intelligent Interaction, pages 582–587, 2013. 72

[Abelson 62] Robert P Abelson and Vello Sermat. *Multidimensional scaling of facial expressions.* Journal of Experimental Psychology, 63(6):546, 1962. 5

[Ahn 21] Youngdo Ahn, Sung Joo Lee, and Jong Won Shin. *Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation*. IEEE Signal Processing Letters, 28:1190–1194, 2021. 106

[Akhtar 19] Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. *All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework*. IEEE Transactions on Affective Computing, 13(1):285–297, 2019. 53

[Ali 17] Abdallah El Ali, Torben Wallbaum, Merlin Wasmann, Wilko Heuten, and Susanne Boll. *Face2Emoji: Using Facial Emotional Expressions to Filter Emojis*. In Conference on Human Factors in Computing Systems, pages 1577–1584. ACM, 2017. 124

[Aspandi 21] Decky Aspandi, Federico Sukno, Björn W. Schuller, and Xavier Binefa. *An Enhanced Adversarial Network with Combined Latent Features for Spatio-temporal Fa-*

cial Affect Estimation in the Wild. In VISIGRAPP (4: VISAPP), pages 172–181. SCITEPRESS, 2021. 50, 75

[Baevski 20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, Volume 33, pages 12449–12460. Curran Associates, Inc., 2020. 44

[Baltrušaitis 13] Tadas Baltrušaitis, Ntombikayise Banda, and Peter Robinson. *Dimensional affect recognition using continuous conditional random fields*. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–8. IEEE, 2013. 49

[Baltrušaitis 15] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. *Cross-dataset learning and person-specific normalisation for automatic action unit detection*. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Volume 6, pages 1–6. IEEE, 2015. 97, 145, 146

[Baltrušaitis 18] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. *Openface 2.0: Facial behavior analysis toolkit*. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 59–66. IEEE, 2018. xxix, xxxiii, 93, 94, 97, 98

[Baltrušaitis 19] 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, 2019. 55

[Barrett 99] Lisa Feldman Barrett and James A. Russell. *The Structure of Current Affect: Controversies and Emerging Consensus*. Current Directions in Psychological Science, 8(1):10–14, 1999. 26

[Barrett 14] Lisa Feldman Barrett. *The conceptual act theory: A précis*. Emotion review, 6(4):292–297, 2014. 128

[Barrett 19] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. *Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements*. Psychological science in the public interest, 20(1):1–68, 2019. 15, 23, 72, 154

[Barros 18] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. *The OMG-Emotion Behavior Dataset*. In 2018 International Joint Conference on Neural Networks (IJCNN), pages 1–7, 2018. 37

[Bartko 66] John J Bartko. *The intraclass correlation coefficient as a measure of reliability*. Psychological Reports, 19(1):3–11, 1966. 31

[Bartlett 05] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. *Recognizing facial expression: machine learning and application to spontaneous behavior*. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Volume 2, pages 568–573. IEEE, 2005. 28

[Batliner 03] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. *How to find trouble in communication*. Speech Communication, 40(1-2):117–143, 2003. 28

[Bechara 00] Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. *Emotion, Decision Making and the Orbitofrontal Cortex*. Cerebral Cortex, 10(3):295–307, 03 2000. 2, 3

[Bengio 05] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. *The curse of dimensionality for local kernel machines*. Techn. Rep, 1258(12):1, 2005. 9

[Bengio 07] Yoshua Bengio and Yann LeCun. *Scaling learning algorithms towards AI*. Large-scale kernel machines, 34(5):1–41, 2007. 10

[Benitez-Quiroz 16] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martínez. *EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 5562–5570. IEEE Computer Society, 2016. 34

[Berridge 19] Kent C Berridge. *Affective valence in the brain: modules or modes?* Nature Reviews Neuroscience, 20(4):225–234, 2019. 8

[Bertasius 21] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. *Is Space-Time Attention All You Need for Video Understanding?* In Proceedings of the International Conference on Machine Learning (ICML), July 2021. 162

[Berthelot 19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. *MixMatch: A Holistic Approach to Semi-Supervised Learning*. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, Volume 32. Curran Associates, Inc., 2019. 47

[Bolger 03] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. *Diary methods: Capturing life as it is lived*. Annual review of psychology, 54(1):579–616, 2003. 25

[Borod 83] Joan C Borod, Elissa Koff, and Betsy White. *Facial asymmetry in posed and spontaneous expressions of emotion*. Brain and Cognition, 2(2):165–175, 1983. 27

[Bradley 94] 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. 30

[Brady 16] 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. 131

[Breuer 17] Ran Breuer and Ron Kimmel. *A deep learning perspective on the origin of facial expressions*. arXiv preprint arXiv:1705.01842, 2017. 41

[Bringmann 13] Laura F Bringmann, Nathalie Vissers, Marieke Wichers, Nicole Geschwind, Peter Kuppens, Frenk Peeters, Denny Borsboom, and Francis Tuerlinckx. *A network approach to psychopathology: new insights into clinical longitudinal data*. PloS one, 8(4):e60188, 2013. 25

[Brugman 04] Hennie Brugman, Albert Russel, and Xd Nijmegen. *Annotating Multi-media/Multi-modal Resources with ELAN*. In LREC, pages 2065–2068, 2004. 30

[Buades 05] Antoni Buades, Bartomeu Coll, and J-M Morel. *A non-local algorithm for image denoising*. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Volume 2, pages 60–65. IEEE, 2005. 133

[Bulat 17] Adrian Bulat and Georgios Tzimiropoulos. *How Far Are We From Solving the 2D & 3D Face Alignment Problem? (And a Dataset of 230,000 3D Facial Landmarks)*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1021–1030, Oct 2017. 89

[Bulat 22] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. *Pre-training Strategies and Datasets for Facial Representation Learning*. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, Computer Vision – ECCV 2022, pages 107–125, Cham, 2022. Springer Nature Switzerland. 97, 137

[Buller 94] David B Buller, Judee K Burgoon, Cindy H White, and Amy S Ebesu. *Interpersonal deception VII: Behavioral profiles of falsification, equivocation, and concealment*. Journal of Language and Social Psychology, 13(4):366–395, 1994. 28

[Buller 96] David B Buller and Judee K Burgoon. *Interpersonal deception theory*. Communication Theory, 6(3):203–242, 1996. 28

[Bustos 23] Cristina Bustos, Carles Civit, Brian Du, Albert Solé-Ribalta, and Agata Lapedriza. *On the use of Vision-Language models for Visual Sentiment Analysis: a study on CLIP*. In 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8. IEEE, 2023. 163

[Calvo 15] R.A. Calvo, S. D'Mello, J.M. Gratch, and A. Kappas. *The Oxford Handbook of Affective Computing*. Oxford Library of Psychology. Oxford University Press, 2015. 2

[Cannon 27] Walter B Cannon. *The James-Lange theory of emotions: A critical examination and an alternative theory*. The American Journal of Psychology, 39(1/4):106–124, 1927. 4

[Careaga 19] Chris Careaga, Brian Hutchinson, Nathan O. Hodas, and Lawrence Phillips. *Metric-Based Few-Shot Learning for Video Action Recognition*. ArXiv, abs/1909.09602, 2019. 106

[Caridakis 06] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. *Modeling Naturalistic Affective States via Facial and Vocal Expressions Recognition*. In Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI '06, page 146–154, New York, NY, USA, 2006. Association for Computing Machinery. 24

[Caruana 93] Rich Caruana. *Multitask Learning: A Knowledge-Based Source of Inductive Bias*. In International Conference on Machine Learning, 1993. 88

[Castellano 07] Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. *Recognising Human Emotions from Body Movement and Gesture Dynamics*. In Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg, 2007. 52

[Chao 14] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. *Multi-Scale Temporal Modeling for Dimensional Emotion Recognition in Video*. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, page 11–18, New York, NY, USA, 2014. Association for Computing Machinery. 49, 50, 131

[Chen 15] Shizhe Chen and Qin Jin. *Multi-Modal Dimensional Emotion Recognition Using Recurrent Neural Networks*. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC 2015, Brisbane, Australia, October 26, 2015, AVEC '15, page 49–56, New York, NY, USA, 2015. Association for Computing Machinery. 21, 49

[Chen 17] 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. 49, 88, 131, 163

[Chen 20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, Volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 89, 90, 122

[Chen 21] 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, 2021. 44

[Choi 18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 45

[Choi 20a] Dong Yoon Choi, Deok-Hwan Kim, and Byung Cheol Song. *Multimodal Attention Network for Continuous-Time Emotion Recognition Using Video and EEG Signals*. IEEE Access, 8:203814–203826, 2020. 54

[Choi 20b] Dong Yoon Choi and Byung Cheol Song. *Semi-Supervised Learning for Continuous Emotion Recognition Based on Metric Learning*. IEEE Access, 8:113443–113455, 2020. 21, 47

[Chopra 05] S. Chopra, R. Hadsell, and Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification*. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Volume 1, pages 539–546 vol. 1, 2005. 87

[Ciubotaru 19] Anca-Nicoleta Ciubotaru, Arnout Devos, Behzad Bozorgtabar, Jean-Philippe Thiran, and Maria Gabrani. *Revisiting Few-Shot Learning for Facial Expression Recognition*. arXiv preprint arXiv:1912.02751, 2019. 106

[Cohen 60] Jacob Cohen. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, 20(1):37–46, 1960. 31

[Cohen 03] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang. *Facial expression recognition from video sequences: temporal and static modeling*. Computer Vision and Image Understanding, 91(1):160–187, 2003. Special Issue on Face Recognition. 39

[Cohn 04] Jeffrey F. Cohn and Karen L. Schmidt. *The Timing of Facial Motion in Posed and Spontaneous Smiles*. Int. J. Wavelets Multiresolution Inf. Process., 2(2):121–132, 2004. 27, 28

[Cohn 10] Jeffrey F. Cohn. *Advances in Behavioral Science Using Automated Facial Image Analysis and Synthesis [Social Sciences]*. IEEE Signal Processing Magazine, 27(6):128–133, 2010. 13, 31

[Cootes 01] T.F. Cootes, G.J. Edwards, and C.J. Taylor. *Active appearance models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):681–685, 2001. 38, 92

[Cortiñas-Lorenzo 23] Karina Cortiñas-Lorenzo and Gerard Lacey. *Toward Explainable Affective Computing: A Review*. IEEE Transactions on Neural Networks and Learning Systems, pages 1–0, 2023. 164

[Cowie 00] Roddy Cowie, E. Douglas-Cowie, Suzie Savvidou, E. McMahon, M. Sawey, and M. Schröder. *'FEELTRACE': An instrument for recording perceived emotion in real time*. In ISCA tutorial and research workshop (ITRW) on speech and emotion, 01 2000. 30, 162

[Cowie 12] Roddy Cowie, Gary McKeown, and Ellen Douglas-Cowie. *Tracing emotion: an overview*. International Journal of Synthetic Emotions (IJSE), 3(1):1–17, 2012. 30, 162

[Damasio 98] Antonio R Damasio. *Emotion in the perspective of an integrated nervous system1*. Brain Research Reviews, 26(2-3):83–86, 1998. 3

[Darwin 72] 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, 1872. 4, 5, 6, 22

[de Gelder 15] B. de Gelder, A.W. de Borst, and R. Watson. *The perception of emotion in body expressions*. WIREs Cognitive Science, 6(2):149–158, 2015. 7

[DePaulo 03] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. *Cues to deception.* Psychological Bulletin, 129(1):74, 2003. 29

[Dhall 12] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. *Collecting Large, Richly Annotated Facial-Expression Databases from Movies*. IEEE MultiMedia, 19(03):34–41, jul 2012. xxvii, 21, 28, 29, 34, 35, 46, 63

[Dhall 15] Abhinav Dhall, O. V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. *Video and Image based Emotion Recognition Challenges in the Wild:EmotiW 2015*. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015, pages 423–426. ACM, 2015. 40, 46

[Dibeklioglu 12] Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers. *Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles*. In Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III, Volume 7574 of *Lecture Notes in Computer Science*, pages 525–538. Springer, 2012. xxvii, 8, 28

[Dibeklioğlu 15] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. *Recognition of Genuine Smiles*. IEEE Transactions on Multimedia, 17(3):279–294, 2015. xxvii, 8

[Ding 18] Hui Ding, Kumar Sricharan, and Rama Chellappa. *ExprGAN: Facial Expression Editing With Controllable Expression Intensity*. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 6781–6788. AAAI Press, 2018. 45

[D'Mello 12] Sidney K. D'Mello and Jacqueline M. Kory. *Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies*. In International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012, pages 31–38. ACM, 2012. 55

[D'mello 15] Sidney K. D'mello and Jacqueline Kory. *A Review and Meta-Analysis of Multimodal Affect Detection Systems*. ACM Comput. Surv., 47(3), feb 2015. 55

[Dosovitskiy 21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. *An Image is Worth 16x16*

Words: Transformers for Image Recognition at Scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. 43

[Douglas-Cowie 07] 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 Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2, pages 488–500. Springer, 2007. 21, 34

[Du 14] 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. xxvii, 4

[D'Mello 16] Sidney K. D'Mello. *On the Influence of an Iterative Affect Annotation Approach on Inter-Observer and Self-Observer Reliability*. IEEE Transactions on Affective Computing, 7(2):136–149, 2016. 29

[Eaton 02] Leslie Eaton and David Funder. *Emotional Experience in Daily Life: Valence, Variability, and Rate of Change*. Emotion (Washington, D.C.), 1:413–21, 01 2002. 128

[Ekman 69] Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. *Pan-Cultural Elements in Facial Displays of Emotion*. Science, 164(3875):86–88, 1969. 4, 5, 8, 21, 22, 72

[Ekman 71] Paul Ekman and Wallace V Friesen. *Constants across cultures in the face and emotion.* Journal of Personality and Social Psychology, 17(2):124, 1971. 6, 12, 128

[Ekman 73] Paul Ekman. *Cross-cultural studies of facial expression*. Darwin and Facial Expression: A Century of Research in Review, 169222(1), 1973. 144

[Ekman 78] Paul Ekman and Wallace V Friesen. *Facial action coding system*. Environmental Psychology & Nonverbal Behavior, 1978. 6, 8

[Ekman 82] Paul Ekman and Wallace V. Friesen. *Felt, false, and miserable smiles*. Journal of Nonverbal Behavior, 6:238–252, 1982. 27

[Ekman 92] Paul Ekman. *An argument for basic emotions*. Cognition and Emotion, 6(3-4):169–200, 1992. 5, 13, 144

[Ekman 94] Paul Ed Ekman and Richard J Davidson. *The nature of emotion: Fundamental questions*. Oxford University Press, 1994. 3

[Ekman 03] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, Volume 10. Ishk, 2003. 22

[Ekman 04] Paul Ekman. *Darwin, Deception, and Facial Expression*. Annals of the New York Academy of Sciences, 1000:205–21, 01 2004. 27

[Ekman 11] Paul Ekman and Daniel Cordaro. *What is Meant by Calling Emotions Basic*. Emotion Review, 3(4):364–370, 2011. 15, 23, 72, 154

[Eleftheriadis 17] Stefanos Eleftheriadis, Ognjen Rudovic, Marc Peter Deisenroth, and Maja Pantic. *Variational gaussian process auto-encoder for ordinal prediction of facial action units*. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, pages 154–170. Springer, 2017. 88

[Eyben 09] Florian Eyben, Martin Wöllmer, and Björn W. Schuller. *OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit*. In Affective Computing and Intelligent Interaction, Third International Conference and Workshops, ACII 2009, Amsterdam, The Netherlands, September 10-12, 2009, Proceedings, pages 1–6. IEEE Computer Society, 2009. 51

[Fasel 00] Beat Fasel and Juergen Luettin. *Recognition of Asymmetric Facial Action Unit Activities and Intensities*. In 15th International Conference on Pattern Recognition, ICPR'00, Barcelona, Spain, September 3-8, 2000, pages 5100–5103. IEEE Computer Society, 2000. 39, 46

[Feng 23] Kexin Feng and Theodora Chaspari. *Few-Shot Learning in Emotion Recognition of Spontaneous Speech Using a Siamese Neural Network With Adaptive Sample Pair Formation*. IEEE Transactions on Affective Computing, 14(2):1627–1633, 2023. 28, 106

[Fernandez Rojas 23] Raul Fernandez Rojas, Niraj Hirachan, Nicholas Brown, Gordon Waddington, Luke Murtagh, Ben Seymour, and Roland Goecke. *Multimodal physiological sensing for the assessment of acute pain*. Frontiers in Pain Research, 4, 2023. 31

[Fleiss 04] Joseph Fleiss, Bruce Levin, and Myunghee Paik. Statistical methods for rates and proportions, third edition, pages 50 – 63. John Wiley & Sons, 01 2004. 31

[Forgas 08] Joseph P. Forgas. *Affect and Cognition*. Perspectives on Psychological Science, 3(2):94–101, 2008. PMID: 26158876. 2

[Fragopanagos 05] N. Fragopanagos and J.G. Taylor. *Emotion recognition in human–computer interaction*. Neural Networks, 18(4):389–405, 2005. Emotion and Brain. 24

[Frantzidis 10] Christos A Frantzidis, Charalampos Bratsas, Manousos A Klados, Evdokimos Konstantinidis, Chrysa D Lithari, Ana B Vivas, Christos L Papadelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis D Bamidis. *On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications*. IEEE Transactions on Information Technology in Biomedicine, 14(2):309–318, 2010. 52

[Friesen 78] E Friesen and Paul Ekman. *Facial action coding system: a technique for the measurement of facial movement*. Palo Alto, 3(2):5, 1978. 31, 46, 93, 150

[Gastaldi 17] Xavier Gastaldi. *Shake-Shake regularization*. arXiv preprint arXiv:1705.07485, abs/1705.07485, 2017. 77, 79, 91

[Gendron 18] Maria Gendron, Carlos Crivelli, and Lisa Feldman Barrett. *Universality reconsidered: Diversity in making meaning of facial expressions*. Current Directions in Psychological Science, 27(4):211–219, 2018. 104

[Ghimire 13] Deepak Ghimire and Joonwhoan Lee. *Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines*. Sensors, 13(6):7714–7734, 2013. 38

[Goodfellow 13] 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 Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20, pages 117–124. Springer, 2013. 40

[Goodfellow 16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 10, 100

[Graver 02] Margaret R. Graver, editor. *Cicero on the Emotions: Tusculan Disputations 3 and 4*. University of Chicago Press, 2002. 4

[Grimm 07] Michael Grimm and Kristian Kroschel. *Emotion estimation in speech using a 3d emotion space concept*. Robust Speech Recognition and Understanding, pages 281–300, 2007. 51

[Grimm 08] Michael Grimm, Kristian Kroschel, and Shrikanth S. Narayanan. *The Vera am Mittag German audio-visual emotional speech database*. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, June 23-26 2008, Hannover, Germany, pages 865–868. IEEE Computer Society, 2008. 11

[Gu 08] Yuan Gu, Su-Lim Tan, Kai-Juan Wong, Moon-Ho Ringo Ho, and Li Qu. *Emotion-aware technologies for consumer electronics*. In 2008 IEEE International Symposium on Consumer Electronics, pages 1–4, 2008. 52

[Gu 12] Wenfei Gu, Cheng Xiang, Y. V. Venkatesh, Dong Huang, and Hai Lin. *Facial expression recognition using radial encoding of local Gabor features and classifier synthesis*. Pattern Recognition, 45(1):80–91, 2012. 39

[Gu 23] Albert Gu and Tri Dao. *Mamba: Linear-time sequence modeling with selective state spaces*. arXiv preprint arXiv:2312.00752, 2023. 162

[Gunes 10] Hatice Gunes and Maja Pantic. *Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners*. In Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10, pages 371–377. Springer, 2010. 54

[Gunes 13] Hatice Gunes and Björn Schuller. *Categorical and dimensional affect analysis in continuous input: Current trends and future directions*. Image and Vision Computing, 31(2):120–136, 2013. 23, 26, 31, 72

[Gupta 14] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth S. Narayanan. *Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions*. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, Orlando, Florida, USA, November 7, 2014, pages 33–40. ACM, 2014. 38

[Haag 04] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. *Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System*. In Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings, Volume 3068 of *Lecture Notes in Computer Science*, pages 36–48. Springer, 2004. 11

[Hajarolasvadi 21] Noushin Hajarolasvadi, Enver Bashirov, and Hasan Demirel. *Video-based person-dependent and person-independent facial emotion recognition*. Signal, Image and Video Processing, 15(5):1049–1056, 2021. 73

[Hamaker 15] Ellen L Hamaker, Eva Ceulemans, Raoul PPP Grasman, and Francis Tuerlinckx. *Modeling affect dynamics: State of the art and future challenges*. Emotion Review, 7(4):316–322, 2015. 25

[Han 17] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn W. Schuller. *Prediction-based learning for continuous emotion recognition in speech*. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017, pages 5005–5009. IEEE, 2017. 51

[Handrich 20] Sebastian Handrich, Laslo Dinges, Ayoub Al-Hamadi, Philipp Werner, and Zaher Al Aghbari. *Simultaneous prediction of valence/arousal and emotions on affectnet, aff-wild and afew-va*. Procedia Computer Science, 170:634–641, 2020. 41, 124, 132

[Happy 12] S L Happy, Anjith George, and Aurobinda Routray. *A real time facial expression classification system using Local Binary Patterns*. In 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pages 1–5, 2012. 39

[Hassani 17] Behzad Hassani and Mohammad H. Mahoor. *Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2278–2288. IEEE Computer Society, 2017. 132

[Hassani 22] Behzad Hassani, Pooran Singh Negi, and Mohammad H. Mahoor. *BReG-NeXt: Facial Affect Computing Using Adaptive Residual Networks With Bounded Gradient*. IEEE Transactions on Affective Computing, 13(2):1023–1036, 2022. 41

[Hastie 09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer, 2009. 144, 145

[Haxby 00] James V Haxby, Elizabeth A Hoffman, and M.Ida Gobbini. *The distributed human neural system for face perception*. Trends in Cognitive Sciences, 4(6):223–233, jun 2000. 6

[Hayale 19] Wassan Hayale, Pooran Negi, and Mohammad Mahoor. *Facial expression recognition using deep siamese neural networks with a supervised loss function*. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1–7. IEEE, 2019. 91

[Hayale 23] Wassan Hayale, Pooran Singh Negi, and Mohammad H. Mahoor. *Deep Siamese Neural Networks for Facial Expression Recognition in the Wild*. IEEE Transactions on Affective Computing, 14(2):1148–1158, 2023. 87

[He 16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 42, 76, 134

[Hemenover 03] Scott H Hemenover. *Individual differences in rate of affect change: studies in affective chronometry*. Journal of Personality and Social Psychology, 85(1):121, 2003. 128

[Hemenover 18] Scott H. Hemenover and Nicholas D. Bowman. *Video games, emotion, and emotion regulation: expanding the scope*. Annals of the International Communication Association, 42(2):125–143, 2018. 160

[Hilliard 18] Nathan Hilliard, L. Phillips, Scott Howland, Artëm Yankov, Court D. Corley, and Nathan Oken Hodas. *Few-Shot Learning with Metric-Agnostic Conditional Embeddings*. ArXiv, abs/1802.04376, 2018. 106

[Hinton 06] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. *A Fast Learning Algorithm for Deep Belief Nets*. Neural Computation, 18(7):1527–1554, jul 2006. 10

[Hochreiter 98] Sepp Hochreiter. *The vanishing gradient problem during learning recurrent neural nets and problem solutions*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02):107–116, 1998. 133

[Hochreiter 01] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. *Learning to Learn Using Gradient Descent*. In Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11, pages 87–94. Springer, Springer Berlin Heidelberg, 2001. 106

[Hoffmann 12] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C. Traue, and Henrik Kessler. *Mapping discrete emotions into the dimensional space: An empirical approach*. In 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 3316–3320, 2012. 24

[Hogg 07] Michael A. Hogg and Dominic Abrams. *Social cognition and attitudes*. In G. Neil Martin, Neil R. Carlson, and William Buskist, editors, Psychology. Third Edition, pages 684–721. Pearson Education Limited, 2007. 1

[Hu 21] Min Hu, Qian Chu, Xiaohua Wang, Lei He, and Fuji Ren. *A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video*. IEEE Signal Processing Letters, 28:698–702, 2021. 50, 76

[Huang 15] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. *An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction*. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15, page 41–48, New York, NY, USA, 2015. Association for Computing Machinery. 104, 162, 163

[Huang 16] Xiaohua Huang, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen, and Matti Pietikäinen. *Multi-modal emotion analysis from facial expressions and electroencephalogram*. Computer Vision and Image Understanding, 147:114–124, 2016. 54

[Hurri 02] Jarmo Hurri and Aapo Hyvärinen. *Temporal coherence, natural image sequences, and the visual cortex*. Advances in Neural Information Processing Systems, 15, 2002. 50, 131

[Iren 23] Deniz Iren, Ediz Yildirim, and Krist Shingjergji. *Ethical Risks, Concerns, and Practices of Affective Computing: A Thematic Analysis*. In 11th International Conference on Affective Computing and Intelligent Interaction, 2023. 12, 161

[Irons 95] David Irons. *Descartes and modern theories of emotion*. The Philosophical Review, 4(3):291–302, 1895. 4

[James 84] William James. *What is an Emotion?* Mind, 9(34):188–205, 1884. 4

[Jang 19] Youngkyoon Jang, Hatice Gunes, and Ioannis Patras. *Registration-free Face-SSD: Single shot analysis of smiles, facial attributes, and affect in the wild*. Computer Vision and Image Understanding, 182:17–29, 2019. 123

[Jeong 22] Euiseok Jeong, Geesung Oh, and Sejoon Lim. *Multi-task Learning for Human Affect Prediction with Auditory–Visual Synchronized Representation*. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2437–2444, 2022. 54, 88

[Jessen 11] S. Jessen and S.A. Kotz. *The temporal dynamics of processing emotions from vocal, facial, and bodily expressions*. NeuroImage, 58(2):665–674, 2011. 54

[Ji 13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. *3D Convolutional Neural Networks for Human Action Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):221–231, 2013. 132

[Ji 23] Yanli Ji, Yuhan Hu, Yang Yang, and Heng Tao Shen. *Region Attention Enhanced Unsupervised Cross-Domain Facial Emotion Recognition*. IEEE Transactions on Knowledge and Data Engineering, 35(4):4190–4201, 2023. 48

[Jia 21] Shan Jia, Shuo Wang, Chuanbo Hu, Paula J Webster, and Xin Li. *Detection of genuine and posed facial expressions of emotion: databases and methods*. Frontiers in Psychology, 11:580287, 2021. 28

[Jing 19] Longlong Jing and Yingli Tian. *Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43:4037–4058, 2019. 89, 122

[Juslin 05] Patrik N Juslin, Klaus R Scherer, J Harrigan, and R Rosenthal. *Vocal expression of affect*. The New Handbook of Methods in Nonverbal Behavior Research, pages 65–135, 2005. 7

[Juslin 18] Patrik N Juslin, Petri Laukka, and Tanja Bänziger. *The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion*. Journal of Nonverbal Behavior, 42:1–40, 2018. 28

[Kaltwang 16] 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, 2016. 69

[Kanade 00] T. Kanade, J.F. Cohn, and Yingli Tian. *Comprehensive database for facial expression analysis*. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pages 46–53, 2000. 31

[Kanwisher 97] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. *The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception*. Journal of Neuroscience, 17(11):4302–4311, 1997. 6

[Katsimerou 16] Christina Katsimerou, Joris Albeda, Alina Huldtgren, Ingrid Heynderickx, and Judith A Redi. *Crowdsourcing empathetic intelligence: The case of the annotation of EMMA database for emotion and mood recognition*. ACM Transactions on Intelligent Systems and Technology (TIST), 7(4):1–27, 2016. xxvii, 29, 35

[Keren 17] Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Björn Schuller. *End-to-end learning for dimensional emotion recognition from physiological signals*. In 2017 IEEE International Conference on Multimedia and Expo (ICME), pages 985–990. IEEE, 2017. 52

[Khademi 14] Mahmoud Khademi and Louis-Philippe Morency. *Relative facial action unit detection*. In IEEE Winter Conference on Applications of Computer Vision, pages 1090–1095. IEEE, 2014. 88

[Khan 20] Gulraiz Khan, Sahar Samyan, Muhammad Usman Ghani Khan, Muhammad Shahid, and Samyan Qayyum Wahla. *A survey on analysis of human faces and facial expressions datasets*. International Journal of Machine Learning and Cybernetics, 11:553–571, 2020. 33

[Khorrami 16] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S. Huang. *How deep neural networks can improve emotion recognition on video data*. In 2016 IEEE International Conference on Image Processing (ICIP), pages 619–623, 2016. 49, 131

[Khosrowabadi 10] Reza Khosrowabadi, Hiok Chai Quek, Abdul Wahab, and Kai Keng Ang. *EEG-based emotion recognition using self-organizing map for boundary detection*. In 2010 20th International Conference on Pattern Recognition, pages 4242–4245. IEEE, 2010. 52

[Kim 17] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. *Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild*. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 529–535, 2017. 47, 131

[Kim 21] Daeha Kim and Byung Cheol Song. *Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition*. Proceedings of the AAAI Conference on Artificial Intelligence, 35(7):5948–5956, May 2021. 41, 45

[Kim 22a] Daeha Kim and Byung Cheol Song. *Emotion-Aware Multi-View Contrastive Learning For Facial Emotion Recognition*. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII, page 178–195, Berlin, Heidelberg, 2022. Springer-Verlag. 41, 87

[Kim 22b] Daeha Kim and Byung Cheol Song. *Optimal Transport-based Identity Matching for Identity-invariant Facial Expression Recognition*. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022. 42

[Kingma 14] Diederik P Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014. 78, 97, 137

[Kipp 01] Michael Kipp. *ANVIL - a generic annotation tool for multimodal dialogue*. In Interspeech, 2001. 30

[Kitaev 20] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. *Reformer: The Efficient Transformer*. In International Conference on Learning Representations, 2020. 162

[Kollias 18a] Dimitrios Kollias and Stefanos Zafeiriou. *Aff-wild2: Extending the aff-wild database for affect recognition*. arXiv preprint arXiv:1811.07770, 2018. 46, 72

[Kollias 18b] Dimitrios D. Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. *Deep Neural Network Augmentation: Generating Faces for Affect Analysis*. International Journal of Computer Vision, 128:1455–1484, 2018. 46, 72, 80, 123, 124

[Kollias 19a] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. *Deep Affect Prediction In-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond*. Int. J. Comput. Vision, 127(6–7):907–929, jun 2019. 68, 76, 80, 104

[Kollias 19b] Dimitrios Kollias and Stefanos Zafeiriou. *Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace*. In BMVC, page 297. BMVA Press, 2019. xxvii, 21, 25, 36, 42, 66, 67

[Kollias 20] Dimitrios Kollias and Stefanos Zafeiriou. *VA-StarGAN: continuous affect generation*. In Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20, pages 227–238. Springer, 2020. 45, 46

[Kollias 21] Dimitrios Kollias and Stefanos Zafeiriou. *Exploiting Multi-CNN Features in CNN-RNN Based Dimensional Emotion Recognition on the OMG in-the-Wild Dataset*. IEEE Transactions on Affective Computing, 12(3):595–606, 2021. 49, 131

[Kollias 23] Dimitrios Kollias. *ABAW: Learning From Synthetic Data & Multi-Task Learning Challenges*. In Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, page 157–172, Berlin, Heidelberg, 2023. Springer-Verlag. 40, 158

[Kossaifi 17] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. *AFEW-VA Database for Valence and Arousal Estimation in-the-Wild*. Image and Vision Computing, 65(C):23–36, sep 2017. xxviii, 15, 28, 35, 41, 46, 63, 69, 72, 73, 76, 105, 106, 114, 124, 154

[Kossaifi 19] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. *SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43:1022–1040, 2019. xxvii, 29, 36, 46, 76

[Kossaifi 20a] Jean Kossaifi, Zachary C Lipton, Arinbjorn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. *Tensor regression networks*. Journal of Machine Learning Research, 21(123):1–21, 2020. 41

[Kossaifi 20b] 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 (CVPR), pages 6060–6069, June 2020. 10, 41, 50, 72, 76, 77, 80, 106, 122, 124, 131

[Larsen 87] Randy J. Larsen and Ed Diener. *Affect intensity as an individual difference characteristic: A review*. Journal of Research in Personality, 21(1):1–39, 1987. 21

[LeCun 15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. *Deep learning*. Nature, 521(7553):436–444, 2015. 10

[Lee 05] Chul Min Lee and S.S. Narayanan. *Toward detecting emotions in spoken dialogs*. IEEE Transactions on Speech and Audio Processing, 13(2):293–303, 2005. 28

[Lee 18] Jiyoung Lee, Sunok Kim, Seungryong Kiim, and Kwanghoon Sohn. *Spatiotemporal Attention Based Deep Neural Networks for Emotion Recognition*. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1513–1517, 2018. 50, 131

[Lerner 15] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. *Emotion and Decision Making*. Annual Review of Psychology, 66(1):799–823, 2015. PMID: 25251484. 3

[Lewis 10] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. *Handbook of Emotions*. Guilford Press, 2010. 22

[Li 19] Dahua Li, Zhe Wang, Chuhan Wang, Shuang Liu, Wenhao Chi, Enzeng Dong, Xiaolin Song, Qiang Gao, and Yu Song. *The Fusion of Electroencephalography and Facial Expression for Continuous Emotion Recognition*. IEEE Access, 7:155724–155736, 2019. 52, 54

[Li 21] Zhongjie Li, Gaoyan Zhang, Jianwu Dang, Longbiao Wang, and Jianguo Wei. *Multi-Modal Emotion Recognition Based On deep Learning Of EEG And Audio Signals*. In

2021 International Joint Conference on Neural Networks (IJCNN), pages 1–6, 2021. 54

[Li 22] Shan Li and Weihong Deng. *Deep Facial Expression Recognition: A Survey*. IEEE Transactions on Affective Computing, 13(3):1195–1215, 2022. 8, 24

[Li 23] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. *A Spontaneous Driver Emotion Facial Expression (DEFE) Dataset for Intelligent Vehicles: Emotions Triggered by Video-Audio Clips in Driving Scenarios*. IEEE Transactions on Affective Computing, 14(1):747–760, 2023. 28

[Lian 18] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. *Speech Emotion Recognition via Contrastive Loss under Siamese Networks*. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data, ASMMC-MMAC'18, page 21–26, New York, NY, USA, 2018. Association for Computing Machinery. 87

[Lin 89] Lawrence I-Kuei Lin. *A Concordance Correlation Coefficient to Evaluate Reproducibility*. Biometrics, 45(1):255–268, 1989. 79

[Liu 16] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. *Large-Margin Softmax Loss for Convolutional Neural Networks*. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, page 507–516. JMLR.org, 2016. 98

[Liu 19] Xiaofeng Liu, B.V.K. Vijaya Kumar, Ping Jia, and Jane You. *Hard negative generation for identity-disentangled facial expression recognition*. Pattern Recognition, 88:1–12, 2019. 42

[Liu 20a] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. *SAANet: Siamese action-units attention network for improving dynamic facial expression recognition*. Neurocomputing, 413:145–157, 2020. 87, 92

[Liu 20b] Yang Liu, Xingming Zhang, Yubei Lin, and Haoxiang Wang. *Facial Expression Recognition via Deep Action Units Graph Network Based on Psychological Mechanism*. IEEE Transactions on Cognitive and Developmental Systems, 12(2):311–322, 2020. 96

[Lu 20] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. *Few-Shot Scene-Adaptive Anomaly Detection*. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V, page 125–141, Berlin, Heidelberg, 2020. Springer-Verlag. 163

[Lucey 10] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pages 94–101, 2010. 33, 93, 120, 150

[Ma 19] Jia-Xin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. *Emotion Recognition using Multimodal Residual LSTM Network*. In Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, pages 176–183. ACM, 2019. 163

[Ma 22] Bowen Ma, Rudong An, Wei Zhang, Yu Ding, Zeng Zhao, Rongsheng Zhang, Tangjie Lv, Changjie Fan, and Zhipeng Hu. *Facial Action Unit Detection and Intensity Estimation from Self-supervised Representation*. arXiv preprint arXiv:2210.15878, 2022. 44

[Ma 23a] Bowen Ma, Wei Zhang, Feng Qiu, and Yu Ding. *A Unified Approach to Facial Affect Analysis: The MAE-Face Visual Representation*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5923–5932, 2023. 44, 46

[Ma 23b] Fuyan Ma, Bin Sun, and Shutao Li. *Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion*. IEEE Transactions on Affective Computing, 14(2):1236–1248, 2023. 44

[Makhmudkhujaev 19] Farkhod Makhmudkhujaev, Mohammad Abdullah-Al-Wadud, Md Tauhid Bin Iqbal, Byungyong Ryu, and Oksam Chae. *Facial expression recognition with local prominent directional pattern*. Signal Processing: Image Communication, 74:1–12, 2019. 39

[Malik 24] Harshit Malik, Hersh Dhillon, Ravikiran Parameshwara, Roland Goecke, and Ramanathan Subramanian. *Examining the Influence of Personality and Multimodal Behavior on Hireability Impressions*. In Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP '23, New York, NY, USA, 2024. Association for Computing Machinery. 163

[Mao 19] Qirong Mao, Qing Zhu, Qiyu Rao, Hongjie Jia, and Sidian Luo. *Learning Hierarchical Emotion Context for Continuous Dimensional Emotion Recognition From Video Sequences*. IEEE Access, 7:62894–62903, 2019. 49, 131

[Marrero-Fernández 19] Pedro D. Marrero-Fernández, Fidel A. Guerrero-Peña, Tsang Ing Ren, and Alexandre Cunha. *FERAtt: Facial Expression Recognition With Attention Net*. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, pages 837–846. Computer Vision Foundation / IEEE, 2019. 43

[Martínez 13] Héctor Perez Martínez, Yoshua Bengio, and Georgios N. Yannakakis. *Learning Deep Physiological Models of Affect*. IEEE Computational Intelligence Magazine, 8(2):20–33, 2013. 52

[Mathias 14] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. *Face Detection without Bells and Whistles*. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision - ECCV 2014 - 13th European Con-

ference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV, Volume 8692 of *Lecture Notes in Computer Science*, pages 720–735. Springer, 2014. 67

[Mavadati 13] Seyed Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. *DISFA: A Spontaneous Facial Action Intensity Database*. IEEE Transactions on Affective Computing, 4(2):151–160, 2013. xxvii, 27, 69, 92

[Mavadati 16] Seyed Mohammad Mavadati, Peyten Sanger, and Mohammad H. Mahoor. *Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016, pages 1452–1459. IEEE Computer Society, 2016. xxvii, 27, 29

[Mehrabian 96] Albert Mehrabian. *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament*. Current Psychology, 14(4):261–292, 1996. 22

[Mehrabian 17] Albert Mehrabian. *Communication without words*. In Communication Theory, pages 193–200. Routledge, 2017. 37

[Meng 13] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Al-Shuraifi, and Yunhong Wang. *Depression recognition based on dynamic facial and vocal expression features using partial least square regression*. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC@ACM Multimedia 2013, Barcelona, Spain, October 21, 2013, pages 21–30. ACM, 2013. 38

[Meng 16] Hongying Meng, Nadia Bianchi-Berthouze, Yangdong Deng, Jinkuang Cheng, and John Paul Cosmas. *Time-Delay Neural Network for Continuous Emotional Dimension Prediction From Facial Expression Sequences*. IEEE Transactions on Cybernetics, 46(4):916–929, 2016. 49

[Meng 17] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. *Identity-Aware Convolutional Neural Network for Facial Expression Recognition*. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 558–565, 2017. 42, 46

[Meng 22a] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Yuanyuan Deng, Ruichen Li, Yannan Wu, Jinming Zhao, Fengsheng Qiao, Qin Jin, and Chuanhe Liu. *Multi-modal Emotion Estimation for in-the-wild Videos*. ArXiv, abs/2203.13032, 2022. 44

[Meng 22b] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. *Valence and Arousal Estimation based on Multimodal Temporal-Aware Features for Videos in the Wild*. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022, pages 2344–2351. IEEE, 2022. 50, 54, 131

[Metallinou 11] Angeliki Metallinou, Athanassios Katsamanis, Yun Wang, and Shrikanth S. Narayanan. *Tracking changes in continuous emotion states using body language and prosodic cues*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, pages 2288–2291. IEEE, 2011. 53

[Metallinou 13] Angeliki Metallinou and Shrikanth Narayanan. *Annotation and processing of continuous emotional attributes: Challenges and opportunities*. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–8, 2013. 29, 72

[Mikels 05] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. *Emotional category data on images from the International Affective Picture System*. Behavior Research Methods, 37(4):626, 2005. xxix, 91, 92

[Miller 95] George A Miller. *WordNet: a lexical database for English*. Communications of the ACM, 38(11):39–41, 1995. 34

[Mitenkova 19] 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, 2019. 41, 80, 87, 114, 124

[Moberly 08] Nicholas J Moberly and Edward R Watkins. *Ruminative self-focus and negative affect: an experience sampling study.* Journal of abnormal psychology, 117(2):314, 2008. 25

[Mollahosseini 16] Ali Mollahosseini, Behzad Hassani, Michelle J. Salvador, Hojjat Abdollahi, David Chan, and Mohammad H. Mahoor. *Facial Expression Recognition from World Wild Web*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016, pages 1509–1516. IEEE Computer Society, 2016. 35

[Mollahosseini 19] 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, 2019. 16, 35, 64, 72, 84, 105, 106, 123, 151, 157, 158

[Moors 14] Agnes Moors. *Flavors of Appraisal Theories of Emotion*. Emotion Review, 6(4):303–307, 2014. 128

[Morris 15] R Morris, Daniel McDuff, and R Calvo. *Crowdsourcing Techniques for Affective Computing*. In The Oxford Handbook of Affective Computing, pages 384–394. Oxford Univ. Press Oxford, UK, 2015. 30

[Myers 76] Ronald E Myers. *Comparative Neurology of Vocalization and Speech: Proof of a Dichotomy*. Annals of the New York Academy of Sciences, 280(1):745–757, 1976. 27

[Mühlberger 10] Andreas Mühlberger, Matthias J. Wieser, Antje B.M. Gerdes, Monika C.M. Frey, Peter Weyers, and Paul Pauli. *Stop looking angry and smile, please: start and stop of the very same facial expression differentially activate threat- and reward-related brain networks*. Social Cognitive and Affective Neuroscience, 6(3):321–329, 05 2010. 26

[Nadaraya 64] Elizbar Nadaraya. *On Estimating Regression*. Theory of Probability and Its Applications, 9:141–142, 1964. 38

[Nakisa 20] Bahareh Nakisa, Mohammad Naim Rastgoo, Andry Rakotonirainy, Frederic Maire, and Vinod Chandran. *Automatic Emotion Recognition Using Temporal Multimodal Deep Learning*. IEEE Access, 8:225463–225474, 2020. 54

[Narayana 22] Soujanya Narayana, Ramanathan Subramanian, Ibrahim Radwan, and Roland Goecke. *To Improve Is to Change: Towards Improving Mood Prediction by Learning Changes in Emotion*. In Companion Publication of the 2022 International Conference on Multimodal Interaction, ICMI '22 Companion, page 36–41, New York, NY, USA, 2022. Association for Computing Machinery. 42

[Narayana 23a] Soujanya Narayana, Shweta Jain, Harish Katti, Roland Goecke, and Ramanathan Subramanian. *Affective computational advertising based on perceptual metrics*. In Affective Computing in Healthcare: Applications based on biosignals and artificial intelligence, pages 4–1. IOP Publishing Bristol, UK, 2023. 159

[Narayana 23b] Soujanya Narayana, Ibrahim Radwan, Ravikiran Parameshwara, Iman Abbasnejad, Akshay Asthana, Ramanathan Subramanian, and Roland Goecke. *A Weakly Supervised Approach to Emotion-change Prediction and Improved Mood Inference*. In 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8, 2023. 85

[Narayana 23c] Soujanya Narayana, Ramanathan Subramanian, Ibrahim Radwan, and Roland Goecke. *Focus on Change: Mood Prediction by Learning Emotion Changes via Spatio-Temporal Attention*. arXiv preprint arXiv:2303.06632, 2023. 88

[Nicolaou 10] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. *Automatic Segmentation of Spontaneous Data using Dimensional Labels from Multiple Coders*. In Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pages 43–48. German Research Center for AI (DFKI), 2010. 54

[Nicolaou 11] 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. 21, 38

[Nicolaou 12] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. *Output-associative RVM regression for dimensional and continuous emotion prediction*. Image and Vision Computing, 30(3):186–196, 2012. 39, 49

[Nicolle 12] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. *Robust continuous prediction of human emotions using multiscale dynamic cues*. In Proceedings of the 14th ACM international conference on Multimodal interaction, pages 501–508, 2012. 38

[Ortony 90] Andrew Ortony and Terence J Turner. *What's basic about basic emotions?* Psychological review, 97(3):315, 1990. 23

[Pan 10] Sinno Jialin Pan and Qiang Yang. *A Survey on Transfer Learning*. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010. 145

[Pan 23] Jiahui Pan, Weijie Fang, Zhihang Zhang, Bingzhi Chen, Zheng Zhang, and Shuihua Wang. *Multimodal emotion recognition based on facial expressions, speech, and EEG*. IEEE Open Journal of Engineering in Medicine and Biology, 2023. 54

[Panksepp 92] Jaak Panksepp. *A critical role for" affective neuroscience" in resolving what is basic about basic emotions.* Psychological Review, 99(3):554, 1992. 23

[Panksepp 04] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004. 3

[Panksepp 11] Jaak Panksepp. *Cross-Species Affective Neuroscience Decoding of the Primal Affective Experiences of Humans and Related Animals*. PLOS ONE, 6(9):1–15, 09 2011. 2, 8

[Pantic 00] Maja Pantic and Leon J. M. Rothkrantz. *Automatic analysis of facial expressions: The state of the art*. IEEE Transactions on pattern analysis and machine intelligence, 22(12):1424–1445, 2000. 13

[Pantic 05] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. *Web-based database for facial expression analysis*. In 2005 IEEE International Conference on Multimedia and Expo, 2005. 33, 72

[Parameshwara 22] Ravikiran Parameshwara, Soujanya Narayana, Murugappan Murugappan, Ramanathan Subramanian, Ibrahim Radwan, and Roland Goecke. *Automated Parkinson's Disease Detection and Affective Analysis from Emotional EEG Signals*. arXiv preprint arXiv:2202.12936, 2022. 26

[Parameshwara 23a] Ravikiran Parameshwara, Ibrahim Radwan, Akshay Asthana, Iman Abbasnejad, Ramanathan Subramanian, and Roland Goecke. *Efficient Labelling of Affective Video Datasets via Few-Shot & Multi-Task Contrastive Learning*. In Proceedings of the 31st ACM International Conference on Multimedia, MM '23, page 6161–6170, New York, NY, USA, 2023. Association for Computing Machinery. 84, 104

[Parameshwara 23b] Ravikiran Parameshwara, Ibrahim Radwan, Ramanathan Subramanian, and Roland Goecke. *Examining Subject-Dependent and Subject-Independent Human Affect Inference from Limited Video Data*. In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pages 1–6. IEEE, 2023. 71

[Parkhi 15] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. *Deep face recognition*. In BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015. 42

[Parthasarathy 17] Srinivas Parthasarathy and Carlos Busso. *Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning.* In Interspeech, Volume 2017, pages 1103–1107, 2017. 51

[Paszke 19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, Volume 32. Curran Associates, Inc., 2019. 78, 97, 112, 137

[Pei 19] Ercheng Pei, Dongmei Jiang, Mitchel Alioscha-Perez, and Hichem Sahli. *Continuous affect recognition with weakly supervised learning*. Multimedia Tools and Applications, 78:19387–19412, 2019. 47, 85

[Pfister 11] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. *Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework*. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pages 868–875, 2011. 28

[Pfister 14] Tomas Pfister, James Charles, and Andrew Zisserman. *Domain-Adaptive Discriminative One-Shot Learning of Gestures*. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 814–829. Springer International Publishing, 2014. 106

[Piana 16] Stefano Piana, Alessandra Staglianò, Francesca Odone, and Antonio Camurri. *Adaptive Body Gesture Representation for Automatic Emotion Recognition*. ACM Transactions on Interactive Intelligent Systems, 6(1), mar 2016. 53

[Picard 00] Rosalind W Picard. *Affective computing*. MIT press, 2000. 2, 7, 21, 144, 160

[Picard 01] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. *Toward machine emotional intelligence: Analysis of affective physiological state*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10):1175–1191, 2001. 159

[Poria 16a] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. *Fusing audio, visual and textual clues for sentiment analysis from multimodal content*. Neurocomputing, 174:50–59, 2016. 54

[Poria 16b] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. *Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis*. In 2016 IEEE 16th International Conference on Data Mining (ICDM), pages 439–448, 2016. 54

[Posner 05] 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. 1, 22

[Praveen 21] 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. 54

[Praveen 22] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. *A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition*. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2485–2494, 2022. 50, 54, 131

[Praveen 23] R Gnana Praveen, Patrick Cardinal, and Eric Granger. *Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention*. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023. 54

[Preoţiuc-Pietro 16] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. *Modelling Valence and Arousal in Facebook posts*. In Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andres Montoyo, editors, Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 9–15. Association for Computational Linguistics, 2016. 53

[Puccetti 21] Nikki A Puccetti, William J Villano, and Aaron S Heller. *The neuroscience of affective dynamics*. Affect dynamics, pages 33–60, 2021. 8

[Puccetti 22] Nikki A Puccetti, William J Villano, Jonathan P Fadok, and Aaron S Heller. *Temporal dynamics of affect in the brain: Evidence from human imaging and animal models*. Neuroscience & Biobehavioral Reviews, 133:104491, 2022. 8

[Pumarola 18] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. *Ganimation: Anatomically-aware facial animation from a single image*. In Proceedings of the European Conference on Computer Vision (ECCV), pages 818–833, 2018. 45

[Qi 21] Fan Qi, Xiaoshan Yang, and Changsheng Xu. *Emotion Knowledge Driven Video Highlight Detection*. IEEE Transactions on Multimedia, 23:3999–4013, 2021. 124

[Radwan 13] Ibrahim Radwan, Abhinav Dhall, and Roland Goecke. *Monocular Image 3D Human Pose Estimation under Self-Occlusion*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2013. 40

[Radwan 19] Ibrahim Radwan, Nour Moustafa, Byron Keating, Kim-Kwang Raymond Choo, and Roland Goecke. *Hierarchical Adversarial Network for Human Pose Estimation*. IEEE Access, 7:103619–103628, 2019. 40

[Ranganathan 16] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. *Multimodal emotion recognition using deep learning architectures*. In

2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–9. IEEE, 2016. 54, 131

[Redmon 16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. *You only look once: Unified, real-time object detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016. 40, 41

[Reza 04] Ali M Reza. *Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement*. Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, 38(1):35–44, 2004. 74

[Ringeval 13] 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, 2013. 34, 72

[Rouast 19] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. *Deep learning for human affect recognition: Insights and new developments*. IEEE Transactions on Affective Computing, 12(2):524–543, 2019. 10

[Roy 21] Shuvendu Roy and Ali Etemad. *Spatiotemporal contrastive learning of facial expressions in videos*. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8. IEEE, 2021. 87

[Russell 80] James A Russell. *A circumplex model of affect.* Journal of Personality and Social Psychology, 39(6):1161, 1980. xxvii, 5, 6, 21, 24

[Russell 03] James A Russell. *Core affect and the psychological construction of emotion.* Psychological review, 110(1):145, 2003. 21

[Sato 07] Wataru Sato, Motoko Noguchi, and Sakiko Yoshikawam. *Emotion elicitation effect of films in a Japanese sample.* Social Behavior and Personality: An International Journal, 35(7):863–874, 2007. 26

[Scherer 05] Klaus R Scherer. *What are emotions? And how can they be measured?* Social Science Information, 44(4):695–729, 2005. 3

[Scherer 09] Klaus R Scherer. *The dynamic architecture of emotion: Evidence for the component process model*. Cognition and Emotion, 23(7):1307–1351, 2009. 128

[Scheurer 20] Sebastian Scheurer, Salvatore Tedesco, Brendan O'Flynn, and Kenneth N Brown. *Comparing Person-Specific and Independent Models on Subject-Dependent and Independent Human Activity Recognition Performance*. Sensors, 20(13):3647, 2020. 73

[Schlosberg 52] Harold Schlosberg. *The description of facial expressions in terms of two dimensions.* Journal of Experimental Psychology, 44(4):229, 1952. 21

[Scotty D. Craig 08] Amy Witherspoon Scotty D. Craig Sidney D'Mello and Art Graesser. *Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning*. Cognition and Emotion, 22(5):777–788, 2008. 30

[Sebe 05] Nicu Sebe, Ira Cohen, and Thomas S Huang. Multimodal emotion recognition, pages 387–409. World Scientific, 2005. 3

[Sekhavat 21] Yoones A Sekhavat, Milad Jafari Sisi, and Samad Roohi. *Affective interaction: Using emotions as a user interface in games*. Multimedia Tools and Applications, 80:5225–5253, 2021. 159

[Shen 22] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. *Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition*. IEEE Transactions on Affective Computing, 2022. 87

[Shokri 20] Mohammad Shokri, Ahad Harati, and Kimya Taba. *Salient object detection in video using deep non-local neural networks*. Journal of Visual Communication and Image Representation, 68:102769, 2020. 133, 134

[Shu 18] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. *A review of emotion recognition using physiological signals*. Sensors, 18(7):2074, 2018. 7

[Shukla 17] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Ramanathan Subramanian. *Evaluating Content-Centric vs. User-Centric Ad Affect Recognition*. In International Conference on Multimodal Interaction, page 402–410, New York, NY, USA, 2017. Association for Computing Machinery. 72

[Shukla 22] Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. *Recognition of Advertisement Emotions With Application to Computational Advertising*. IEEE Transactions on Affective Computing, 13(2):781–792, 2022. 72, 159

[Simonyan 15] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In International Conference on Learning Representations, 2015. 42

[Siriwardhana 20] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. *Multimodal emotion recognition with transformer-based self supervised feature fusion*. IEEE Access, 8:176274–176285, 2020. 54

[Sneddon 11] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. *The belfast induced natural emotion database*. IEEE Transactions on Affective Computing, 3(1):32–41, 2011. 36

[Soleymani 14] Mohammad Soleymani, Sadjad Asghari-Esfeden, Maja Pantic, and Yun Fu. *Continuous emotion detection using EEG signals and facial expressions*. In 2014 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2014. 54

[Song 18] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. *Mask-guided contrastive attention model for person re-identification*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1179–1188, 2018. 163

[Stein 92] Nancy L Stein and Keith Oatley. *Basic emotions: Theory and measurement*. Cognition & Emotion, 6(3-4):161–168, 1992. 23

[Strubell 19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650, 2019. 161

[Sujono 15] Sujono and Alexander A.S. Gunawan. *Face Expression Detection on Kinect Using Active Appearance Model and Fuzzy Logic*. Procedia Computer Science, 59:268–274, 2015. International Conference on Computer Science and Computational Intelligence (ICCSCI 2015). 38

[Sun 21] Xuran Sun, Jiabei Zeng, and Shiguang Shan. *Emotion-aware contrastive learning for facial action unit detection*. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pages 01–08. IEEE, 2021. 87

[Sundberg 11] Johan Sundberg, Sona Patel, Eva Bjorkner, and Klaus R Scherer. *Interdependencies among voice source parameters in emotional speech*. IEEE Transactions on Affective Computing, 2(3):162–174, 2011. 7

[Tang 14] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. *Learning sentiment-specific word embedding for twitter sentiment classification*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1555–1565, 2014. 53

[Tao 05] Jianhua Tao and Tieniu Tan. *Affective computing: A review*. In International Conference on Affective Computing and Intelligent Interaction, pages 981–995. Springer, 2005. 2, 7

[Tellamekala 19] Mani Kumar Tellamekala and Michel Valstar. *Temporally coherent visual representations for dimensional affect recognition*. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–7. IEEE, 2019. 50, 72, 87, 104, 131, 132

[Tellamekala 22] Mani Kumar Tellamekala, Timo Giesbrecht, and Michel Valstar. *Modelling stochastic context of audio-visual expressive behaviour with affective processes*. IEEE Transactions on Affective Computing, 2022. 50, 140, 148

[Thayer 90] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990. 21

[Tian 01] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. *Recognizing action units for facial expression analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2):97–115, 2001. 39

[Titchener 09] Edward Bradford Titchener. *The Experimental Psychology of Thought.* In Lectures on the Experimental Psychology of the Thought-processes, pages 157–194. MacMillan Co, New York, NY, US, 1909. 1

[Toisoul 21] 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. 10, 13, 21, 41, 50, 68, 76, 77, 79, 80, 89, 91, 97, 106, 111, 120, 121, 122, 123, 124, 125, 132, 137, 147, 149, 163

[Tomkins 62] Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer Publishing Company, 1962. 5, 22

[Tran 18] 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. 40, 132

[Tran 23] Minh Tran, Yelin Kim, Che-Chun Su, Cheng-Hao Kuo, and Mohammad Soleymani. *SAAML: A Framework for Semi-supervised Affective Adaptation via Metric Learning*. In Proceedings of the 31st ACM International Conference on Multimedia, pages 6004–6015, 2023. 47

[Tyukin 21] Ivan Y. Tyukin, Alexander N. Gorban, Muhammad H. Alkhudaydi, and Qinghua Zhou. *Demystification of Few-shot and One-shot Learning*. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–7, 2021. 107

[Tzirakis 17] 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. 131

[Tzirakis 18] 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. 51

[Um 12] Eunjoon Um, Jan L Plass, Elizabeth O Hayward, Bruce D Homer, et al. *Emotional design in multimedia learning*. Journal of Educational Psychology, 104(2):485, 2012. 3

[Union 24] European Union. *EU AI Act: first regulation on artificial intelligence*. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2024. Accessed: 2024-01-18. 12

[Valstar 06] Michel Valstar and Maja Pantic. *Fully automatic facial action unit detection and temporal analysis*. In 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pages 149–149. IEEE, 2006. 28

[Valstar 07] Michel F Valstar, Hatice Gunes, and Maja Pantic. *How to distinguish posed from spontaneous smiles using geometric features*. In Proceedings of the 9th International Conference on Multimodal Interfaces, pages 38–45, 2007. 28

[Van der Maaten 08] Laurens Van der Maaten and Geoffrey Hinton. *Visualizing data using t-SNE.* Journal of Machine Learning Research, 9(11), 2008. 80

[van Niekerk 22] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. *A comparison of discrete and soft speech units for improved voice conversion*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6562–6566. IEEE, 2022. 44

[Vaswani 17] 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. 43, 162

[Verma 14] Gyanendra K Verma and Uma Shanker Tiwary. *Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals*. NeuroImage, 102:162–172, 2014. 54

[Vogel 16] Susanne Vogel and Lars Schwabe. *Learning and memory under stress: implications for the classroom*. NPJ Science of Learning, 1(1):1–10, 2016. 3

[Walecki 16] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. *Copula ordinal regression for joint estimation of facial action unit intensity*. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pages 4902–4910, 2016. 88

[Wallbott 98] Harald G Wallbott. *Bodily expression of emotion*. European Journal of Social Psychology, 28(6):879–896, 1998. 7

[Wang 04] Yubo Wang, Haizhou Ai, Bo Wu, and Chang Huang. *Real time facial expression recognition with adaboost*. In Proceedings of the 17th International Conference on Pattern Recognition, Volume 3, pages 926–929. IEEE, 2004. 39

[Wang 17] Shu-hui Wang and Chiou-Ting Hsu. *AST-Net: An Attribute-based Siamese Temporal Network for Real-Time Emotion Recognition*. In British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017. BMVA Press, 2017. 87

[Wang 18] 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. 133, 134, 139, 162

[Wang 20] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. *Generalizing from a Few Examples: A Survey on Few-Shot Learning*. ACM Computing Surveys, 53(3), 2020. 105, 106

[Wang 21] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng. *OAENet: Oriented attention ensemble for accurate facial expression recognition*. Pattern Recognition, 112:107694, 2021. 43

[Wang 22] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. *A systematic review on affective computing: Emotion models, databases, and recent advances*. Information Fusion, 83:19–52, 2022. 38

[Ware 18] Shweta Ware, Chaoqun Yue, Reynaldo Morillo, Jin Lu, Chao Shang, Jayesh Kamath, Athanasios Bamis, Jinbo Bi, Alexander Russell, and Bing Wang. *Large-scale automatic depression screening using meta-data from wifi infrastructure*. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(4):1–27, 2018. 159

[Watson 85] David Watson and Auke Tellegen. *Toward a consensual structure of mood*. Psychological Bulletin, 98(2):219, 1985. 21

[Wehrle 00] Thomas Wehrle, Susanne Kaiser, Susanne Schmidt, and Klaus R Scherer. *Studying the dynamics of emotional expression using synthesized facial muscle movements.* Journal of Personality and Social Psychology, 78(1):105, 2000. 26

[Wei 20] Gou Wei, Li Jian, and Sun Mo. *Multimodal (audio, facial and gesture) based emotion recognition challenge*. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 908–911. IEEE, 2020. 54

[Wöllmer 08] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. *Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies*. Interspeech 2008, 2008:597, 2008. 24, 51

[Wöllmer 10] Martin Wöllmer, Björn Schuller, Florian Eyben, and Gerhard Rigoll. *Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening*. IEEE Journal of Selected Topics in Signal Processing, 4(5):867–881, 2010. 51

[Wu 18] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. *Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5177–5186, 2018. 106

[Wu 20] Jinting Wu, Yujia Zhang, Xiaoguang Zhao, and Wenbin Gao. *A Generalized Zero-Shot Framework for Emotion Recognition from Body Gestures*. arXiv preprint arXiv:2010.06362, abs/2010.06362, 2020. 53

[Wu 23] Yujin Wu, Mohamed Daoudi, and Ali Amad. *Transformer-based self-supervised multimodal representation learning for wearable emotion recognition*. IEEE Transactions on Affective Computing, 2023. 52

[Wundt 02] Wilhelm Max Wundt and Charles Hubbard Judd. *Outlines of psychology*. W. Engelmann, 1902. 1, 5

[Xia 15] 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. 88

[Xie 19] Siyue Xie, Haifeng Hu, and Yongbo Wu. *Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition*. Pattern Recognition, 92:177–191, 2019. 43

[Xie 22] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. *An overview of facial micro-expression analysis: Data, methodology and challenge*. IEEE Transactions on Affective Computing, 2022. 159

[Xing 19] Baixi Xing, Hui Zhang, Kejun Zhang, Lekai Zhang, Xinda Wu, Xiaoying Shi, Shanghai Yu, and Sanyuan Zhang. *Exploiting EEG signals and audiovisual feature fusion for video emotion recognition*. IEEE Access, 7:59844–59861, 2019. 54

[Xu 15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. *Show, attend and tell: Neural image caption generation with visual attention*. In International Conference on Machine Learning, pages 2048–2057. PMLR, 2015. 42

[Xu 18] Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. *Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training*. In Alexandra Balahur, Saif M. Mohammad, Veronique Hoste, and Roman Klinger, editors, Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 292–298, Brussels, Belgium, 2018. Association for Computational Linguistics. 53

[Yang 16] Zhaojun Yang and Shrikanth S Narayanan. *Modeling dynamics of expressive body gestures in dyadic interactions*. IEEE Transactions on Affective Computing, 8(3):369–381, 2016. 53

[Yannakakis 18] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. *The ordinal nature of emotions: An emerging approach*. IEEE Transactions on Affective Computing, 12(1):16–35, 2018. 88

[Yosinski 14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. *How transferable are features in deep neural networks?* Advances in Neural Information Processing Systems, 27, 2014. 42, 139

[Yu 04] Chen Yu, Paul M. Aoki, and Allison Woodruff. *Detecting user engagement in everyday conversations*. In INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004, pages 1329–1332. ISCA, 2004. 24

[Zafeiriou 17] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. *Aff-wild: valence and arousal'In-the-Wild'challenge*. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops, pages 34–41, 2017. 36, 66

[Zen 14] Gloria Zen, Enver Sangineto, Elisa Ricci, and Nicu Sebe. *Unsupervised domain adaptation for personalized facial emotion recognition*. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 128–135, 2014. 48

[Zeng 07] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. *A survey of affect recognition methods: audio, visual and spontaneous expressions*. In Proceedings of the 9th International Conference on Multimodal Interfaces, pages 126–133, 2007. 13, 28

[Zhang 16a] Biqiao Zhang, Georg Essl, and Emily Mower Provost. *Automatic recognition of self-reported and perceived emotion: Does joint modeling help?* In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 217–224, 2016. 131

[Zhang 16b] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. *Joint face detection and alignment using multitask cascaded convolutional networks*. IEEE Signal Processing Letters, 23(10):1499–1503, 2016. 74

[Zhang 18] Zhilu Zhang and Mert Sabuncu. *Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels*. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, Volume 31. Curran Associates, Inc., 2018. 98

[Zhang 19] Shiqing Zhang, Xiaoming Zhao, and Qi Tian. *Spontaneous speech emotion recognition using multiscale deep convolutional LSTM*. IEEE Transactions on Affective Computing, 13(2):680–688, 2019. 28

[Zhang 21] Shiqing Zhang, Xin Tao, Yuelong Chuang, and Xiaoming Zhao. *Learning deep multimodal affective features for spontaneous speech emotion recognition*. Speech Communication, 127:73–81, 2021. 28

[Zhang 22] Tianyi Zhang, Abdallah El Ali, Alan Hanjalic, and Pablo Cesar. *Few-shot learning for fine-grained emotion recognition using physiological signals*. IEEE Transactions on Multimedia, 2022. 106

[Zhang 23a] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. *Multi-Task Learning Framework for Emotion Recognition In-the-Wild*. In European Conference on Computer Vision, pages 143–156. Springer, 2023. 88

[Zhang 23b] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. *Facial Affect Recognition based on Transformer Encoder and Audiovisual Fusion for the ABAW5 Challenge*. arXiv preprint arXiv:2303.09158, 2023. 44

[Zhao 11] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. *Facial expression recognition from near-infrared videos*. Image and Vision Computing, 29(9):607–619, 2011. 33

[Zhao 18] 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. 49

[Zhao 22] Rui Zhao, Tianshan Liu, Zixun Huang, Daniel PK Lun, and Kin-Man Lam. *Spatial-Temporal Graphs Plus Transformers for Geometry-Guided Facial Expression Recognition*. IEEE Transactions on Affective Computing, 2022. 80

[Zhou 17] Yuqian Zhou and Bertram Emil Shi. *Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder*. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 370–376. IEEE, 2017. 45

[Zhou 18] Zhi-Hua Zhou. *A brief introduction to weakly supervised learning*. National Science Review, 5(1):44–53, 2018. 47, 85

[Zhou 23] Enting Zhou, You Zhang, and Zhiyao Duan. *Learning Arousal-Valence Representation from Categorical Emotion Labels of Speech*. arXiv preprint arXiv:2311.14816, 2023. 24

[Zhu 17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. *Unpaired image-to-image translation using cycle-consistent adversarial networks*. In Proceedings of the IEEE International Conference on Computer Vision, pages 2223–2232, 2017. 46

[Zhu 18] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. *Emotion Classification with Data Augmentation Using Generative Adversarial Networks*. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, Advances in Knowledge Discovery and Data Mining, pages 349–360. Springer International Publishing, 2018. 45

[Zou 22] Xinyi Zou, Yan Yan, Jing-Hao Xue, Si Chen, and Hanzi Wang. *When Facial Expression Recognition Meets Few-Shot Learning: A Joint and Alternate Learning Frame-*

work. Proceedings of the AAAI Conference on Artificial Intelligence, 36(5):5367–5375, Jun. 2022. 106

[Zou 23] Peng Zou, Rui Wang, Kehua Wen, Yasi Peng, and Xiao Sun. *Spatial-temporal transformer for affective behavior analysis*. arXiv preprint arXiv:2303.10561, 2023. 44