University of Nottingham
UK | CHINA | MALAYSIA

Towards Uncertainty-Aware and Label-Efficient Machine Learning of Human Expressive Behaviour

Thesis submitted to the University of Nottingham for the degree of **Doctor of Philosophy, November 15, 2022.**

Mani Kumar Tellamekala

14342775

Supervised by

Prof. Michel Valstar
Prof. Andrew French

Signature _____

Date _____ / _____ / _____

Abstract

The ability to recognise emotional expressions from non-verbal behaviour plays a key role in human-human interaction. Endowing machines with the same ability is critical to enriching human-computer interaction. Despite receiving widespread attention so far, human-level automatic recognition of affective expressions is still an elusive task for machines. Towards improving the current state of machine learning methods applied to affect recognition, this thesis identifies two challenges: *label ambiguity* and *label scarcity*.

Firstly, this thesis notes that it is difficult to establish a clear one-to-one mapping between inputs (face images or speech segments) and their target emotion labels, considering that emotion perception is inherently subjective. As a result, the problem of label ambiguity naturally arises in the manual annotations of affect. Ignoring this fundamental problem, most existing affect recognition methods implicitly assume a one-to-one input-target mapping and use deterministic function learning. In contrast, this thesis proposes to learn non-deterministic functions based on uncertainty-aware probabilistic models, as they can naturally accommodate the one-to-many input-target mapping. Besides improving the affect recognition performance, the proposed uncertainty-aware models in this thesis demonstrate three important applications: adaptive multimodal affect fusion, human-

in-the-loop learning of affect, and improved performance on downstream behavioural analysis tasks like personality traits estimation.

Secondly, this thesis aims to address the challenge of scarcity of affect labelled datasets, caused by the cumbersome and time-consuming nature of the affect annotation process. To this end, this thesis notes that audio and visual feature encoders used in the existing models are label-inefficient i.e. learning them requires large amounts of labelled training data. As a solution, this thesis proposes to pre-train the feature encoders using unlabelled data to make them more label-efficient i.e. using as few labelled training examples as possible to achieve good emotion recognition performance. A novel self-supervised pre-training method is proposed in this thesis by posing hand-engineered emotion features as task-specific representation learning priors. By leveraging large amounts of unlabelled audiovisual data, the proposed self-supervised pre-training method demonstrates much better label efficiency compared to the commonly employed pre-training methods.

Acknowledgements

I am incredibly grateful to my supervisor, Prof. Michel Valstar, without whom this thesis would not have been possible. I sincerely thank him for sharing his wisdom and encouraging me to be thoughtful and to take risks. He provided me with the opportunities to collaborate with some amazing people in the field of Affective Computing. Thank you Michel for your kind support throughout my PhD, and for generously supporting me in my research career.

Many thanks to Dr. Timo Giesbrecht from Unilever, my second supervisor Prof. Andrew French, and my annual review panel: Dr. Xin Chen and Dr. Isaac Triguero, for their critical feedback on my work. Also I am especially grateful to Dr. Enrique Sánchez-Lozano for his valuable guidance and advice in developing some important ideas presented in this thesis. Thank you so much Kike for helping me learn how to rigorously and critically validate a research idea.

I had an amazing opportunity to closely work with Prof. Elisabeth André and Prof. Björn Schuller, and spend four wonderful months at the University of Augsburg. During this time, I was fortunate to work with some very nice colleagues: Dr. Shahin Amiriparian, Dr. Tobias Baur, Dr. Ömer Sümer, and Dominik Schiller. For their nice company and great support during my PhD, I thank all the CVL colleagues: Keerthy, Siyang, Dimitris, Aaron, Shashank, Joy, Zane, Ioanna, and others.

I would like to specially thank SK Mahammad Rafi, Dr. KRS Chandra Kumar, Dr. N. Ramakrishna Reddy, and Dr. Jayachandra Dakala for encouraging me to pursue a career in research. Finally, I thank my family members for their valuable support during the tough times of my PhD, and I dedicate this thesis to my parents, without whom none of this would have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Nothing in life is to be feared.
>
> It is only to be understood.
>
> *Marie Curie*

Humans are inherently social creatures. In regulating our social interactions, expressing emotions through nonverbal behaviour is an integral part. Reeves and Naas, in their seminal work 'The Media Equation' [Reeves and Nass, 1996], notes that our emotional expressions are not confined to just human-human interactions; we tend to treat computers as real people and implicitly exhibit a social attitude towards computers too. With the rapidly expanding role of computing devices in managing our lives, the idea of enabling machines to recognise emotional expressions has been becoming more and more relevant ever since the first attempt made by [Parke, 1974] in 1974.

But why do machines need to learn about human emotional expressive behaviour? Endowing computers with the ability to recognise users' apparent emotions and choose their responses accordingly holds the potential

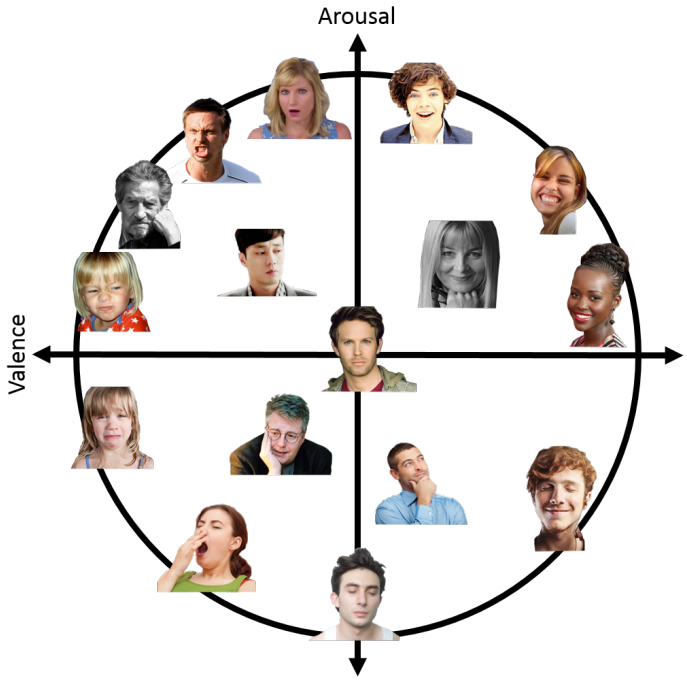


Figure 1.1: Circumplex model of dimensional affect [Russell, 1980] (Image source: AffectNet corpus [Mollahosseini et al., 2017])

to enrich human-computer interactions. Further, making computers aware of human emotional behaviour is critical to build technology with a positive societal impact through applications ranging from diagnostic tools for mental health and well-being problems [Valstar et al., 2016, Yannakakis, 2018] to personalised and interactive educational tools [Wu et al., 2016, Yadegaridehkordi et al., 2019].

With the objective of making machines emotionally intelligent, Affective Computing [Picard, 1997] aims to develop tools that provide machines with the abilities to perceive, analyse and respond to human affective expressions. The fundamental notion of 'Affect' broadly encompasses a person's basic sense of feeling, and it is not specific to just emotions. In its dimensional representation [Russell, 1980], affect characterises a feeling in terms of how pleasant or unpleasant it is (valence), and how active or passive it is (arousal), as illustrated in Fig. 1.1. Unlike the commonly used basic emotion categories such as happy, angry, disgust, etc [Ekman, 1999], the

dimensional affect model composed of valence and arousal axes can capture a wider spectrum of complex and nuanced emotions (e.g. depression, content, etc).

Automatic recognition of affective states (valence and arousal levels) from non-verbal behaviour lies at the core of building affect-aware interfaces. Considering that affective expressions are inherently multimodal phenomena [Zeng et al., 2008], collecting and integrating behavioural cues from multiple channels is essential for building reliable affect recognition systems. To this end, facial (visual) and vocal (audio) modalities evolved as the most favourable channels for recognising affective states [Zeng et al., 2007], given that facial and vocal expressions are the two most dominant modalities that humans use in communicating their affective states, and due to the ubiquity of video cameras and microphones in human-computer interactions. Although wearable sensors also could be used for recognising affect from other modalities such as electrodermal Activity (EDA) and electrocardiogram (ECG), video and audio modalities are more preferable in practice due to their unobtrusive nature, ease-of-use and scalablity advantages.

From a mathematical standpoint, training a machine learning model to recognise affect involves essentially learning a function $f : X \rightarrow Y$ that maps face videos and/or speech signals (X) to manually annotated valence and arousal vectors (Y). The performance of a trained machine learning (ML) model is measured w.r.t how well it generalises to unseen data. Here the notion of 'generalisation' refers to how well the learned function f performs when presented with novel inputs (X'). In recent years, data-driven end-to-end ML powered by Deep neural networks (DNNs) [LeCun et al., 2015] as universal function approximators, demonstrated impressive results on a wide range of perceptual tasks such as image classification [Krizhevsky

et al., 2012], object detection [Szegedy et al., 2013a], etc. Leveraging the advancements in data-driven ML, contemporary affect recognition models also demonstrated good generalisation performance by applying DNNs to face and voice data collected in naturalistic conditions (e.g. [Tzirakis et al., 2017]).

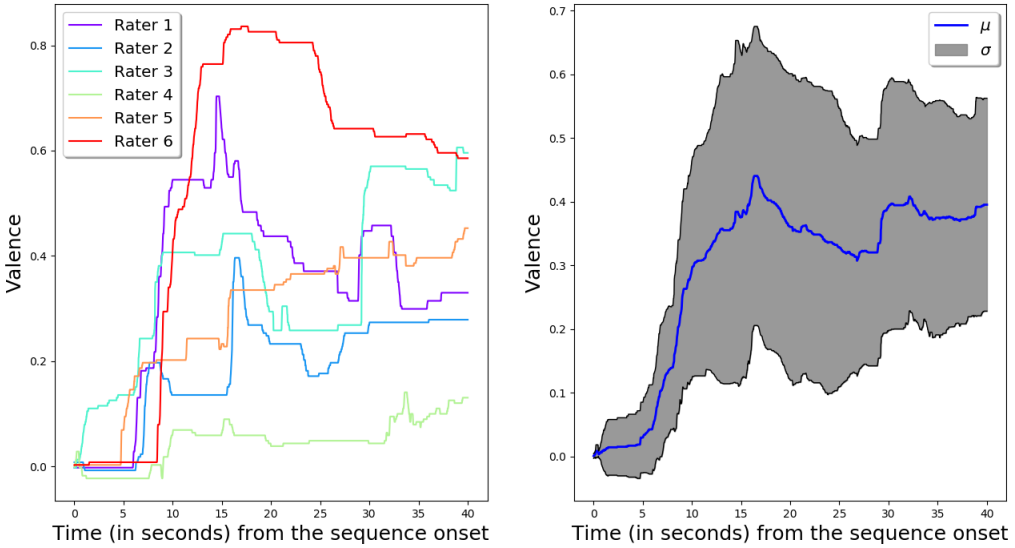
However, most existing affect recognition approaches largely ignore a fundamental difference between general perceptual tasks and apparent affect recognition: ambiguity in the ground truth labels. Given an input signal X_i, the ambiguity in manually assigning its ground truth label refers to the condition in which several label classes or values $(Y_i^a, Y_i^b, ...)$ are likely to be correct. In tasks such as object recognition the human supervision is largely unambiguous and objective in nature, whereas in affect recognition the ground truth labels are strongly influenced by the subjective nature of emotion experience, expression or communication and perception processes in humans [LeDoux and Hofmann, 2018].

Since emotions are inherently latent (not directly observable) psychological constructs that are highly context-sensitive [Barrett et al., 2011], the perception of the same expressed emotion is likely to differ significantly from rater to rater. To give an example, if we ask two different annotators to label the identities of clearly visible objects (e.g. bicycle and car) in an image, it is unlikely that the two annotators will provide different labels for the same object, considering the unambiguous nature of label classes i.e. object identities. But when annotating valence and arousal levels, it is common to notice disagreements among different annotators [Busso et al., 2008, Devillers et al., 2005, Douglas-Cowie et al., 2005]. For instance, as Fig. 1.2 illustrates, in the continuous-valued ratings of valence and arousal labelled by six different annotators in the SEMAINE corpus [McKeown et al., 2010], we can clearly see that the valence and arousal annotations have high vari-

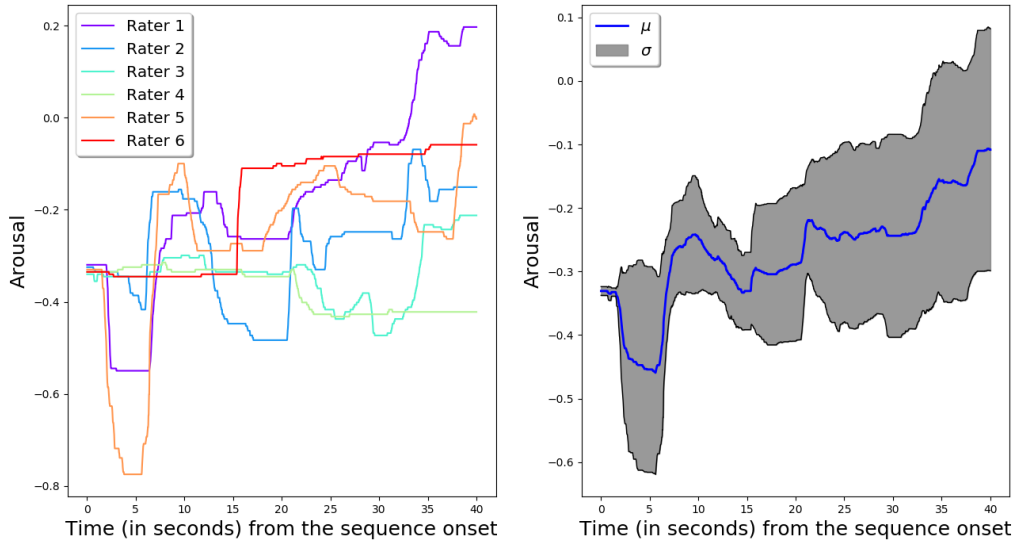
ability from rater to rater, inducing ambiguity into the final ground truth labels (wide variance ranges around the mean curves) [Mower et al., 2009]. It is interesting to note that while all annotators agreed that the valence increased at around 6 seconds, they disagreed about the exact timing of the rise, and how strong the rise was. This disagreement among the raters in the event of sudden changes in emotional states, has been widely studied in the literature of dimensional affect recognition (e.g. [Cowie et al., 2012]).

Another major source of affect annotation ambiguities lies in the less accurate emotion representation models. [Sethu et al., 2019] highlights the limitations of existing emotion representation models and strongly advocates the need for accommodating ambiguity into the existing models. Based on an adapted version of Brunswik's function lens model [Scherer, 2003], they present a theoretical framework, as illustrated in Fig. 1.3, that delineates the role of affect annotation ambiguity at different stages of the annotation process (experience, expression and perception), and the uncertainty introduced into the the machine learning models trained for emotion recognition.

In contrast to all the aforementioned theoretical arguments, existing affect recognition datasets [Kossaifi et al., 2019, Ringeval et al., 2019, Busso et al., 2008] ignore the affect annotation ambiguities, and assume that the variability among the emotion annotations can be modelled as mere *noise*. Guided by this assumption, the process of affect annotation is designed generally to minimise the *label noise* (variance of ratings), with the goal of maximising inter-rater-reliability score. First, a small of pool of raters annotate each video and/or audio input with affect ratings. Then, the ground truth labels of affect are prepared by averaging the annotations collected from all the raters, using techniques such as evaluation-weighted-estimation [Ringeval et al., 2017, 2018, 2019].



(a)



(b)

Figure 1.2: Rater-wise annotations of (a). valence and (b). arousal for an example sequence from the SEMAINE corpus [McKeown et al., 2010]

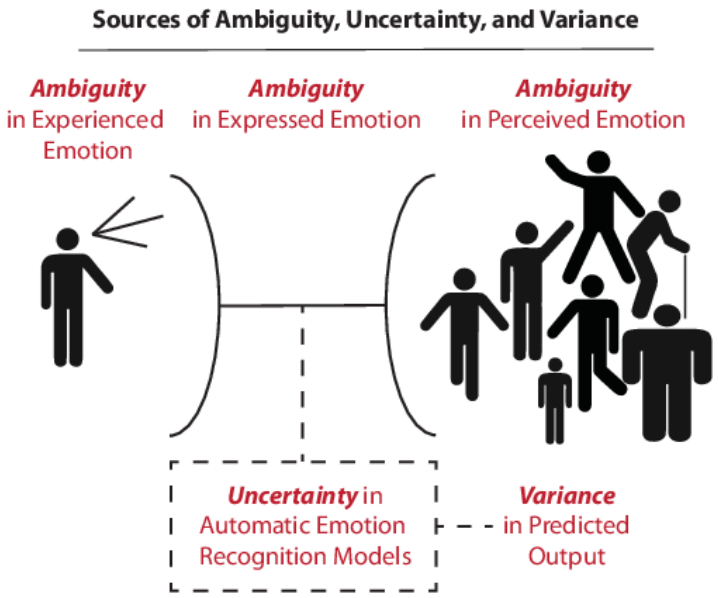


Figure 1.3: Brunswik's functional lens model of ambiguities in emotion experience, expression and perception stages, and the uncertainty introduced into the emotion recognition models (Image source: [Sethu et al., 2019]).

Consequently, annotating the ground truth labels of affect, especially for large-scale datasets, is often a highly laborious and time-consuming process. Thus, when training end-to-end ML models for affect recognition for novel use cases (e.g. driving environments), affect label scarcity is a commonly encountered problem. Furthermore, despite collecting ratings from multiple annotators, it is less likely that the final ground truth labels are completely unambiguous, given the sub-optimal nature of simple weighted averaging techniques which do not model the variability among the raters systematically. The ML models trained using such ambiguous supervision signals of affect are bound to have poor generalisation performance. Therefore, this thesis argues that when training ML models for affect recognition it is important to account for the problems of *Label Ambiguity and Label Scarcity*, in order to advance the current state of affect recognition models.

1.1 Research Objectives

Most existing ML models applied to affect recognition tasks can be characterised as deterministic function learning models, which assume that the ground truth labels are completely unambiguous. Such models expect a clear one-to-one functional mapping between the inputs and target labels. On the contrary, the input-label mapping in the case of affect recognition does not naturally fit into the one-to-one type, for the reasons discussed above. Darwin, in his work on emotions [Darwin, 1948], notes this fundamental problem: "*the observation of Expression is by no means easy. Hence it is difficult to determine, with certainty, what are the movements of the features and of the body, which commonly characterize certain states of the mind*". Motivated by these observations, this thesis explores non-deterministic function learning models [Del Coz et al., 2009] through uncertainty modelling which allows us to map an input video or audio signal to a range of affect labels.

On the other hand, to deal with the problem of label scarcity in affect recognition it is important to make the end-to-end affect recognition models label-efficient i.e. achieve good generalisation performance using small amounts of labelled training data. In order to make the affect recognition models label-efficient, this thesis proposes to leverage the natural supervision signals embedded in the audiovisual data. To this end, by building on the recent advancements in self-supervised representation learning [Jing and Tian, 2020], this work proposes a novel self-supervised pre-training method for improving the label-efficiency of affect recognition models.

In summary, this thesis aims to address the above discussed two fundamental challenges in automatic affect recognition, label ambiguity and scarcity, by exploring two key ideas: 1. Uncertainty-aware learning – to learn non-

deterministic ML models that are aware of label ambiguity through probabilistic temporal modelling and 2. Label-efficient learning – to minimise the requirement of labelled data for learning expressive features from high-dimensional face and voice data. The following section presents a brief summary of the approaches explored towards uncertainty-aware and label-efficient affect recognition.

1.2 Proposed Solutions

Uncertainty Modelling in the Temporal Context of Affect. This thesis proposes to leverage the temporal context information in face videos and speech signals, to partially account for the affect label ambiguity problem. Particularly, novel probabilistic temporal models are proposed for non-deterministic function learning, in which a single input signal is mapped to multiple output affect values by predicting distributions, in stead of points estimates. To this end, the uncertainty of latent states and output states is modelled in the temporal affect recognition models.

First, for latent uncertainty modelling, this thesis proposes two probabilistic frameworks: Calibrated and Ordinal Latent Distributions (COLD) and Affective Processes (APs). Both these frameworks aim to capture uncertainty in the temporal context of affect signals, but with different assumptions about the underlying temporal context distributions. Further, applications of latent state uncertainty to audiovisual affect fusion and human-in-the-loop affect learning are demonstrated. Then, for predictive or output uncertainty modelling, an approach to quantify epistemic and aleatoric predictive uncertainties is presented and its application to an important downstream behavioural analysis task, apparent personality traits

recognition is demonstrated.

Label-Efficient Affect Representation Learning. This thesis proposes to leverage unlabelled audiovisual data for label-efficient affect recognition i.e for reducing the requirements of labelled examples for representation learning – extracting low dimensional features from high dimensional raw inputs. To this end, this thesis proposes to pre-train the audio and visual feature encoders from unlabelled audiovisual data using a novel self-supervised learning approach, which involves learning a proxy task for which the labels are automatically derived by exploiting the intrinsic structure of face and voice data. In contrast to the affect-agnostic priors used in the existing proxy tasks of self-supervised learning, the proposed proxy task based on deep temporal clustering exploits hand-engineered emotion features by posing them as task-specific representation learning priors. Compared to the generic learning priors like temporal predictability, the proxy task proposed in this thesis demonstrates superior label-efficiency results.

Thus, towards advancing the current state of automatic affect recognition, this thesis presents uncertainty-aware temporal models and label-efficient feature encoders to overcome an important challenge in affective behaviour analysis, ambiguity and scarcity of manual supervision.

1.3 Thesis Outline

This thesis is structured as follows:

- Chapter 2 first reviews the standard computational models of affect representation, different machine learning approaches applied to affect recognition and their limitations. Then, it discusses an important

trade-off between label ambiguity and label scarcity in affect recognition tasks. Towards addressing the affect label ambiguity problem, various non-deterministic temporal function learning methods are reviewed. With the objective of making affect recognition models label-efficient, this chapter reviews different self-supervised representation learning methods applied to affect recognition.

- Chapter 3 introduces a latent uncertainty modelling step in temporal networks based on the canonical recurrent models. In particular, this work proposes a non-deterministic temporal model for uncertainty-aware audiovisual fusion for affect recognition – 'COLD Fusion: Calibrated Ordinal Latent Distributions Fusion'. In this method the vector form hidden state (or context) in RNNs is replaced with a distribution form hidden state whose variance is constrained by the calibration and ordinal ranking properties. This work demonstrates that multimodal affect fusion performance can be improved significantly by adopting the non-deterministic temporal context learning through uncertainty modelling.

- Chapter 4 proposes another latent uncertainty modelling method, 'Affective Processes (APs)', as a more efficient alternative to the COLD fusion model. APs build on recently proposed neural latent variable models using an encoder-decoder composition [Garnelo et al., 2018b,a], and learn a global latent variable for learning the non-deterministic temporal context. It demonstrates an application of APs to audiovisual affect fusion and it shows that by means of learning a global latent variable, APs outperform the COLD fusion method and other standard model-agnostic fusion baselines. Further, an application of APs to Cooperative Machine Learning of affect recognition is proposed.

- Chapter 5 discusses a predictive or output uncertainty model with the aim to capture holistic uncertainty in video-based affect recognition models. Further, it demonstrates an application of predictive uncertainty estimates of the valence and arousal dimensions to an important downstream behavioural analysis task, apparent personality traits estimation.

- Chapter 6 presents a novel self-supervised representation learning approach to improve the label-efficiency of affect recognition from face and voice data. The proposed self-supervised pre-training method, dubbed 'CHeF: Clustering of Hand-engineered Emotion Features', leverages large amounts of unlabelled data by making use of their hand-crafted audiovisual features of affect as task-specific representation learning priors. Compared to the existing self-supervised baselines guided by generic representation learning priors that are affect-agnostic, the proposed CHEF pre-training demonstrates superior label-efficiency results in both visual and audio modalities.

- Chapter 7 concludes this thesis by summarising its key contributions made towards advancing the current state of machine learning models applied to automatic affect recognition.

1.4 Publications

1. **Mani Kumar Tellamekala**, Timo Giesbrecht, Michel Valstar *Modelling Stochastic Context of Audio-Visual Expressive Behaviour with Affective Processes*, IEEE Transactions on Affective Computing 2022 [*Link*]

2. **Mani Kumar Tellamekala**, Timo Giesbrecht, Michel Valstar *Di-*

mensional Affect Uncertainty Modelling and its Application to Personality Recognition, IEEE Transactions on Affective Computing, 2022. [*Link*]

3. **Mani Kumar Tellamekala**, Timo Giesbrecht, Michel Valstar *Apparent Personality Recognition from Uncertainty-Aware Facial Emotion Predictions using Conditional Latent Variable Models*, IEEE FG 2021 [*Link*]

4. **Mani Kumar Tellamekala**, Enrique Sanchez, Georgios Tzimiropoulos, Timo Giesbrecht, Michel Valstar *Stochastic process regression for cross-cultural speech emotion recognition*, INTERSPEECH 2021[***Best student paper award candidate*] [*Link*].

5. Enrique Sanchez, **Mani Kumar Tellamekala**, Michel Valstar, Georgios Tzimiropoulos, *Affective Processes: stochastic modelling of temporal context for emotion and facial expression recognition*, CVPR 2021.[*Link*]

6. **Mani Kumar Tellamekala**, Michel Valstar, Michael Pound, Timo Giesbrecht, *Audio-visual predictive coding for self-supervised visual representation learning*, ICPR 2020.[*Link*]

7. **Mani Kumar Tellamekala**, Michel Valstar, *Temporally coherent visual representations for dimensional affect recognition*, Affective Computing and Intelligent Interaction 2019.[*Link*]

8. S. Song, E. S. Lozano, **Mani Kumar Tellamekala**, Linlin Shen, Alan Johnston, Michel Valstar, *Dynamic facial models for video-based dimensional affect estimation*, ICCV-W, 2019, [*Link*]

9. Vincent Karas, **Mani Kumar Tellamekala**, Adria Mallol-Ragolta, Michel Valstar, Bjorn W. Schuller *Time-Continuous Audiovisual Fu-*

sion with Recurrence vs Attention for In-The-Wild Affect Recognition, CVPR-W, 2022. [*Link*]

10. **Mani Kumar Tellamekala**, Shahin Amiriparian, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar *COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition* ArXiv preprint arXiv:2206.05833, 2022. [*Link*]

11. **Mani Kumar Tellamekala**, Ömer Sümer, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar *Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?* ArXiv preprint arXiv:2207.01113, 2022. [*Link*]

12. *CHeF: Clustering of Hand-Engineered Features for Self-supervised Learning of Expressive Behaviour from Faces and Voices*, [Manuscript in preparation].

Chapter 2

Background and Motivation

> All learning has an emotional base
>
> *Plato*

In general, intelligent behaviour can be defined as the ability to perceive, learn, and adapt to an external environment. This common view of intelligence fails to account for an important set of abilities human beings naturally demonstrate with ease in their daily lives - emotional intelligence [Salovey and Mayer, 1990], which constitutes the abilities to have, communicate, perceive and process affective states[1]. Conveying affective states by modulating nonverbal cues (e.g. facial and vocal expressions, body gestures), plays a crucial role in not only enriching human communication but also in motivating human actions.

Human-human communication heavily depends on nonverbal cues for conveying affective states. A famous work by [Mehrabian, 1968] posited that when the spoken word and the expressed behaviour seem to contradict each

[1] *Affect* and *emotion* are synonymously used throughout this work

other, emotion information communicated in a message relies 55% on facial expressions, 38% on vocal utterances, and only 7% on spoken words. This trend clearly shows the importance of perceiving paralinguistic signals of a message in communicating the affective states through facial and vocal expressions.

As computing devices are becoming more and more ubiquitous, making human-machine communication as natural as possible, has evolved as an important research problem. To solve this problem, the field of Affective Computing [Picard, 1997] aims to develop computing tools specialised in human affective behavioural analysis. Automatic recognition of apparent affect from face and voice data is the central problem in affective computing. This thesis notes two fundamental challenges, label ambiguity and label scarcity, towards solving the problem of naturalistic affect recognition.

Chapter Summary. This chapter first reviews the standard computational models used for quantitatively representing the affective states, followed by a discussion of the key trends in machine learning approaches applied to affect-related feature extraction and temporal modelling of affective signals. Then, the focus shifts to dissecting an interesting challenge posed by human perception uncertainty in annotating apparent emotions: a trade-off between label ambiguity and label scarcity problems. To deal with the former problem, this thesis explores a class of machine learning models with non-deterministic function learning abilities, as they allow learning from ambiguous supervision signals. This chapter discusses various existing non-deterministic ML models and their limitations. To cope with the label scarcity problem, a recently emerging paradigm of Self-Supervised Learning is proposed as a solution. A review of the existing affect recognition approaches that leveraged self-supervision so far and their limitations are discussed at the end.

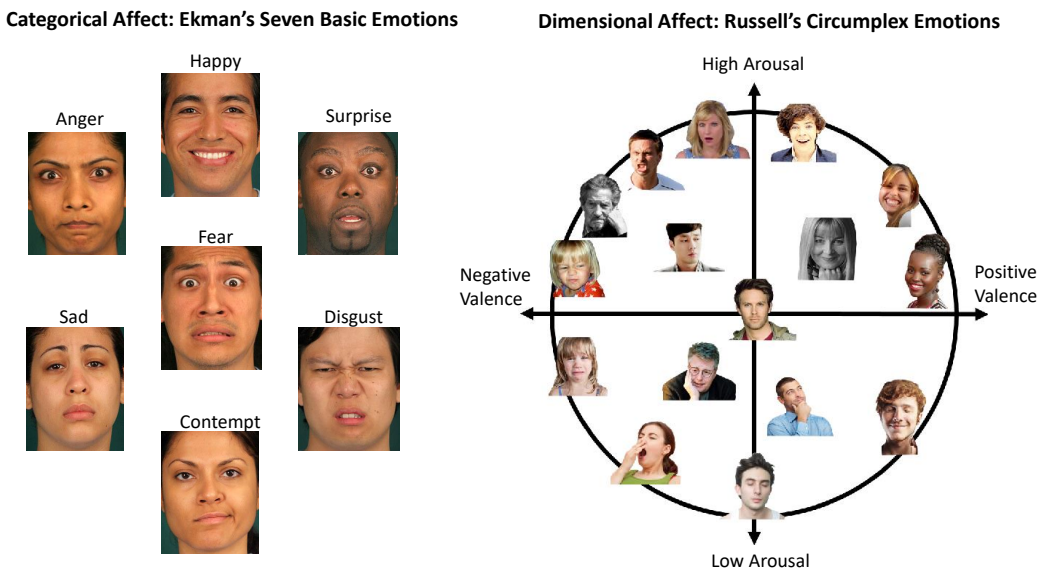


Figure 2.1: Affect Representations Models: Categorical vs. Dimensional

2.1 Computational Models of Affect

The process of human experience, expression and perception of emotions is inherently subjective and latent in nature [Devillers et al., 2005, Liscombe et al., 2003, Mower et al., 2009], hence, quantitatively representing affective states is a fundamental challenge. While no universal descriptive model of emotions is found yet in the literature, two popular models, categorical and dimensional, are widely adopted to mathematically represent emotional states in a machine.

Categorical Model of emotions uses the nominal approach to associate a data instance with a discrete label representing a particular emotion class. Ekman's model of seven basic emotions [Ekman, 1999], as illustrated in Fig. 2.1, is a classic categorical model that was widely used in the early works of affect recognition from facial and vocal expressions [Pantic and Rothkrantz, 2003]. Despite achieving impressive predictive accuracy, affect recognition models based on categorical emotions have not much impact in real-world applications. They were mostly evaluated on data containing

prototypical and exaggerated emotional displays of fixed intensity, whereas the emotions encountered in natural human interactions are often non-prototypical, highly nuanced and continually changing in intensity with time. Thus, there are two main limitations for the basic categorical models in capturing emotional expressions that occur in naturalistic interactions: 1. emotional states are rarely independent and mutually exclusive and 2. the intensity of emotions varies continually in time.

In real-world situations, we often experience complex emotions that are composed of overlapping states of basic emotions simultaneously [Cowie and Cornelius, 2003]. To give an example, the combinations of happiness and surprise, and fear and surprise classes are not mutually exclusive and they can co-occur very often, as illustrated in Fig. 2.2. To address this limitation, some works proposed later advocated the use of compound extensions of categorical models [Du et al., 2014] that are defined using primary and secondary class combinations and describe the intensities of each prototypical emotion (e.g. Plutchik's emotion wheel [Plutchik and Kellerman, 2013]). Using such secondary emotions and their intensities as additional targets improved the applicability of emotion recognition models [Du et al., 2014], however, such methods were not actively adopted in the later works due to their highly complex annotation processes.

Dimensional Models of affect use the interval measurement approach for quantifying emotional states which are assumed to be neither independent nor discrete, in contrast to the nominal approach used in the categorical model. Derived from the factor analysis of self-reported emotional attributes, Russell's circumplex model [Russell, 1980] is a popular candidate among the dimensional models, in which an emotional state is modelled as a point on two-dimensional bipolar space composed of two orthogonal axes: valence and arousal. In this model, for a given emotion, the axis of

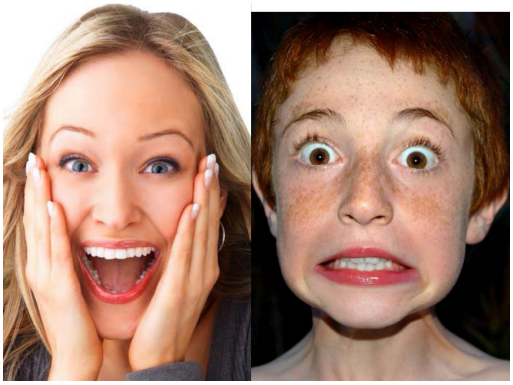


Figure 2.2: Overlapping states of categorical emotion classes (left – happy and surprise, right – fear and surprise), demonstrating that basic emotional states are not mutually exclusive.

valence indicates the degree of its pleasantness (sadness to happiness) and arousal shows the degree of its activeness (sleepiness to excitement). Thus, emotional states are represented as continuous-valued vectors living in a two dimensional space in the dimensional affect model.

In terms of unambiguously representing a wider range of emotional expressions encountered in natural conditions, dimensional affect models are found to be superior to categorical models in general. But, in practice it is common to observe that some of the basic emotions get assigned similar valence and arousal values, as shown in [Sethu et al., 2019]. In such cases, adding dominance as a third dimension may help in resolving the ambiguity. However, as the existing datasets annotated with dimensional affect have very few samples with such overlapped emotional state, not much attention has been paid to the limitations of dimensional affect in terms of distinctly characterising all emotional states. To annotate face videos and speech signals like time-series data with valence and arousal, tools such as FeelTrace [Cowie et al., 2000] have been developed, which output continuous valued 2D vectors defined within the fixed intervals (e.g. [-1, +1]) for training the machine learning models.

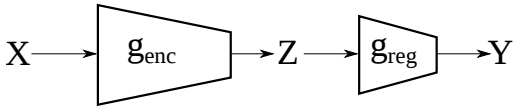


Figure 2.3: Key components of a supervised machine learning model of affect recognition (g_{enc} and g_{reg} denote the feature extraction and temporal regression respectively.)

2.2 Facial and Vocal Affect Recognition

Most existing automatic affect recognition models are primarily based on visual and audio signals, considering the crucial role played by facial and vocal expressions in communicating emotional states [Mehrabian, 1968]. Early attempts to analyse emotions in facial and vocal expressions were first made by [Suwa, 1978] and [Williams and Stevens, 1972] respectively. Since the late 1990s, an increasing number of machine learning approaches were proposed for recognising affect from face images [Mase, 1991, Kobayashi and Hara, 1993] as well as speech signals [Dellaert et al., 1996]. The work done by [Chen et al., 1998] is one of the early efforts towards fusing audio and visual cues for affect recognition. For an exhaustive survey of the early trends in affect recognition, the reader is referred to [Pantic and Rothkrantz, 2003]. Although most of the initial attempts were mainly confined to recognising deliberate affective displays of prototypical emotions, later the focus shifted to spontaneous affect recognition [Zeng et al., 2008]. Refer to [Rouast et al., 2019] for a comprehensive review of the recent developments in spontaneous affect recognition based on advancements in fully supervised machine learning approaches. The task of learning automated recognition of apparent affect from faces and voices, essentially translates to training a machine learning model guided by human-supervision in the form of manually defined emotion labels using either categorical or dimensional models. Most affect recognition methods proposed in the literature are based on the paradigm of supervised machine learning.

Supervised Machine Learning. In this approach, the learning objective is to recover a function map $f : X \rightarrow Z \rightarrow Y$, where X indicates a raw input face image sequence or a speech segment, Y denotes its apparent affective state label and Z represents a facial or vocal feature vector into which affect-specific information is distilled from its corresponding raw input. Similar to any other supervised ML problem formulation, as illustrated in Fig. 2.3, affect recognition can decomposed into two sub problems: (1). **Feature Extraction:** Extracting the most informative low-dimensional feature representations from high-dimensional raw face or voice data ($X \rightarrow Z$) and (2). **Temporal Modelling:** Predicting the target affective states from the temporal dynamics of low-dimensional representations ($Z \rightarrow Y$).

2.2.1 Trends in Affect Feature Extraction: From Engineering to Learning

Facial Affect Features

Feature Engineering. Early approaches to facial affect recognition were mainly based on handcrafted feature representations, which can be broadly grouped into two categories: geometric and appearance. As geometric features, explicit knowledge of face salient points or popularly known as 2D face landmarks, and shapes of individual facial structures (mouth, nose, eyes, etc.) were heavily used. Since facial actions like mouth dimpler cause changes in the appearance only, it may be theoretically impossible to detect all facial displays using geometric features alone. Typical examples of facial affect analysis methods that adopted such geometric facial features include [Chang et al., 2006], [Pantic and Rothkrantz, 2004, Pantic and Bartlett, 2007] and [Kotsia and Pitas, 2006].

Whereas the appearance features of facial expressions were computed using either low-level descriptors of edge distributions or template models. Gabor wavelets used in [Bartlett et al., 2002, 2005, 2006] and [Guo and Dyer, 2005], and Haar features used in [Whitehill and Omlin, 2006] are some notable examples in the former group. Whereas the template-based appearance features include holistic spatial ratio (e.g. [Anderson and McOwan, 2006]) and temporal face templates (e.g. [Valstar et al., 2004]), etc. Later, several works found that fusing both geometric and appearance features helps in improving the facial affect recognition performance. Active Appearance Model (AAM) of faces used in [Lucey et al., 2007] is one such fusion method that combines the key features of face shape and appearance for analysing facial expressions. In spite of showing promising affect recognition performance, most of the hand-engineered facial affect features were found to be not suitable for a wider range of head poses. It is commonly observed that facial appearance features based on low-level descriptors of edge distributions show poor performance when applied to the face videos recorded in in-the-wild conditions, and naturalistic non-frontal head poses [Bartlett et al., 2006]. Refer to [Sariyanidi et al., 2014] for a detailed review of the limitations of different hand-engineered facial affect features.

Feature Learning. Overcoming the limitations of image feature engineering, AlexNet [Krizhevsky et al., 2012], a Deep Convolutional Neural Network (CNN), demonstrated the potential of (labelled) data-driven spatial feature learning in 2012. Based on the premise of end-to-end learning of hierarchical image representations through multi-layered neural network architectures, CNNs transformed the field of Computer Vision. The application of CNNs to facial expression analysis was already explored in some of the early works [Fasel, 2002a,b], much before the revival of deep representation learning of image data in 2012 by AlexNet. However, the

lack of sufficiently large-scale datasets annotated with affect labels constrained such early efforts in automatic facial feature extraction. With the success of AlexNet [Krizhevsky et al., 2012] in image classification tasks, gradually the focus shifted to creating large-scale face expression analysis datasets such as AffectNet [Mollahosseini et al., 2017], SEWA [Kossaifi et al., 2019], etc. Such large-scale datasets coupled with GPU-enabled parallel computing technology allowed facial expression recognition methods to rapidly[2] adopt advanced CNN models for spatial feature learning. Various standard CNN architectures like VGGNet [Simonyan and Zisserman, 2014], ResNet [He et al., 2016], etc that were originally proposed for image classification and object detection tasks, have been widely applied to the face expression analysis tasks. For a comprehensive survey of deep CNN models applied to the facial expression tasks, the reader is referred to [Li and Deng, 2020].

Though CNN-based facial affect recognition methods achieved impressive generalisation performance, they are severely constrained by the limited amount of affect labelled image data, unlike in the case of general perceptual tasks such as image classification and object detection, etc. As a result, when using large-scale CNNs with millions of parameters, face expression recognition models tend to suffer from the over fitting problem. Some of the early works (e.g. [Jaiswal and Valstar, 2016]) that successfully adopted CNNs for tasks such as facial action analysis, relied on shallow CNNs to learn the appearance and shape information of different facial regions. But, the facial affect analysis methods proposed later based on deep CNN models adopted different transfer learning techniques to alleviate the impact of over fitting. Several prior works (e.g. [Kaya et al., 2017] and [Ng et al., 2015]) reported the effectiveness of transfer learning in improving the face

[2]129 studies on facial expression analysis used CNNs between 2012 and 2017 [Rouast et al., 2019]

affect recognition performance. A commonly used transfer learning method is to initialise the network parameters from pre-trained models on datasets like ImageNet [Deng et al., 2009] (14 million images with the labels of 1000 classes) and VGG-Face [Parkhi et al., 2015] (2.6 million face images annotated with identity labels). Though such pre-trainined model parameters reduce the over fitting problem (e.g. improved accuracy from 39% to 42% in [Chen et al., 2016]), effectively training CNNs using small amounts of labelled data is still largely an unsolved problem in deep learning.

Vocal Affect Features

Information communicated in a speech signal can be decomposed into two major components: linguistic and paralinguistic. While the former component is concerned with the actual words spoken by the speaker, the latter describes the way those words are spoken. Besides the static factors such as identity, age, gender, etc, paralinguistic information is mainly influenced by the speaker's affective states that evolve dynamically with time. The objective of speech-based emotion recognition is to extract affect-specific paralinguistic features from a raw speech signal and map those features to categorical or dimensional emotion attributes. The efforts made towards distilling vocal affect features, similar to the facial affect feature extraction, can be grouped into the categories of engineering- and learning-based approaches, as discussed below.

Feature Engineering. The role of affective states in modulating the vocal parameters of human speech is a widely studied problem in the literature [Anagnostopoulos et al., 2015, El Ayadi et al., 2011]. Early works on speech emotion recognition have heavily relied on the knowledge of such affect-modulated vocal parameters for designing compact discriminative

feature sets. A speech signal is essentially an acoustic signal whose energy stems from vocal cord vibrations. Pitch is an important property of speech, which corresponds to the fundamental frequency of vocal cord vibrations, and prosodic features of speech signals refer to the acoustic variations in the voice pitch and intensity variables that serve linguistic functions.

Several works in speech emotion recognition consistently demonstrated that short-term prosodic, energy and spectral features of speech signals are informative about the underlying affective states. In the form of *Low-Level Descriptors (LLDs)* (e.g. [Eyben et al., 2010, Schuller et al., 2013, Eyben et al., 2015]), such speech features have been widely used as hand-engineered features for vocal affect recognition. Given a speech segment as input, LLDs are typically computed from overlapping audio *frames* with a short duration (e.g. window size set to 25 ms sliding at 10 ms rate). ComPare [Schuller et al., 2013] and eGeMAPS [Eyben et al., 2015] are two widely used standard sets of LLDs, which are composed of frequency-based parameters (pitch, jitter, etc), energy-related information (loudness and shimmer), spectral parameters and cepstral parameters such as Mel-Frequency Cepstral Coefficient values, etc. Development of open-source feature extraction tools like openEAR [Eyben et al., 2009] and openSMILE [Eyben et al., 2010], played a vital role in accelerating voice affect recognition research, by offering simplified procedures for computing the standardised LLDs. Most vocal affect recognition methods based on the aforementioned hand-engineered features assume that temporal variations in LLDs are more important for emotion recognition than the static values of per-frame LLDs. Motivated by this assumption, simple statistical functions (e.g. mean, max, variance, etc) are computed to describe the temporal variations and contours of per-frame LLDs.

Feature Learning. In recent years, the focus of voice affect feature ex-

traction approaches shifted to end-to-end representation learning, following the developments in other speech-related learning tasks like Automatic Speech Recognition. Rapid adoption of learning vocal affect features directly from raw signals, was motivated by the findings reported in [Mirsamadi et al., 2017], which demonstrated improved recognition performance through learning LLDs directly from raw spectral representations of individual audio frames. Given a raw audio waveform, the spectrogram representations are computed using either Fourier Transform or log Mel-Frequency Cepstral representations. Later, several speech emotion recognition methods relied on computing spectrogram features of multiple frames, as it allows interpreting speech segments as 2D images and using convolutional neural network architectures for feature learning. Thus, several recent works (e.g. [Huang et al., 2014, Badshah et al., 2017]) in speech representation learning adopted various shallow-variants of CNN architectures that were originally proposed for image classification and object recognition tasks.

Similarly, the idea of learning speech features directly from raw audio waveform data was first explored in 2011 in [Jaitly and Hinton, 2011]. Later, [Trigeorgis et al., 2016] demonstrated the application of 1D CNNs to speech feature learning for emotion recognition. By using a shallow-CNN with just two convolution layers for feature learning, [Trigeorgis et al., 2016] achieved significant improvements over LLDs, almost doubling the correlation between predicted and ground truth arousal labels. Following such initial developments, state-of-the-art speech emotion recognition methods completely shifted to the paradigm of learning vocal features, which requires large amounts of affect labelled audio data.

To combat the problem of limited availability of affect labelled audio data, transfer learning of speech representations received more attention in re-

cently proposed affect recognition methods. In such transfer learning methods, the parameters of DNNs are initialised with that of models pre-trained on closely related paralinguistic tasks [Gideon et al., 2017], different affect representations [Zhang et al., 2017b], and various standard datasets [Deng et al., 2013], etc. For instance, SoundNet [Aytar et al., 2016] is one of the notable examples of such pre-training strategies for effective transfer learning of speech representations for emotion recognition [Pini et al., 2017]. Despite these advancements in transfer learning of pre-trained representations, vocal affect recognition using state-of-the-art deep learning models with as few labelled data points as possible is still a major challenge in affective computing.

2.2.2 Trends in Temporal Affect Prediction: From Local to Global

Recognising affective states from face and voice data is an inherently continuous temporal phenomenon, hence, effectively modelling the temporal dynamics of low-dimensional facial and vocal affect features is crucial for reliably recognising affective states. By leveraging the availability of sequence-level context information, temporal models of affect recognition demonstrated superior generalisation performance compared to the static (frame-level) affect recognition models [Ebrahimi Kahou et al., 2015]. In temporal affect recognition the first step is to extract the spatial features for each frame (see Sec. 2.2.1) and then model the feature dynamics across the frames of a face video or a speech segment.

Several sequential data processing models have been adopted to the affect recognition tasks, ranging from classical Hidden Markov Models (HMMs) [Rabiner and Juang, 1986], Recurrent Neural Networks (RNNs) [Rumelhart

et al., 1985] to state-of-the-art self-attention [Vaswani et al., 2017] models. The efforts made so far in the temporal affect recognition literature can be broadly grouped into two categories, as discussed below:

Local Temporal Modelling refers to aggregating short-range temporal dynamics of feature representations. Given a short sequence of video or audio frames, CNN architectures are commonly used for modelling such short-range dynamics of spatio-temporal representations jointly. In the case of face image sequences, 3D CNN models [Barros et al., 2015] are widely used for modelling the temporal context directly from an input volume of face image. But, due to the large number of trainable parameters to be learned in 3D CNNs, local temporal modelling is usually limited to very short sequences (typically less than 10 frames).

In the case of speech signals, 1D CNNs [Trigeorgis et al., 2016] are typically used to hierarchically learn the spatial features and their short-range temporal dynamics. For this purpose, 1D CNNs are designed such that their low-level (i.e. first few) convolution layers capture the spatial (frame-level) characteristics of audio signals and the high-level layers model the temporal structure. In [Trigeorgis et al., 2016], the first convolution layer coupled with pooling is used for learning the spatial characteristics of raw audio signals, and the second layer is composed of convolutional kernels spanning 500 ms for modelling the temporal dynamics. Similar approaches have been explored using 2D CNN architectures for modelling the local dynamics speech representations. For instance, in [Zhang et al., 2017a], AlexNet [Krizhevsky et al., 2012] is adopted for processing the log Mel Spectral segments computed from short sequences of audio frames.

Global Temporal Modelling involves learning long-range temporal dynamics or temporal context from the feature representations of consecutive visual

or audio frames. In the case of facial or vocal affect recognition, the length of such long input sequences can go up to 10 seconds or more [Ringeval et al., 2019]. Traditionally, generative sequence models such as Hidden Markov Models (HMMs) dominated the temporal affect recognition approaches [Rabiner and Juang, 1986]. HMMs offer a highly sophisticated probabilistic model for learning temporal dependencies between per-frame affect features. However, due to their high computational complexity and their ability to learn only discrete latent states, application of HMMs to naturalistic affect recognition showed limited success. Building on the advancements in deep neural networks for sequential data processing, RNNs and self-attention [Vaswani et al., 2017] models have been widely used in recent years to model long-range temporal dynamics from the feature sequences of faces and voices. Recurrent mechanisms typically rely on gated sequential propagation of temporal context encoded into a hidden state vector. RNNs were explored in some early works for modelling the sequential dynamics of hand-engineered facial features like 2D landmark locations and optical flow. Similarly, learning the temporal context from audio LLDs coupled with RNNs was proposed in [Mirsamadi et al., 2017] for utterance-level affect recognition. In recent years, the CNN-RNN combination emerged as a prominent approach for global temporal affect modelling in both face (e.g. [Kollias and Zafeiriou, 2018b]) and voice domains (e.g. [Lim et al., 2016]).

In theory RNNs are capable of handling arbitrarily long sequences, however, in practice they are found to suffer from the vanishing gradients problem as the sequence length grows. On the other hand, attention models bypass the sequential propagation of information and directly attend to the past inputs. Thus, attention models can easily capture long-range temporal contingencies by circumventing the problem of vanishing gradients.

By controlling the information flow, gated variants of RNNs perform better than vanilla RNNs in capturing long-range dependencies. But, due to their fixed dimensional latent state to hold past information, unlike in the attention models, gated RNNs still fall short in practice in modelling long-range temporal context. For this reason, self-attention models have been recently explored in the domain of affect recognition [Wagner et al., 2011, Chandran et al., 2020, Sanchez et al., 2021]. However, this advantage with attention models comes at the cost of their poor (quadratic) scalability with the sequence length, which is not the case with RNNs. Furthermore, attention models can operate only within a fixed temporal context window whereas the RNNs can easily handle unbounded context, at least in theory. Hence, the CNN-RNN (its gated variants) combination still continues to be a popular architectural choice for global temporal modelling of affective dynamics.

Although extensive efforts have been made over the last three decades to improve the feature extraction and temporal modelling methods, automatic recognition of affective states encountered in naturalistic interactions has not been able to reach human-level parity so far. Facial and vocal affect recognition in in-the-wild operating conditions is still largely an unsolved problem. The question is: why? Towards answering this question, this thesis notes a fundamental challenge in affect recognition model training w.r.t. procuring unambiguous human supervision in large quantities, as discussed in the following section.

2.3 Challenges in Affect Labelling: A Trade-off Between Ambiguity and Scarcity

All the aforementioned supervised learning models assume the availability of reliable ground truth labels to be used as supervision signals. Here, label reliability refers to how consistently the affect labels are annotated by the human raters. Human judgements of perceived emotions can differ drastically due to the influence of various subjective and contextual factors involved in the emotion expression and perception processes [Devillers et al., 2005, Liscombe et al., 2003, Mower et al., 2009]. As a consequence, the labels provided by different human raters on the same data instances often tend to have high variance. The level of inter-rater-disagreement is particularly very high when the annotators are presented with non-prototypical and subtle emotional expressions. For instance, when annotating discrete affect labels, the neutral category is one of the most ambiguous classes for human raters [Kim and Provost, 2015]. Whereas in continuous dimensional affect annotation, the disagreement among the raters is significantly high when there is sudden and substantial rise or fall in the affect intensity [Yannakakis et al., 2018]. To give an example, in the SEMAINE corpus [McKeown et al., 2010] annotations of valence and arousal, this particular trend in the inter-rater-disagreement can be clearly observed, as illustrated in Fig. 1.2. Thus, given the inherently subjective nature of human emotion perception and its high sensitivity to different contextual factors, preparing reliable affect labels is a fundamental challenge in the field of affective computing.

To cope with the problem of inter-rater-variability, label aggregation is a commonly employed technique in most affective computing problems. With the goal of normalising subjective components in the affect ratings,

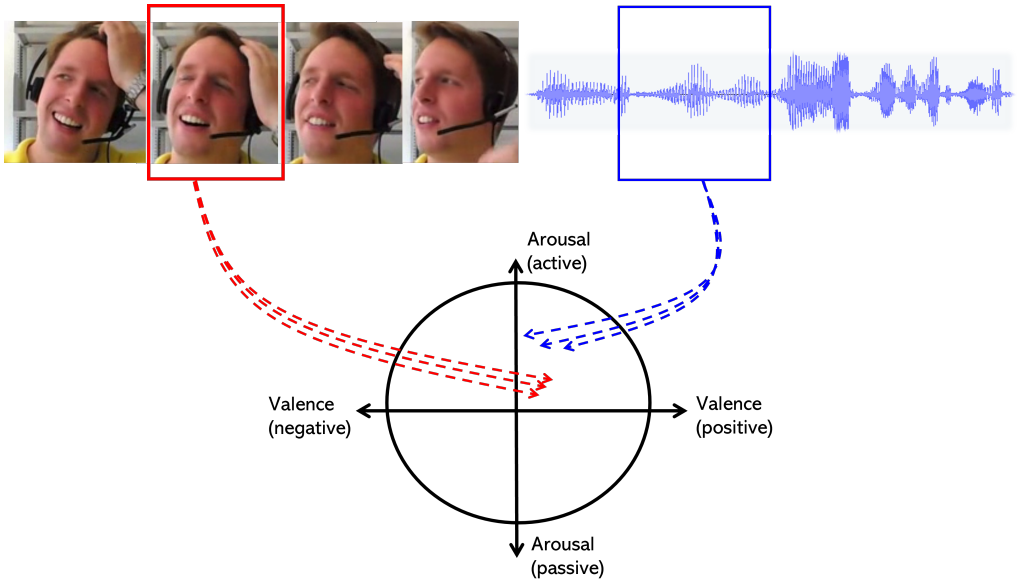


Figure 2.4: Illustration of one-to-many function mappings in facial and vocal affect recognition

the label preparation step aims to maximise the inter-rater agreement score, hence the label consistency. To this end, first a pool of human raters are employed to label the same set of data instances. Then, by applying the simple averaging or majority voting to multiple annotations, the labels provided by all the raters are aggregated into a final ground truth label. More sophisticated aggregation techniques such as Estimator Weighted Evaluation (EWE) [Grimm and Kroschel, 2005] have been adopted for maximising the inter-rater-agreement scores. Regardless of the size and composition of the pool of raters and the aggregation technique applied, it is impractical to completely get rid of the ambiguity induced by the systematic disagreements among the raters.

A key assumption implicitly made in the existing affect label aggregation techniques is that inter-rater variability is simply a noise component, and minimising this does not cause any information loss. But, it is worth noting that clearly explainable systematic disagreements among the raters may also induce variability, which may reflect the influence of dependencies

such as their personality, mood, socio-demogrpahic factors, etc on their interpretations of perceived affect. Such important nuances play a vital role in enriching emotional expressions in natural human interactions. By treating the variability induced by the key contextual factors as mere noise, the existing label aggregation methods are likely to fail in capturing the richness of affect labels provided by individual raters.

By contesting the commonly held view of treating 'variability-as-noise', several recent works proposed to embrace the inter-rater-disagreement as a learning signal. Instead of aggregating multiple annotations of affect into a hard label, methods proposed in [Han et al., 2017, 2020] explored the use of soft labels in systematically capturing the inter-rater-disagreements. Another important approach is based on ensemble learning of affect [Fayek et al., 2016, Fornaciari et al., 2021], in which different models are trained using the emotion labels provided by individual raters. All these methods demonstrated improved generalisation performance by using inter-rater-disagreement as an additional supervision signal for systematically modelling the label ambiguity. However, the application of soft-labelling and ensemble learning models has been mainly confined to small-scale corpora, and scaling them up to large-scale in-the-wild datasets is severely constrained by various factors such as the size and composition of the pool of human raters, high computational complexity of deep ensemble models etc.

Trade-off Between Label Ambiguity and Label Scarcity: Assuming a fixed number of annotator-hours, the objective of making the affect labels less ambiguous by recruiting a large pool of raters leads to scarcely annotated datasets, as observed in the case of SEWA [Kossaifi et al., 2019]. Similarly, the objective of annotating large amounts of data using the same number of annotator-hours can lead to more ambiguous affect labels. Thus,

Figure 2.5: Visual perception in the absence of contextual information (Image source:[Gregory, 2005])

the trade off between the label ambiguity and label scarcity is one of the most important bottlenecks thwarting the progress of affect recognition systems. This thesis aims to relax this trade off by developing ML models that separately attempt to cope with the label ambiguity and label scarcity problems, with the number of annotator-hours unaltered.

From the above discussion, it is clear that the process of manually annotating affect labels is inherently ambiguous. To deal with the label ambiguity problem, one natural approach is to identify the easily accessible contextual cues that can help in partly resolving the ambiguity, considering that emotional expressions do not occur in isolation [Kosti et al., 2017]. Fundamentally, perception of an auditory or visual stimulus involves deriving meaningful abstract representations or features from raw acoustic or pixel data. In the absence of any contextual information, auditory or visual perception tasks can be highly ambiguous. For example, describing the visual stimulus presented in Fig. 2.5, without knowing any contextual information, demands more cognitive effort to resolve the ambiguity. On the other hand, we can effortlessly process any of this visual input if it is accompanied with enough contextual information. The contextual signals could

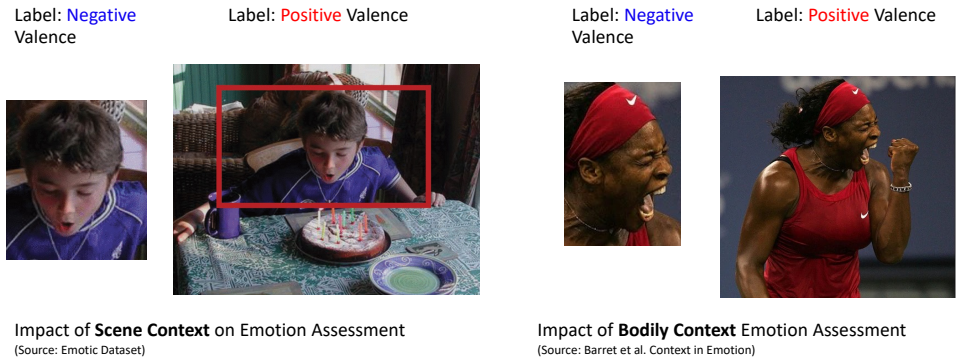


Figure 2.6: Resolving ambiguity in emotion perception using context in (a). scene-level cues (Image source: [Kosti et al., 2017]) and (b). bodily cues (Image source: [Barrett et al., 2011])

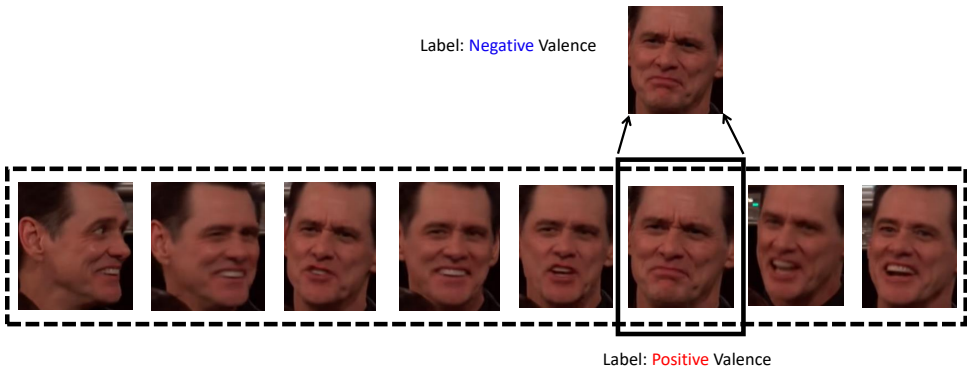


Figure 2.7: Using temporal contextual cues to resolve the affect labelling ambiguities

be presented either through the same modality (like a sequence of images) or some other modality (text, audio, etc.). To give an example, revisiting Fig. 2.5 after reading its caption, 'A dog sniffing around the base of a tree', immediately resolves the perception uncertainty encountered earlier. This simple example clearly demonstrates the importance of the contextual cues in minimising the label ambiguity.

Contextual cues that play a crucial role in human affect perception [Barrett et al., 2011] can be broadly grouped into two categories, according their accessibility:

a. Type I: Characteristics of the input stimulus like the temporal evolu-

tion of emotions expressed in a face video or speech signal, multi-modal correspondences between audio and visual modalities, body gesture signals embedded in the 2D or 3D pose data, and interactions with other people and objects in a given scene, etc.

b. Type II: Characteristics of the perceiver's state like expressivity-related latent factors (personality traits, mood, etc) that vary from individual to individual, biases induced by various socio-demographic factors and the degree of cultural familiarity, etc.

Compared to the Type II contextual cues, Type I cues such as scene-level and bodily context (see Fig. 2.6) are more easily accessible for in-the-wild emotion recognition tasks. As this thesis mainly focuses on temporal affect recognition from face and voice data, among the Type I cues, temporal evolution patterns or temporal context (see Fig. 2.7) is used the main contextual cue for partly resolving the ambiguity, as temporal dynamics are more readily accessible compared to the remaining cues listed above. Thus, by effectively leveraging the temporal contextual cues, this thesis aims to address the following problems:

1. The label ambiguity problem by proposing uncertainty-aware temporal regression models that can naturally handle ambiguous supervision signals (ground truth labels of affect) through probabilistic temporal modelling.

2. The label scarcity problem by proposing label-efficient representation learning models for face and voice data, which exploit the freely available supervision signals for self-supervised pre-training.

2.4 Dealing with Label-Ambiguity: Deterministic vs. Non-deterministic Temporal Context Learning

By definition, the problem of label ambiguity refers to the condition in which an input is likely to have multiple output states, as illustrated in Fig. 2.4. Such input-output mappings are not commonly encountered when preparing the ground truth labels for general perceptual learning tasks such as image classification and object detection, etc. Most existing affect recognition models, ignoring this fundamental difference, directly adopt the ML models that were originally developed for tasks dealing with unambiguous input-output correspondences. It is important to note that such models implicitly assume a deterministic learning function between the inputs and outputs. As a result, label-ambiguity-unaware ML models are bound to show poor generalisation performance when applied to affect recognition tasks which need to deal with ambiguous supervision. This raises a fundamental question - how to choose an ML model that can account for the label ambiguity without an explicit supervision signal in the form of inter-rater-disagreement scores, etc. Answering this question, this thesis proposes to learn affect recognition by using ML models that allow **non-deterministic** function learning.

Supervised ML models, based on either parametric or non-parametric formulations, can be broadly characterised as deterministic and non-deterministic function learning approaches. In supervised learning, given a set of dependent variables (X) and their corresponding independent variables (Y), training an ML model boils down to learning an approximate functional map f between X and Y i.e. $f : X \rightarrow Y$. The goal here is to recover an

approximate function f that is as close as possible to the true underlying function f^*. In a deterministic model, a single function is learned from the whole training dataset, and at test time the predictions made by it are point estimates of the targets (Y). From this function modelling viewpoint, the problem of label ambiguity implies the existence of one-to-many functional mapping between an independent variable (X) and multiple dependent variables (Y). Thus, when the true underlying function f^* between X and Y tends to have high variability induced by the annotation ambiguities, as in the case of affect recognition, a deterministic function map is bound to have poor predictive performance. State-of-the-art DNNs applied to face and voice recognition tasks, despite their over-parameterised implementations, can be grouped into the deterministic function learning category, which explains their poor generalisation performance when the supervision signals are ambiguous.

On the other hand, non-deterministic function learning models assume that the functional mapping between X and Y can be modelled better using a probability distribution over the function space ($P(f)$) or a function ensemble ($\{f_i\}$). As a result, at test time, the predictions made by non-deterministic models are probability distributions over the target values, instead of the fixed point estimates. Thus, non-deterministic function learning approaches, at least in theory, possess the natural ability to cope with the ambiguous supervision signals, unlike the deterministic function learning models.

This thesis argues that adopting non-deterministic function learning is critical to make the affect recognition models capable of coping with the annotation ambiguities. In particular, this work explores a class of non-deterministic models based on learning a distribution over function space $P(f)$, in which the most probable function corresponds to the target Y

value with highest likelihood value. When applied to the tasks of affect recognition from a face video or speech signal, this probabilistic functional viewpoint translates to learning a distribution of temporal functions, rather than learning a single temporal function. In the ML literature, learning distributions of temporal functions from sequential data using probabilistic modelling is a widely studied topic. Some of the most notable probabilistic temporal function learning algorithms are discussed below.

- **Hidden Markov Models (HMMs)** [Rabiner and Juang, 1986] are based on the idea of augmenting the Markov chain, which were widely used for probabilistic modelling of sequential data. Given sequences of random variables, also called as *states* $\{s_i\}$, a Markov chain computes the state probabilities assuming that predicting a next state in the sequence depends solely on the current state i.e. $P(s_i|s_1....s_{i-1}) = P(s_i|s_{i-1})$. Based on this assumption, an HMM is designed to model the probabilities of not only the observable variables in a sequence but also the *hidden* variables that are assumed to be the causal factors underlying the observable data. Thus, HMMs model the time-series data as a doubly stochastic process by learning a directed probabilistic graphical model. Due to their early success in temporal modelling of speech signals in tasks such as speech recognition [Gales, 1998], HMMs were explored in temporal affect recognition tasks (e.g. [Li et al., 2013]). In spite of their ability to probabilistically model the temporal context, HMMs failed to compete with other temporal models such RNNs, mainly due to their strong Markovian assumption, computationally expensive learning algorithms, and the discrete hidden states which limit their representation capacity.

- **Continuous Conditional Random Fields (CCRFs)** [Qin et al., 2008] extend the classical relational learning model, Conditional Ran-

dom Field (CRF) [Sutton et al., 2012], to capture the temporal relationships of random variables in a sequence. Unlike an HMM which learns a directed graphical model of probabilities of sequences, a CRF is an undirected graphical model of conditional probability distributions of random variables. Thus, CRFs can be arbitrarily structured, unlike the linear-sequence structuring constraint in HMMs. By generalising CRFs to continuous variables, CCRFs were applied to the time-continuous affect recognition tasks (e.g. [Baltrušaitis et al., 2013]). Although CCRFs offer a rich representation framework for capturing the ambiguity in temporal context, as in affect recognition tasks, high computational complexity of training these models severely constrained their application to large-scale emotion recognition datasets.

- **Bayesian Recurrent Neural Networks (BRNNs)** [Fortunato et al., 2017] augment the standard RNN architectures with uncertainty modelling abilities by learning distributions over the weights, rather than learning point estimates. By combining the strengths of RNNs in modelling non-linear temporal dynamics with the advantages of Bayesian learning approaches, BRNNs offer a principled framework for learning non-deterministic temporal functions, however, their application to large-scale models of temporal learning is very limited. Although some affect recognition methods attempted to combine the RNNs with Bayesian filters such as Kalman filters [Pei et al., 2022], direct application of BRNNs to temporal affect modelling is still an under explored solution, mainly due to more complicated training procedures and slower convergence rates of Bayesian Neural Networks, when compared with the standard RNNs with point-valued weights.

- **Gaussian Processes (GPs)** [Rasmussen, 2003] family is a popular choice for non-parametric Bayesian modelling of time-series data. Fundamentally, a Gaussian *process* can be viewed as the generalised formulation of a Gaussian *distribution*. A probability distribution is for describing the scalar or vector-valued random **variables**, whereas a stochastic *process* describes the principles governing the properties of random **functions**. To exploit the richness of GPs w.r.t. capturing the ambiguity or uncertainty in temporal function learning in a principled fashion, some works explored the application of GPs to continuous affect recognition tasks [Atcheson et al., 2017]. Despite their flexibility, data-efficiency and probabilistic nature, the potential of GPs has not been fully exploited for affect recognition. This is due to two key limitations of the existing GP implementations that constrain their applicability to large-scale temporal modelling problems: poor inference scalability (cubical complexity) to high-dimensional feature spaces and the requirement of hand-designed covariance function (kernel function) based on the domain knowledge of problem at hand.

Noting the limitations of above discussed non-deterministic temporal modelling approaches, this thesis aims to develop novel scalable probabilistic temporal models for face and speech emotion recognition tasks, with two main objectives:

1. To improve the generalisation performance of in-the-wild naturalistic affect recognition by effectively fusing the audiovisual affect information

2. To capture temporal predictive uncertainty associated with the estimated affective states for the benefit of downstream behavioural

analysis tasks

2.5 Dealing with Label-Scarcity: Direct supervision vs. Self-supervision for Representation Learning

State-of-the-art affect recognition models, as discussed in Sec. 2.2, heavily rely on large amounts of labelled data for learning facial and vocal features. Affect annotation is a highly time-consuming and expensive process. Due to the label ambiguity problem in affect recognition tasks, collecting annotations from multiple trained human raters that have a socio-cultural understanding of the dataset context is essential. Most existing affect labelled datasets are prepared by collecting labels from at least five to six human raters [Kossaifi et al., 2019, Kollias et al., 2020, McKeown et al., 2010]. As a consequence, preparing large-scale labelled datasets for emotion recognition is a challenge, leading to small or scarcely labelled corpora. To give an example, SEWA [Kossaifi et al., 2019], one of the largest in-the-wild datasets annotated with time-continuous affect labels, has approximately 33 hours of raw audio-visual recordings. But, only 14 % of the total data in SEWA could be labelled with affect annotations, which indicates the prohibitively expensive nature of affect labelling process. Compared to the other large-scale video datasets such as the YouTube-8M [Abu-El-Haija et al., 2016] that has over 350,000 hours of audio-visual data fully annotated with ground truth labels, affect recognition datasets such as SEWA are considerably small for end-to-end video representation learning. Hence, relying solely on directly or fully supervised approaches for learning facial and vocal representations severely limits the generalisation performance

of existing affect recognition methods. Towards promoting label-efficient feature learning for affect recognition, this thesis aims to develop an alternative approach that can perform well using fewer human annotations for model training.

In the absence of large amounts of labelled data, to minimise the model over fitting, it is a common practice to use pre-trained models such as ImageNet [Deng et al., 2009] or VGGFace [Parkhi et al., 2015] to initialise the weights of the CNN feature encoders, as discussed in Sec. 2.2.1. Since the CNNs are designed to learn representations in a hierarchical fashion, it is reasonable to presume that the early layers' features tend to be task-agnostic. Hence, the first few layers of the pre-trained models may transfer well to the task at hand in spite of the fact that the pre-trained models are often based on completely different datasets and tasks. As a result, initializing the feature encoders with pre-trained model weights helps in improving the performance. However, such pre-training techniques do not leverage abundantly available unlabelled audio-visual data. To this end, self-supervised representation learning for model pre-training has emerged as a promising alternative in recent years, and it demonstrated great potential in facilitating label-efficient representation learning [Hénaff et al., 2019].

Self-Supervised Representation Learning relies on natural supervision signals that are embedded in unlabelled data in the form of data point correspondences [Jing and Tian, 2020]. These natural supervision cues are exploited to define a surrogate or proxy learning task, which is designed in such a way that its target labels can be automatically generated from the structure of the unlabelled data. By training a model to learn the proxy task, several methods demonstrated the possibility of learning the underlying semantic representations embedded in the unlabelled data without

using any manual annotations.

Motivated by the success of self-supervised learning models such as BERT [Devlin et al., 2018] in natural language processing, a wide range of proxy tasks [Jing and Tian, 2020] have been developed in recent years for both visual and audio modalities as well. For example, colorization [Larsson et al., 2017] is a well-studied visual proxy task in which the model is trained to predict the color values of the image pixels in the corresponding gray scale image input. Since color information is strongly correlated with the image semantics, this proxy task indirectly encourages the model to learn general-purpose semantic features without using any labelled data. Similarly, Wav2Vec [Schneider et al., 2019] is a popular proxy task designed for audio representation learning, in which the model is tasked with predicting the raw waveform data of unseen segments based on the context gleaned from the already seen audio segments.

In recent years, building on the advancements in self-supervised visual and audio feature learning tasks, several attempts have been made to extend them to facial and vocal features for affect recognition. For instance, given an unlabelled video corpus, to learn discriminative features for facial action unit (AU) analysis, [Li et al., 2020a] proposed a Twin-cycle Auto Encoder (TAE). TAE is trained in a self-supervised manner with the goal of disentangling pose-induced and action-induced facial movements. Similarly, in [Lu et al., 2020], temporal consistency is used as a natural supervision signal for facial feature learning, and it demonstrated how to use the natural ordering of frames in a face video for defining frame-ranking as a proxy task. The self-supervised facial features learned in this approach achieved impressive performance on facial action unit detection tasks. Unlike these tasks that heavily rely on the temporal dynamics of face videos, [Chang et al., 2021] designed a proxy task for learning features from static face

image data, based on the idea of cycle-consistency, which is used as a constraint in disentangling the facial identity and expression features. In contrast to all the aforementioned works, Emotion-aware Contrastive Learning (EmoCo) [Sun et al., 2021] leveraged the discrete expression labels coupled with Contrastive Learning for AU-related feature learning. This method exploits the fact that procuring the labels of six basic emotions is considerably easier than labelling the AU intensity values.

Self-supervised speech representations are also widely explored using emotion recognition as the downstream task. Most notably, [Shukla et al., 2021] proposed audio-guided face reconstruction as the proxy task to learn speech features for emotion recognition. They also proposed the audio-version of a visual proxy task called Odd-One-Out in which the model is trained for the task of temporal order verification. Affect recognition models based on such self-supervised speech features demonstrated better generalisation performance than the fully supervised models. To comprehensively establish the performance gains of self-supervised pre-training in speech emotion recognition, recently, [Wagner et al., 2022] presented a thorough analysis of some standard self-supervised transformer models. This study delineated the impact of self-supervised pre-training based on methods such as wav2vec 2.0 [Baevski et al., 2020] and HuBERT [Hsu et al., 2021] on speech affect recognition performance. The findings presented in it confirmed that pre-trained speech features learned through self-supervision by leveraging linguistic information, can achieve significantly better results on valence recognition, which is a challenging dimension to infer from speech data alone using fully supervised learning models.

Despite the potential of self-supervised pre-training in facilitating label-efficient learning, its adoption into affective computing is relatively very limited. Furthermore, it is important to note that most existing works

directly adopt the proxy tasks that were originally proposed for general purpose visual or audio representation learning. Considering that affective states are typically weak and noisy signals to capture from high-dimensional in-the-wild recordings, this work argues that to effectively learn affect-related features from face and speech data, the proxy task must be made aware of the downstream task's requirements. Furthermore, this thesis presents a thorough analysis of label-efficiency advantages promised by self-supervised pre-training, towards addressing the label scarcity problem of affect recognition tasks.

To address the above mentioned limitations of existing self-supervised learning approaches applied to affect recognition from face and speech data, this thesis focuses on

1. Learning a novel proxy task that is informed by the properties of a specific downstream task of interest i.e. continuous dimensional emotion recognition, through its hand-engineered feature representations.

2. Comprehensively evaluating the label-efficiency benefits of the proposed self-supervised learning model, in comparison with the state-of-the-art fully supervised learning models.

Chapter 3

COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition

> As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.
>
> *Albert Einstein*

Chapter Summary. Humans rely primarily on visual (faces) and audio (voices) modalities to encode and express their affective or emotional states. Automatically recognising apparent emotions from face and voice

is hard, in part because of various sources of uncertainty, including in the input data and the labels used in a machine learning framework. This chapter introduces an uncertainty-aware audiovisual fusion approach that quantifies modality-wise uncertainty towards emotion prediction. To this end, this thesis proposes a novel fusion framework in which first latent distributions are learned over audio and visual temporal context vectors separately, and then the variance values of unimodal latent distributions are constrained such that they represent the amount of information each modality holds w.r.t. emotion recognition. In particular, the proposed approach imposes **Calibration and Ordinal Ranking** constraints on the temporal context variance vectors of audio and visual latent distributions. When *well-calibrated*, modality-wise uncertainty scores indicate how much their corresponding predictions may differ from the ground truth labels. *Well-ranked* uncertainty scores allow ordinal ranking of different frames across the modalities. To jointly impose both these constraints on the audio and visual latent distributions, this thesis proposes a softmax distributional matching loss. In both classification and regression settings, this chapter compares the proposed uncertainty-aware fusion model with standard feature and prediction fusion models, as well as a temporal context fusion baseline. The experimental evaluation on a spontaneous emotion recognition corpus, AVEC 2019 Cross-cultural Emotion Subchallenge (CES), shows that multimodal emotion recognition can considerably benefit from well-calibrated and well-ranked latent uncertainty measures.

3.1 Introduction

Learning to fuse task-specific information from multiple modalities is a fundamental problem in Machine Learning. At its core, this problem entails

estimating how informative each modality is towards predicting the labels of a target task. Thus, uncertainty-aware information fusion is a natural approach to multimodal learning. This chapter formulates an uncertainty-aware fusion method for an inherently multimodal task – apparent emotion recognition from audiovisual signals (faces and voices). It further proposes a multimodal fusion framework based on probabilistic modelling of unimodal temporal context.

Being an intrinsically temporal and multimodal phenomenon, continuous emotion (valence and arousal) recognition from face videos and speech signals is one of the long-standing challenges in Affective Computing [Schuller et al., 2012, Valstar et al., 2013, Ringeval et al., 2019]. A meta-analysis presented in [D'mello and Kory, 2015] has shown that although emotion recognition can benefit from multimodal fusion in general, performance improvements are not significant when it comes to spontaneous emotions. This chapter argues that uncertainty-aware multimodal fusion may have the potential to address this challenge, considering that the intensity of spontaneous emotions embedded in the facial and vocal expressions are likely to vary dynamically over time [Nicolaou et al., 2011, Zeng et al., 2008].

Although Deep Neural Networks (DNNs) have been extensively applied to audiovisual emotion recognition [Rouast et al., 2019, Noroozi et al., 2017, Schoneveld et al., 2021, Gerczuk et al., 2021], estimating modality-wise uncertainty for improved fusion performance is a relatively under-explored avenue. However, modelling predictive uncertainty (or confidence, its opposite) in DNNs received widespread attention in recent years [Guo et al., 2017, Mukhoti et al., 2020], motivated by the observation that DNNs tend to make over-confident predictions [Nguyen et al., 2015, Szegedy et al., 2013b]. Most existing efforts towards uncertainty or confidence estimation

in DNNs [Guo et al., 2017, Kumar et al., 2018b] focus solely on reducing miscalibration errors, i.e., the mismatch between expected model estimation errors and their corresponding confidence scores. Recently, as an alternative perspective, [Moon et al., 2020] introduced the idea of learning to rank confidence scores for identifying the most reliable predictions.

The objective of the proposed method in this chapter is to estimate the uncertainty scores of unimodal inputs to maximise the multimodal fusion performance. This chapter argues that the predictive uncertainty of an estimator must be simultaneously both *well-calibrated* and *well-ranked (ordinal)*. The former is needed to accurately represent the correctness likelihood of a prediction for **an individual sample**. The latter is essential to effectively order predictions for **a group of samples** according to their correctness likelihoods. In other words, if an uncertainty estimate of an individual sample is well-calibrated, in the absence of its ground truth, the uncertainty score can serve as a proxy for its expected prediction error. If the uncertainty scores associated with different predictions are well-ranked or maintain ordinality, then one can use them to order their corresponding samples in terms of their reliability towards the target prediction, and to distinguish the most informative samples from the least informative samples.

For multimodal temporal learning, it is critical to estimate how informative the predictions made for different frames in different unimodal sequences are, towards estimating a common target label, so that the target-specific information can be reliably integrated [Yang et al., 2017]. This chapter hypothesises that jointly learning these two properties – calibration and ordinality – may lead to more reliable per-frame predictive uncertainty estimates for each modality, facilitating more effective uncertainty-weighted temporal context fusion. Based on this hypothesis, this chapter proposes an

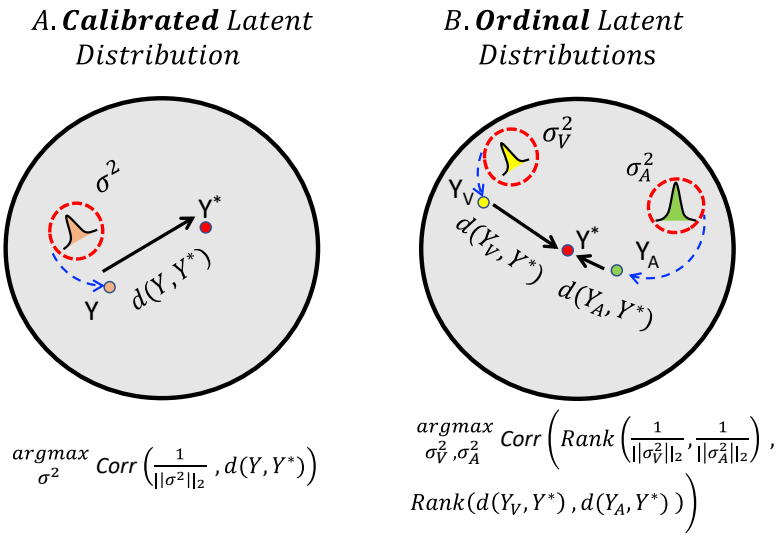


Figure 3.1: Illustration of the proposed latent distribution modelling for multimodal fusion (Y_V and Y_A – unimodal predictions, Y^* – target label, and d – a distance function): ***A. Calibrated Latent Distribution:*** For a given modality, its temporal context is modelled by a latent distribution that is learned under the *calibration constraint* i.e. $\underset{\sigma^2}{\text{argmax}}\ Corr(\frac{1}{\|\sigma^2\|_2}, d(Y, Y^*))$. Thus, the variance σ^2 is learned to represent how informative the temporal context is w.r.t the target label prediction. ***B. Ordinal Latent Distributions:*** The variance values of audio and visual temporal context distributions (σ_V^2 and σ_A^2) are learned under the ordinal ranking constraint i.e. $\underset{\sigma_V^2, \sigma_A^2}{\text{argmax}}\ Corr(Rank(\frac{1}{\|\sigma_V^2\|_2}, \frac{1}{\|\sigma_A^2\|_2}), Rank(d(Y_V, Y^*), d(Y_A, Y^*))$. Thus, the audio and visual modalities are ranked based on how informative they are towards the target prediction.

uncertainty modelling method that imposes the calibration and ordinality constraints jointly, as Figure 3.1 illustrates. The proposed method conditions the unimodal latent distributions' context variance vectors such that they represent how informative different modalities are w.r.t. predicting the target labels. This approach can be viewed as an uncertainty-aware extension of classical late fusion. It proposes to learn uncertainty estimates in terms of higher dimensional (more informative) latent distributions, unlike simple confidence-weighted late fusion which directly models uncertainty over lower-dimensional (less informative) unimodal predictions. This formulation is based on the assumption that modelling uncertainty in the

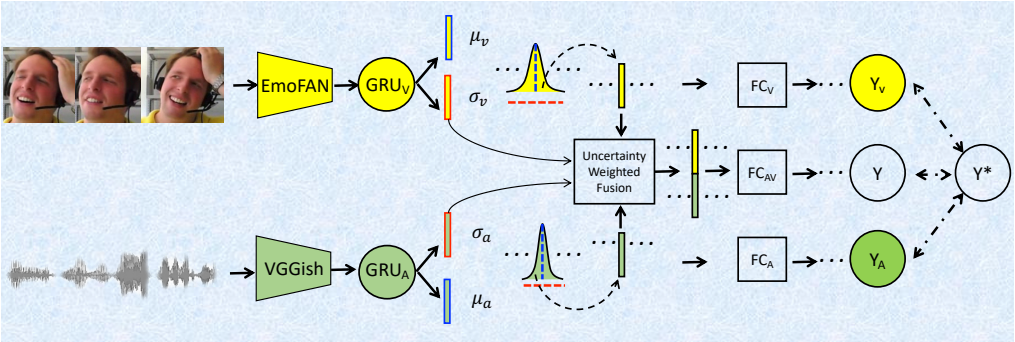


Figure 3.2: Overview of the proposed approach to an uncertainty-aware audiovisual fusion for emotion recognition: Modelling latent distributions over unimodal temporal context vectors to derive modality-wise uncertainty guided fusion weights. A detailed description of the proposed approach is given in Section 3.4.

higher dimensional latent space may be more effective than in the lower dimensional label space.

The proposed framework, denoted as Calibrated Ordinal Latent Distributions (COLD), first learns the latent distributions (multivariate normal distributions) over the temporal context of audio and visual modalities separately, as Figure 3.2 shows. It models the temporal context variance norm values of the audio and visual latent distributions, $\|\sigma_V^2\|_2$ and $\|\sigma_A^2\|_2$, as the confidence measures towards emotion prediction. A novel training objective is designed based on softmax distributional matching to encourage the frame-wise temporal context variance values in each modality to be: (a) directly correlated with the correctness likelihood of the unimodal predictions, and (b) ordinal in nature to effectively rank the frames of both the modalities towards the target prediction. Thus, the calibrated and ordinal unimodal variance scores are learnt for effective uncertainty-weighted fusion, as shown in Figure 3.2.

This chapter evaluates the proposed COLD fusion on an in-the-wild audiovisual corpus, AVEC 2019 CES [Ringeval et al., 2019], for recognising spontaneous emotions in naturalistic interactions. The experimental results

show that the COLD fusion outperforms the standard model-agnostic fusion baselines by considerable margin, with ∼6% average relative improvement in terms of mean correlation score over the best performing fusion baseline trained for emotion regression. Furthermore, this chapter evaluates the robustness of different fusion models at test time by inducing noise into the visual modality through face masking. With the faces masked in 50% of the evaluation sequences, the COLD fusion achieves ∼17% average relative improvement over the best fusion baseline.

The key contributions of this chapter are as follows:

- This chapter proposes an uncertainty-aware multimodal fusion method that dynamically estimates how informative unimodal inputs are w.r.t. the target label prediction.
- The proposed method demonstrates how to jointly learn *well-calibrated* and *well-ranked* unimodal uncertainty measures. For this purpose, it proposes a simple softmax distributional matching loss function that applies to both regression and classification models.
- On an in-the-wild audiovisual emotion recognition database, the proposed uncertainty-aware fusion model outperforms the standard model-agnostic fusion baselines as well as a multimodal transformer baseline.

3.2 Related Work

Audiovisual Dimensional Affect Recognition. Recognising dimensional emotions, valence (how pleasant an emotion is) and arousal (how active an emotion is), from audiovisual modalities is a widely studied problem in various prior works [Zeng et al., 2008, Gunes et al., 2011], ranging from

the almost a decade-long running annual AVEC challenge series [Schuller et al., 2012, Valstar et al., 2013, Ringeval et al., 2019] to the recently introduced MuSe challenge [Stappen et al., 2020, 2021, Christ et al., 2022] and ABAW challenge [Kollias et al., 2020, Kollias and Zafeiriou, 2021]. The reader is referred to [Poria et al., 2017] and [Rouast et al., 2019] for comprehensive surveys of affect recognition in multimodal settings and contemporary deep learning-specific advancements in it. Since the main focus in this chapter is on uncertainty-aware fusion models for emotion recognition, it reviews the literature closely related to the following key research topics: i) uncertainty modelling for emotion and expression recognition, ii) uncertainty-aware multimodal fusion, iii) calibrated uncertainty, and iv) ranking-based uncertainty.

Uncertainty Modelling for Emotion and Expression Recognition. In discrete facial expression recognition tasks, modelling predictive uncertainty is studied in several recent works [She et al., 2021, Zhang et al., 2021, Wang et al., 2020], by estimating uncertainty in the space of low-dimensional feature embedding outputs from a Convolutional Neural Network (CNN) backbone. On the other hand, directly predicting emotion label uncertainty is explored in [Foteinopoulou et al., 2021], but only in unimodal (video-only) settings. For uncertainty-aware multimodal emotion recognition, some prior works applied Kernel Entropy Component Analysis (KECA) [Zeng et al., 2005] and Multi Modal-Hidden Markov Models (MM-HMMs) [Xie and Guan, 2013] by predicting modality-specific uncertainty measures for estimating the fusion weights. Noting the limitations of deterministic function learning in DNNs for uncertainty modelling, [Dang et al., 2017] explored the application of Gaussian Process (GP) Regression to the fusion of emotion predictions.

All the aforementioned methods demonstrated the potential of uncertainty-

aware emotion recognition models over their uncertainty-unaware counterparts in general. However, they ignore two important aspects in uncertainty modelling: calibration and ordinality (ranking). This chapter aims to demonstrate the significance of these two properties by hypothesising that learning well-calibrated and well-ranked uncertainty estimates is critical for improving the audiovisual emotion recognition performance.

Uncertainty-Aware Multimodal Fusion. In general, for multimodal sensor fusion, several prior works [Zeng et al., 2005, Schörgendorfer and Elmenreich, 2006, Große et al., 2008, Papandreou et al., 2009] explored uncertainty-aware or confidence-weighted averaging techniques for classic machine learning models before the advent of Deep Neural Networks (DNNs). Recently, [Subedar et al., 2019] applied Bayesian DNNs for uncertainty-aware audiovisual fusion to improve human activity recognition performance. Similarly, [Tian et al., 2020] explored the use of uncertainty estimation in fusing the softmax scores predicted using CNNs for semantic segmentation. Although these approaches demonstrated critical advantages over the models that predict only point estimates, they do not study the calibration properties of the estimated uncertainty scores. Further, such DNN models focus mainly on modelling absolute uncertainty estimates, whereas the focus of this chapter is on jointly **learning the calibrated and relational uncertainty estimates** in an end-to-end fashion introducing a novel loss function based on softmax distributional matching.

Calibrated Uncertainty. As DNNs tend to make overconfident predictions [Nguyen et al., 2015, Szegedy et al., 2013b], confidence calibration has received significant attention in recent years [Nguyen et al., 2015, Szegedy et al., 2013b]. Calibrating confidence or uncertainty estimates involves maximising the correlation between predictive accuracy values and predictive uncertainty scores. A wide variety of calibration techniques, particu-

larly in classification settings, can be broadly categorised into explicit and implicit calibration categories [Wang et al., 2021]. In the former category, two types of post-hoc methods, binning-based and temperature-scaling, are applied to increase the reliability of DNN confidence estimates [Guo et al., 2017, Minderer et al., 2021]. In binning-based methods such as non-parametric histogram binning [Zadrozny and Elkan, 2001], calibrated confidence is estimated based on the average count of positive-class instances in each bin. This method is extended to jointly optimise the bin boundaries and their predictions in Isotonic Regression [Zadrozny and Elkan, 2002]. Temperature-scaling methods can be viewed as generalised versions of Platt scaling [Platt et al., 1999] using logistic regression for calibrating the class probabilities. Here, we use temperature-scaling as a calibration baseline [Hinton et al., 2015, Guo et al., 2017] to compare against the uncertainty calibration performance of the proposed method, due to its simplicity.

Implicit calibration methods mainly focus on tailoring the training objective of DNNs to minimise the prediction error and calibration error simultaneously. Addressing the limitations of standard cross-entropy loss w.r.t. confidence calibration, various alternative loss functions such as focal loss [Mukhoti et al., 2020], maximum mean calibration error [Kumar et al., 2018b], and accuracy vs uncertainty calibration [Krishnan and Tickoo, 2020], have been investigated recently. Calibrating regression models is relatively under-explored compared to the classification. Some recent works [Kuleshov et al., 2018, Song et al., 2019, Utpala and Rai, 2020] made attempts to extend some of the aforementioned calibration techniques to continuous-valued predictions.

Ordinal or Ranking-based Uncertainty. In the existing uncertainty modelling works, the ordinal property of uncertainty estimates received less

attention compared to the calibration property, which partly motivated the method introduced in this chapter. [Li et al., 2021] proposed to model data uncertainty by inducing ordinality into probabilistic embeddings of face images. Towards uncertainty-aware regression problems, the results reported in [Li et al., 2021] highlighted the key limitations of deterministic unordered embeddings compared to the probabilistic ordinal embeddings. Although not strictly ordinal, relative uncertainty modelling is explored for facial expression recognition in [Zhang et al., 2021].

Other closely related works approached the problem of ordinal ranking of uncertainty estimates with different objectives such as failure prediction [Corbière et al., 2019], out-of-distribution detection [Roady et al., 2019], and selective classification [Geifman and El-Yaniv, 2017]. Fundamentally, all these objectives necessitate a method that can train the model to output well-ranked confidence or uncertainty scores. Among these existing methods, the one most closely related to ours is by [Moon et al., 2020], which proposes a Correctness Ranking Loss (CRL). CRL directly imposes ordinal ranking constraints on the confidence estimates of a DNN classifier. Similar to CRL, our proposed softmax distributional matching loss also constrains the ordinal-ranking property of uncertainty estimates. However, in addition to ordinal ranking, the proposed method imposes the calibration property as well, most importantly by controlling the latent distribution variance, unlike in CRL. Moreover, its formulation generalises the idea of calibrated and ordinal uncertainty estimates to both classification and regression settings, using a common loss function computation.

3.3 Method

Preliminaries and Notations. As Fig. 3.2 illustrates, given a face video clip X_V with N frames and its corresponding speech signal X_A, using overlapping time windows, N speech segments that correspond to the N visual frames are extracted first. This method assumes that both the signals X_V and X_A are annotated with a common dimensional emotion label, $Y^* = [Y^*_{valence}, Y^*_{arousal}]$ (either per-frame or per-sequence). It extracts sequences of per-frame low dimensional features (Z_V, Z_A) from the face video and speech inputs using a two-stream network. This network is composed of a 2D CNN f_V and a 1D CNN f_A for processing the face images and speech segments respectively, $f_V : X_V \rightarrow [z^1_V, z^2_V, ..., z^N_V]$ and $f_A : X_A \rightarrow [z^1_A, z^2_A, ..., z^N_A]$. For unimodal emotion recognition, the temporal context from each modality is processed separately from Z_V and Z_A using different temporal networks $g_V : Z_V \rightarrow Y_V$ and $g_A : Z_A \rightarrow Y_A$ to predict the emotion labels Y_V and Y_A.

Figure 3.3 illustrates the proposed solution to uncertainty-aware multimodal fusion. This section first discusses how to estimate modality-wise uncertainty by learning unimodal latent distributions over the temporal context, and it presents the proposed approach to derive the audiovisual fusion weights based on unimodal temporal context variance. Then, it introduces two key optimisation constraints that are imposed on the variance norms of unimodal latent distributions and discusses their implementation details.

3.3.1 Uncertainty-Aware Audiovisual Context Fusion

Quantifying modality-wise uncertainty towards predicting a common target label is crucial to improve multimodal fusion performance. Here, the objective is to first quantify intramodal uncertainty in the temporal context space, and then use the estimated uncertainty scores to derive the fusion weights. To this end, a method is proposed to learn unimodal latent distributions over the temporal context of the audio and visual modalities separately, as discussed below.

Latent Distributions over Unimodal Temporal Context

Figure 3.2 illustrates how the temporal networks (Gated Recurrent Unit (GRU)-RNNs) g_V and g_A are modified to output the parameters (mean and variance) of multivariate normal distributions $\mathcal{N}(\mu_V^i,\ {\sigma_V^i}^2)$ and $\mathcal{N}(\mu_A^i,\ {\sigma_A^i}^2)$ over the audio and visual temporal context vectors, respectively. Here, the term 'temporal context' refers to the hidden state outputs from the corresponding unimodal GRU blocks (g_A or g_V). For each modality separately, this hidden state output is learned as a multivariate normal distribution, instead of a typical deterministic embedding vector. This approach presumes that these unimodal latent distributions are capable of representing modality-wise emotion information more effectively than deterministic embeddings.

Given a sequence of frames, $[X_1,\ X_2,\ ...,\ X_T]$, in order to predict their corresponding target variables $[Y_1^*,\ Y_2^*,\ ...,\ Y_T^*]$ it is important to learn the underlying temporal context information, which is a function of the frames present in the input sequence as well as the order in which they appear. By modelling the temporal context as a probability distribution, we propose

to use the prediction error $\|Y_i - Y_i^*\|_2$ to constrain the contribution of each frame X_i in terms of its explained variance of the overall temporal context. Here, the idea of frame-wise explained variance of the temporal context refers to how much information a particular frame holds given all the rest of frames, towards predicting the target variable Y_i^*. Thus, the higher the explained variance of a particular frame X_i, the more informative it is for accurately predicting the target variable.

It is important to note the difference between the absolute variance of the temporal context distribution learned from all the frames and the explained temporal context variance of an individual frame. While the former can be thought of as a proxy metric for uncertainty measurement, the latter can be viewed as a per-frame information metric w.r.t the target prediction. For the sake of simplicity, throughout this chapter we use the term 'temporal context variance' in order to refer to the explained variance of temporal context for a given frame in an input sequence. The above argument can be extended to a multimodal fusion setting as well, in which the explained temporal context variance of a particular modality can be used as a proxy for how informative that modality is w.r.t predicting a common target variable.

The proposed method models the variance of a unimodal latent distribution as a proxy for how informative that modality is w.r.t. predicting the target emotion, and it uses the inverse of variance values to quantify how uncertain a particular modality is towards predicting emotion labels. Note that the potential of signal variance-based uncertainty modelling for multimodal fusion was already demonstrated in [Evangelopoulos et al., 2013]. Inspired by this idea, we model the unimodal context variance norm values $\|\sigma_V^2\|_2$ and $\|\sigma_A^2\|_2$ to estimate how certain the audio and visual modalities are about predicting the emotion labels. The proposed approach to de-

rive variance-based fusion weights for integrating the audiovisual temporal context vectors is discussed below.

Context Distribution Variance-Based Fusion Weights

For an input frame with index i, given its unimodal latent distributions $\mathcal{N}(\mu_V^i, {\sigma_V^i}^2)$ and $\mathcal{N}(\mu_A^i, {\sigma_V^i}^2)$ over its audio and visual temporal context separately, COLD fusion first computes the $L2$ norms of their variance values, $\|{\sigma_V^i}^2\|_2$ and $\|{\sigma_A^i}^2\|_2$. As discussed above, these variance norm values are assumed to represent modality-specific certainty w.r.t. predicting the target emotions. By normalising the variance norm values of the audio and visual modalities, this method derives fusion weights that are used in a simple linear fusion model of the audiovisual temporal context (h_{VA}^i) :

$$h_{VA}^i = w_V^i * h_V^i + w_A^i * h_A^i, \tag{3.1}$$

where h_V^i and h_A^i denote the visual and audio temporal context vectors, and w_V^i and w_A^i denote their corresponding weight values. The temporal context vectors h_V^i and h_A^i are sampled from their respective latent distributions, $h_V^i \sim \mathcal{N}(\mu_V^i, {\sigma_V^i}^2)$ and $h_A^i \sim \mathcal{N}(\mu_A^i, {\sigma_A^i}^2)$ during training. At test time, h_V^i and h_A^i are set to their corresponding mean vectors μ_V^i and μ_A^i for evaluation purpose.

Based on the unimodal context variance norm values ($\|{\sigma_V^i}^2\|_2$ and $\|{\sigma_A^i}^2\|_2$), the weight values w_V^i and w_A^i in Equation (3.1) are computed as:

$$w_V^i = \frac{\|{\sigma_V^i}^2\|_2}{(\|{\sigma_V^i}^2\|_2 + \|{\sigma_A^i}^2\|_2)}, w_A^i = \frac{\|{\sigma_A^i}^2\|_2}{(\|{\sigma_V^i}^2\|_2 + \|{\sigma_A^i}^2\|_2)}. \tag{3.2}$$

Context variance modelling seems to be a simple yet effective approach to uncertainty-aware audiovisual fusion, yet learning audiovisual latent distributions with well-conditioned variance ranges is non-trivial in practice, as shown later in the experiments. To condition the variance values that can effectively capture intramodal uncertainty w.r.t. predicting the target labels, this chapter defines a more principled model training approach that applies two key optimisation constraints: Calibration and Ordinality.

3.3.2 COLD: Calibrated and Ordinal Latent Distributions

To effectively learn the unimodal latent distributions for uncertainty-aware fusion, COLD fusion proposes to condition their variance values by applying optimisation constraints to the model training objective. It achieves this conditioning by imposing two key constraints: Calibration and Ordinality (or ranking) on the latent distribution variance vectors. When well-calibrated, an uncertainty score acts as a proxy for the correctness likelihood of its prediction for an individual input from a specific modality. In other words, well-calibrated uncertainty indicates the expected estimation error, i.e., how far the predicted emotion is expected to lie from its ground truth. Given the predictions made for a set of frames from different modalities, when their uncertainty scores are well-ranked or maintain ordinality, this approach can effectively arrange the input unimodal frames according to their reliability for predicting a target emotion. Figure 3.1 illustrates the definitions of both these constraints. *It is important to note the fundamental difference between these two constraints: while the calibration constraint is applied individually for each unimodal frame, the ordinality or ranking constraint is imposed jointly for a set of frames from different modalities.*

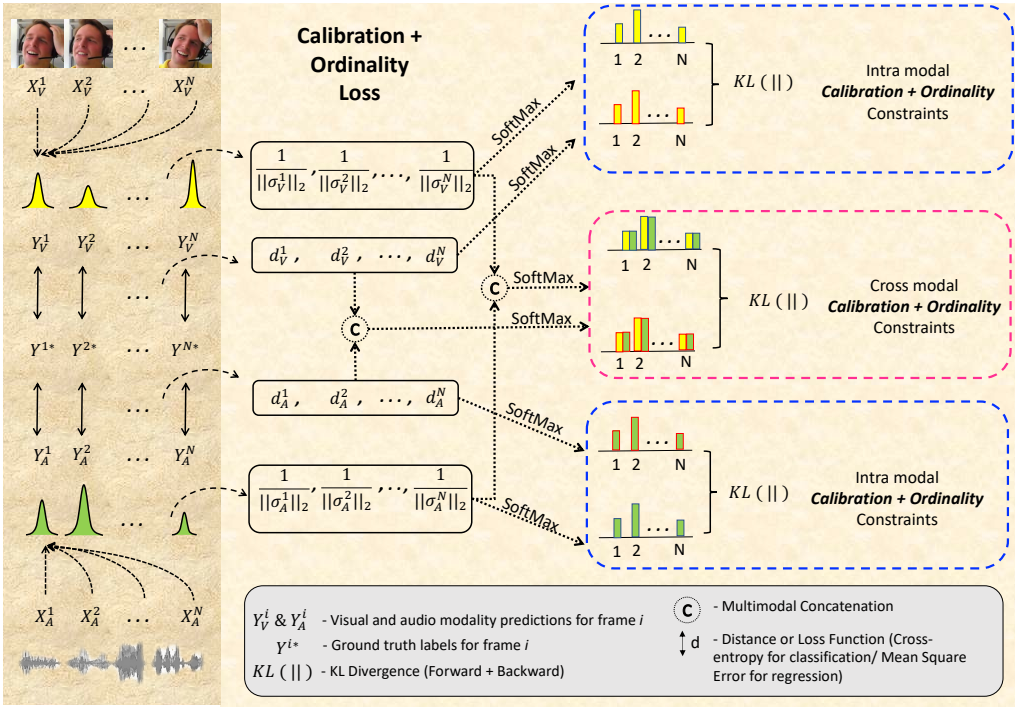


Figure 3.3: COLD fusion loss function: To simultaneously impose the calibration and ordinality constraints on the unimodal latent distributions' variance vectors, COLD fusion minimises the softmax distributional matching loss (KL divergence) between the distance vectors $[d^i]$ and variance-norm vectors $[\frac{1}{||\sigma^{i\,2}||_2}]$, in both intramodal and crossmodal settings.

Calibration Constraint – this is imposed by regularising the unimodal context variance norms, $\|{\sigma_V^i}^2\|_2$ and $\|{\sigma_A^i}^2\|_2$, such that their values are strongly correlated with the correctness likelihood values of target emotion classes. In regression models, this constraint can be implemented by forcing the variance norm values to correlate with the Euclidean distance between their corresponding unimodal predictions Y_V and Y_A and their ground truth labels Y^*, as shown in Fig. 3.1. In other words, the context variance values are learnt as reliability measures indicating how far the emotion predictions are expected to lie from their ground truth labels. To impose this property on the variance values of both modalities, COLD fusion applies the following regularisation constraints,

$$\begin{aligned} &\underset{\sigma_V^2}{\text{argmax}} \;\; Corr(\frac{1}{\|\sigma_V^2\|_2},\, d(Y_V, Y^*)) \\ &\underset{\sigma_A^2}{\text{argmax}} \;\; Corr(\frac{1}{\|\sigma_A^2\|_2},\, d(Y_A, Y^*)) \end{aligned} \tag{3.3}$$

where $d(.)$ denotes the distance function that measures the target emotion estimation error. Cross-entropy and Mean Squared Error (MSE) are used as the distance functions for the classification and regression models respectively.

Ordinality Constraint – this is applied to rank the frames of unimodal sequences, so that their uncertainty measures indicate how reliable different multimodal frames are w.r.t. each other. This ranking operation can be implemented as a simple ordering constraint which jointly regularises the unimodal context variance norm values, $\|{\sigma_V^i}^2\|_2$ and $\|{\sigma_A^i}^2\|_2$. Here, modality-wise reliability is again computed in terms of the distance values (see Equation (3.3)) between different unimodal predictions and the

ground truth labels:

$$\underset{\sigma_V^2,\sigma_A^2}{\text{argmax}} \; Corr(Rank(\frac{1}{\|\sigma_V^2\|_2}, \frac{1}{\|\sigma_A^2\|_2}),\; Rank(d(Y_V, Y^*), d(Y_A, Y^*))) \quad (3.4)$$

Implementation: Calibration and Ordinality Constrained Training for Audiovisual Emotion Recognition

Classification models of dimensional emotion recognition are trained in addition to the standard regression models used in the literature. In both cases, the underpinning principles of the COLD fusion are the same, but implementations of the training objective differ slightly. To train the temporal context fusion models by imposing the above-described calibration and ordinality constraints, the network is optimised to jointly minimise a loss function composed of the following components:

Emotion Prediction Loss (L_{emo}) is computed using the standard cross-entropy function for training the classification models. For the regression models training, similar to [Kossaifi et al., 2020], inverse Concordance Correlation Coefficient (CCC) loss (1.0 - CCC) is used in addition to MSE. This loss is computed for the predictions from unimodal (Y_V and Y_A) and multimodal (Y_{AV}) branches jointly (Figure 3.2).

Calibration and Ordinality Loss (L_{CO}) combines the aforementioned constraints, defined in Equation (3.3) and Equation (3.4), into a single training objective using differentiable operations. Figure 3.3 shows the steps involved in implementing this component: given an input sequence with N frames, their unimodal latent distributions followed by their corresponding unimodal predictions are inferred. To impose the calibration and

ordinality constraints, two sets of vectors for each modality are computed:

Distance Vectors. The scalar distance values (d_V^i and d_A^i) between the unimodal predictions (Y_V^i and Y_A^i) and the ground truth labels (Y^{i*}) are computed using either cross-entropy (classification) or MSE (regression) as the distance function. This step produces N-dimensional distance vectors, $D_V = [d_V^1, d_V^2.., d_V^N]$ and $D_A = [d_A^1, d_A^2.., d_A^N]$.

Variance-Norm Vectors. The inverted unimodal context variance norm values are collected into another set of N-dimensional vectors, S_V and S_A, as shown below:

$$\begin{aligned} S_V &= [\frac{1}{\|\sigma_V^1\|_2}, \frac{1}{\|\sigma_V^2\|_2}, .., \frac{1}{\|\sigma_V^N\|_2}] \\ S_A &= [\frac{1}{\|\sigma_A^1\|_2}, \frac{1}{\|\sigma_A^2\|_2}, .., \frac{1}{\|\sigma_A^N\|_2}]. \end{aligned} \tag{3.5}$$

Softmax Distributional Matching for Calibration and Ordinal Ranking. Note that the distance vectors and variance-norm vectors contain scalar values that summarise the properties of different embedding spaces, emotion labels, and temporal context, respectively. Hence, it is assumed that matching their properties by imposing the calibration and ordinality constraints directly in their original spaces, is not optimal. For this reason, as illustrated in Figure 3.3, the softmax operation is applied to the distance vectors and variance-norm vectors separately to generate the softmax distributions. Then, the calibration and ordinality constraints are imposed by minimising the mismatch between softmax distributions of the variance-norm vectors and distance vectors. This approach to calibration and ordinality loss computation based on soft-ranking is inspired by [Bruch et al., 2019] in which softmax cross-entropy is used for ordinal regression.

As Figure 3.3 shows, in both intramodal and crossmodal settings, soft-

max distributions of distance vectors (P_{D_V}, P_{D_A}, and $P_{D_{AV}}$) and variance-norm vectors (P_{S_V}, P_{S_A}, and $P_{S_{AV}}$) are computed. Note that in the crossmodal case, the audio and visual distance vectors and variance-norm vectors are concatenated separately, i.e., $D_{AV} = [d_A^1, d_V^1, .., d_A^N, d_V^N]$ and $S_{AV} = [s_A^1, s_V^1, ..., s_A^K, s_V^N]$. Then, the softmax operation is applied to the concatenated list which is $2N$ dimensional. Thus, the crossmodal softmax distributions capture the relational information across both modalities. Now, to impose the calibration constraint, the proposed method minimises the KL divergence (both forward and backward) between the distance distributions and variance-norm distributions in both intramodal and crossmodal settings, as shown below:

$$L_{CO} = KL(P_D||P_S) + KL(P_S||P_D), \tag{3.6}$$

where P_D represents P_{D_V} and P_{D_A}, and P_S represents P_{S_V} and P_{S_A} in the intramodal loss computation. In the crossmodal case, P_D and P_S denote $P_{D_{AV}}$ and $P_{S_{AV}}$, respectively.

Variance Regularisation Loss (L_{regu}). Prior works [Chang et al., 2020, Sanchez et al., 2021] on latent distribution learning in high-dimensional input spaces such as images, have reported that the variance collapse is a commonly encountered problem. Variance collapse occurs mainly because the network is encouraged to predict small variance σ^2 values to suppress the unstable gradients that arise while training the latent distribution models using Stochastic Gradient Descent. To prevent this problem, the regularisation term proposed in [Chang et al., 2020] is included in the training

objective:

$$\begin{aligned} L_{regu} &= KL(\mathcal{N}(\mu,\, \sigma^2)||\mathcal{N}(0,\, \mathbf{I})) \\ &= -\frac{1}{2}(1 + log\sigma^2 - \mu^2 - \sigma^2), \end{aligned} \tag{3.7}$$

where I denotes an identity variance matrix. Note that this regularisation term is applied to the audio and visual context latent distributions, separately.

In summary, the COLD fusion training objective composed of the above-discussed loss components, is as follows:

$$\begin{aligned} L_{total} = L_{emo} + \lambda_{CO_V} \cdot L_{CO_V} + \lambda_{CO_A} \cdot L_{CO_A} + \\ \lambda_{CO_{AV}} \cdot L_{CO_{AV}} + \lambda_R \cdot L_{regu}, \end{aligned} \tag{3.8}$$

where λ_{CO_V} (for visual-only), λ_{CO_A} (for audio-only), $\lambda_{CO_{AV}}$ (for audio and visual combined), and λ_R (for regularisation) are the optimisation hyperparameters that control the strength of each regularisation constraint. Here, the crossmodal loss term $L_{CO_{AV}}$ is computed by replacing P_D and P_S in Eq. 3.6 with $P_{D_{AV}}$ and $P_{S_{AV}}$ respectively.

3.4 Model-Agnostic Fusion Baselines

Before discussing the experimental evaluation of the proposed method, this section briefly discusses general multimodal fusion techniques w.r.t. audiovisual emotion recognition. One fundamental question in multimodal learning concerns the optimal stage to perform fusion [Baltrušaitis et al., 2018]. This chapter considers the following three typical model-agnostic

fusion methods as the standard baselines: feature fusion, temporal context fusion, and prediction fusion.

Feature Fusion or early fusion integrates frame-level emotion cues present in the audiovisual features Z_V and Z_A (e.g., [Zhang et al., 2017a]), not accounting for commonly encountered temporal misalignment between different modalities [Lingenfelser et al., 2016]. These per-frame audiovisual features are concatenated into a single sequence, $Z = [Z_V, Z_A]$, which are passed to a common temporal network $g_{AV} : Z \rightarrow Y$ to predict emotion labels.

Decision Fusion combines the unimodal emotion predictions Y_V and Y_A (e.g., [Ringeval et al., 2015]). Here, its implementations applies predictive confidence based weighted averaging to perform the late fusion. Unlike early fusion, late fusion does not leverage the low-level correspondences among the emotion cues distributed over the audio and visual streams [Baltrušaitis et al., 2018].

Temporal Context Fusion or simply context fusion integrates sequence-level emotion information aggregated in the form of audiovisual temporal context vectors h_V^i and h_A^i for frame i, produced by the temporal networks g_V and g_A respectively. This method is also referred to as 'feature fusion with RNNs' or 'mid-level' fusion in some prior works [Rouast et al., 2019, Tzirakis et al., 2017]. It is important to note that here, temporal context or simply context at i^{th} frame refers to the emotion information present in frame i w.r.t. the emotion information carried by remaining frames in the input sequence. As a result, unlike early fusion, context fusion is bound to suffer less from the temporal misalignment between the emotion-related semantics of audio and visual feature sequences. Further, context fusion benefits from the low-level audiovisual correspondences in the emotion space,

in contrast to late fusion.

Considering the above mentioned critical advantages of temporal context fusion, the proposed method aims to learn an uncertainty-aware context fusion model for multimodal emotion recognition as discussed above.

3.5 Experiments

This section first discusses the details of the dimensional emotion recognition dataset used for evaluating the proposed COLD fusion model. Then, it presents the regression and classification models' performance evaluation metrics, along with a standard uncertainty calibration error metric that applies to the classification models. Network architectural details of the visual and audio stream models and their fusion implementations, and their optimisation details can be found in Appendix A.

3.5.1 Dataset

For Spontaneous Emotion Recognition, the AVEC 2019 CES challenge corpus [Ringeval et al., 2019] is used, which was designed for in-the-wild emotion recognition in cross-cultural settings as part of the SEWA project [Kossaifi et al., 2019]. This corpus is composed of 8.5 hours of audiovisual recordings collected from German, Hungarian, and Chinese participants. All videos in this corpus are annotated with continuous-valued valence and arousal labels in the range [-1, 1]. Note that the train and validation partitions are composed of only German and Hungarian cultures. As the labels for the test set (which has the Chinese culture in addition) are not publicly available, the results are reported on the validation set.

The proposed audiovisual fusion models are trained on the AVEC 2019 CES dataset in regression as well as classification settings. Continuous-valued labels as targets in the range [-1, 1] are used to train the regression models. For classification, the continuous emotion values are mapped to three different classes for valence (positive, neutral, negative) and arousal (high, neutral, low) individually. For this binning, the thresholds of -0.05 and 0.05 are chosen to draw the boundaries between the three above-mentioned bins, such that they minimise the imbalances in the resultant class-wise label distribution.

Addressing Imbalanced Emotion Class Label Distributions. Although the binning thresholds are tuned carefully, the class-wise label distributions still have significant imbalances, as shown in Figure 3.4. To mitigate the effect of this problem, two general techniques are applied while training the classification models: a. non-uniform sampling of the training instances for different classes and b. class-weighted cross-entropy loss. In the former, the sampling criteria is modified to oversample for the minority classes and undersample for the majority classes based on the number of examples available for each class in the train set. In the latter technique, the cross-entropy loss values for different classes are divided by their relative bin size (in the train set).

3.5.2 Evaluation Metrics

Regression models' performance is measured using Lin's Concordance Correlation Coefficient (CCC) [Lawrence and Lin, 1989] between the predicted emotions y^o and their ground truth labels y^*

$$CCC = \frac{\rho_{y^* y^o}.\sigma_{y^*}.\sigma_{y^o}}{\left(\mu_{y^*} - \mu_{y^o}\right)^2 + {\sigma_{y^*}}^2 + {\sigma_{y^o}}^2}, \tag{3.9}$$

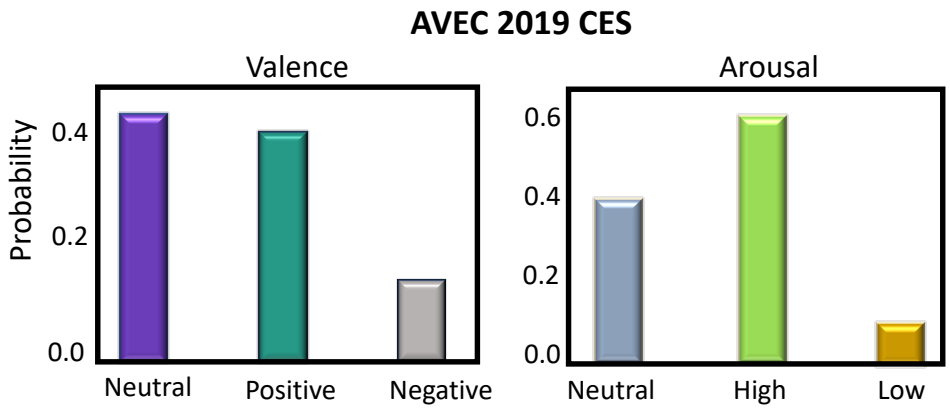


Figure 3.4: Class imbalances in the distribution of valence and arousal labels prepared for 3-way classification on the AVEC 2019 CES dataset.

where $\rho_{y^*y^o}$ denotes the Pearson's coefficient of correlation between y^* and y^o, and (μ_{y^*}, μ_{y^o}) and $(\sigma_{y^*}, \sigma_{y^o})$ denote their mean and standard deviation values, respectively.

Classification models are evaluated using precision, recall, and F1 score metrics, given that accuracy is not a reliable metric because of the imbalanced class distributions (see Figure 3.4). In all three metrics, the unweighted values or macro average values of the three emotion classes are computed for the valence and arousal dimensions separately.

Uncertainty Calibration Errors of the classification models are measured to analyse the deviations between the true class likelihoods p and the predicted class confidence estimates $\hat{p}$. Reliability diagrams [Guo et al., 2017] are used as empirical approximations to visually represent the confidence calibration errors. For plotting these diagrams, first, the accuracy and confidence axes are binned into equally-sized intervals and then, for each interval mean accuracy values are plotted against their corresponding mean confidence scores. For a perfectly calibrated model, the reliability diagram is supposed to be an identity function, i.e., accuracy and confidence should have the same values. Expected Calibration Error (ECE), a scalar summary statistic of the reliability diagram, computes the weighted

average of calibration errors over all the intervals in a reliability diagram.

$$ECE = \sum_{m=1}^{M} \frac{|I_m|}{N} |Acc(I_m) - Conf(I_m)|, \tag{3.10}$$

where I_m denotes the m^{th} interval, M is the total number of intervals, and N is the total number of samples.

3.6 Results and Discussion

This section first presents the results of dimensional emotion regression and classification models based on different audiovisual fusion techniques. By inducing visual noise through face masking, it investigates the robustness of the proposed COLD fusion compared to the standard fusion baselines. Then, an analysis of the uncertainty calibration performance of the COLD fusion model is presented, particularly in classification settings. To validate the improvements achieved by COLD fusion over the remaining fusion models, statistical significance tests are performed. Furthermore, a comparison between the proposed COLD fusion and a multimodal transformer baseline [Tsai et al., 2019] is provided. Finally, an ablation study of different components in the COLD fusion formulation is conducted, by nullifying different hyperparameters to modify the COLD training objective (Equation (3.8)).

3.6.1 Dimensional Emotion Recognition Results

Regression performance of different unimodal (Aud-branch and Vis-branch) and multimodal (AV) predictions are presented in Table 3.1. From these re-

Model	Valence CCC ↑	Arousal CCC ↑	Avg. CCC ↑
AVEC Winners:Aud [Zhao et al., 2019]	0.388	0.518	0.453
Aud-branch	0.369	0.465	0.417
AVEC Winners:Vis [Zhao et al., 2019]	0.579	0.594	0.586
Vis-branch	0.511	0.514	0.512
AVEC CES Winners:AV [Zhao et al., 2019]	**0.614**	0.645	0.629
AV Feature Fusion	0.515	0.509	0.512
AV Prediction Fusion	0.552	0.617	0.584
AV Context Fusion	0.578	0.620	0.599
AV COLD Fusion	0.611	**0.661**	**0.636**

Table 3.1: Dimensional emotion *regression* results on the **AVEC 2019 CES validation set** (CCC – Concordance Correlation Coefficient).

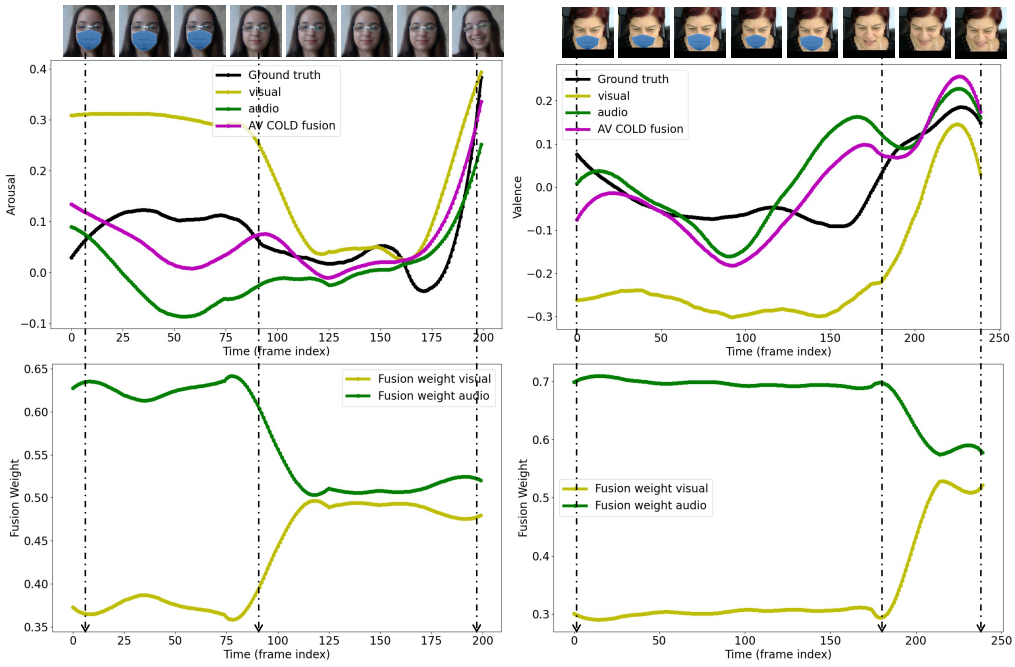


Figure 3.5: Dynamic adaptation of COLD fusion weights when presented with novel noise patterns induced into the visual inputs: At test time, face masking is applied to randomly chosen consecutive frames in the AVEC 2019 CES validation examples. When the visual modality is noisy, i.e., containing faces with masks, AV COLD fusion output relies more on the audio modality (note the gaps between visual predictions and AV COLD fusion predictions, and modality-wise fusion weights). After removing the face masks, the fusion weight values adapt accordingly, hence, the fusion outputs.

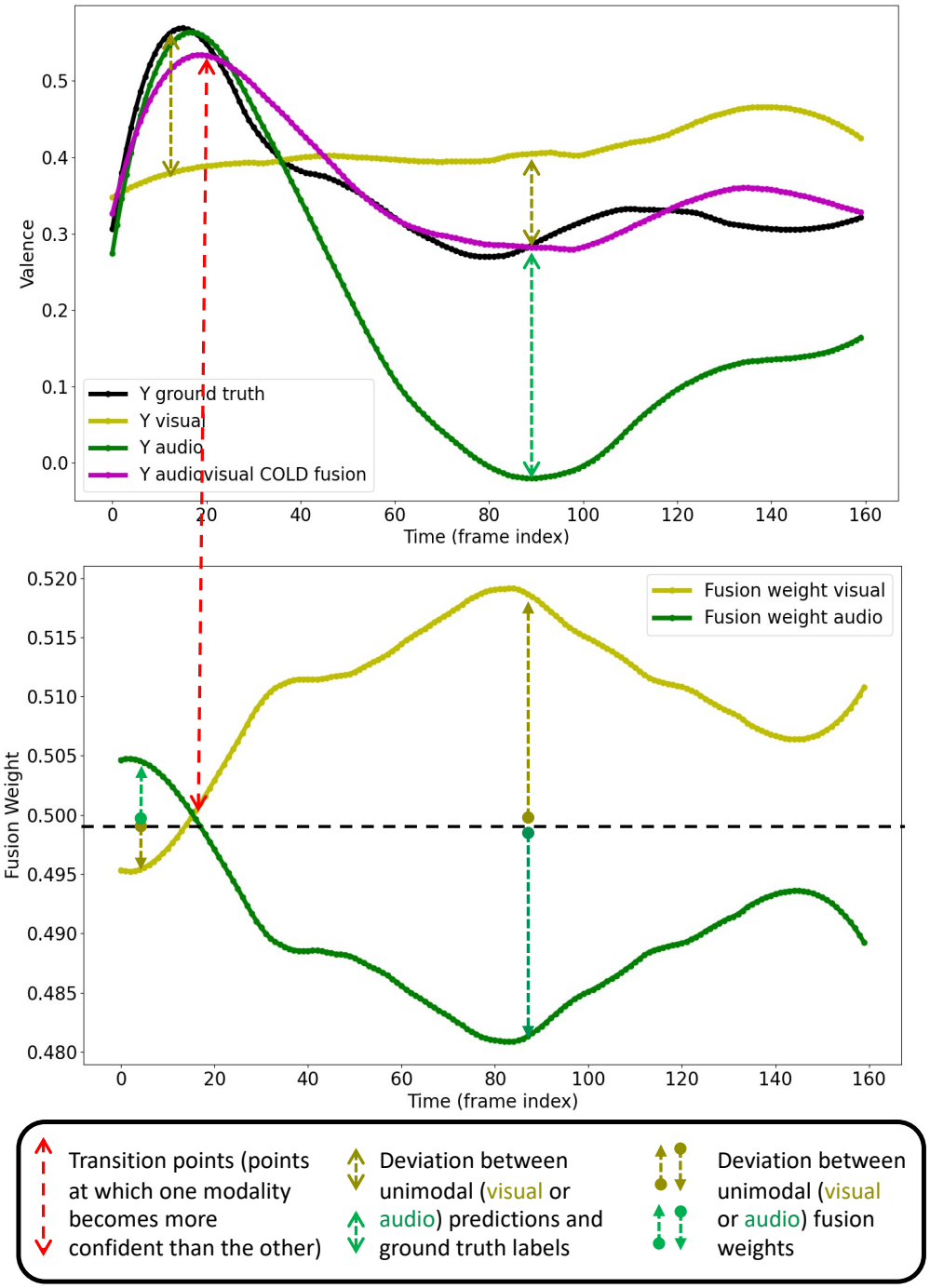


Figure 3.6: Emotion predictions on an example from the AVEC 2019 CES validation set: Unimodal and multimodal valence predictions, and their uncertainty-based fusion weights estimated by the AV COLD fusion predictions. Note that fusion weights of the audio and visual modalities demonstrate (a) the calibration property – how far their corresponding unimodal predictions are from the ground truth ratings and (b) the ordinal ranking property – how well they can order the audio and visual modalities in terms of their reliability.

	Valence			Arousal		
Model	**P↑**	**R↑**	**F1↑**	**P↑**	**R↑**	**F1↑**
Aud-branch	68.3	48.2	56.6	74.3	50.2	59.9
Vis-branch	70.9	58.1	63.9	76.8	70.3	73.4
AV Feature Fusion	67.8	60.2	63.8	73.4	68.2	70.7
AV Prediction Fusion	68.9	60.5	64.4	77.0	69.4	73.0
AV Context Fusion	75.0	60.6	67.0	77.1	71.1	73.9
AV COLD Fusion	**76.8**	**62.4**	**68.9**	**79.5**	**74.0**	**76.5**

Table 3.2: Dimensional emotion *3-way classification* results (P – Precision, R – Recall, F1 – F1 score) on the **AVEC 2019 CES validation set**.

sults, it can be clearly seen that the COLD fusion consistently outperformed the standard fusion baselines (feature, prediction and context) as well as the unimodal results. When compared to the best performing CNN+RNN fusion baselines, on average, COLD fusion achieved ∼6% relative improvement in dimensional emotion regression.

Compared to the winners of the AVEC 2019 challenge, [Zhao et al., 2019], COLD fusion performs well in terms of arousal and mean CCC scores. However, it is slightly worse in the case of valence CCC. Note that [Zhao et al., 2019] use a domain adaptation technique to cope with the cross-cultural variations in the audiovisual emotion expressions. However, this chapter's focus is not on coping with the cross-cultural variations, but primarily on improving the fusion performance. It is important to note that the proposed fusion technique is, in principle, complementary to the domain adaptation used in [Zhao et al., 2019].

Classification performance on the AVEC 2019 CES corpus is presented in Table 3.2. Similar to the regression results, COLD fusion demonstrates superior emotion classification results than the standard model-agnostic fusion baselines. Note that here the original regression problem is posed as a 3-way classification problem by discretising the continuous emotion

labels. For this reason, there are no existing benchmarks for comparison in this particular classification setting. Nevertheless, it is worth noting that the performance improvements achieved by the COLD fusion are consistent for both valence and arousal in terms of all three metrics.

Analysis of Fusion Baselines. Among the fusion methods that are evaluated here, temporal context or simply context fusion is found to be the second-best performing method after the proposed COLD fusion. Note that the temporal context refers to the output of the unimodal GRU block, and unimodal predictions are generated by applying a shallow fully connected network to the unimodal context vector. Thus, the context vectors can be viewed as higher-dimensional descriptors of the final unimodal predictions. Based on this assumption, in theory, the performance of context fusion is bound to be either better or at least as good as the prediction fusion, justifying the trends observed in our experimental results.

We can clearly notice that the feature fusion performance is inferior to all the remaining fusion techniques, and prediction fusion performs better than feature fusion. This result is consistent with an observation that prediction fusion achieves better results compared to feature fusion in general, as reported in the existing multimodal affect recognition literature [Ringeval et al., 2015]. It is worth noting that the results of feature fusion are worse than that of the best performing unimodal models. This performance degradation may be due to not explicitly correcting the temporal misalignment effects [Lingenfelser et al., 2016], which are heuristically derived in general [Ringeval et al., 2019]. This result indicates that integrating multimodal emotion information at feature-level or frame-level could be suboptimal most likely due to the temporal misalignment issues, given that continuous emotion information is expressed in the audiovisual modalities at different frame rates [Rouast et al., 2019, Tzirakis et al., 2017].

	Valence ECE ↓		Arousal ECE ↓	
Model	**BTS**	**ATS**	**BTS**	**ATS**
Aud-branch	13.6e-2	6.3e-2	3.2e-2	2.8e-2
Vis-branch	8.9e-2	7.1e-2	12.6e-2	3.1e-2
AV Feature Fusion	6.1e-2	5.0e-2	5.5e-2	3.3e-2
AV Prediction Fusion	8.7e-2	5.1e-2	2.5e-2	2.6e-2
AV Context Fusion	6.9e-2	**4.0e-2**	6.3e-2	3.0e-2
AV COLD Fusion	**3.7e-2**	4.3e-2	**1.3e-2**	**0.9e-2**

Table 3.3: Dimensional emotion classification *calibration* results on the **AVEC 2019 CES validation set** (ECE – Expected Calibration Error, BTS – Before Temperature Scaling, ATS – After Temperature Scaling).

Dynamic Adaptation of Fusion Weights in the Presence of Noise. In this experiment, the aim is to understand how different fusion models perform when presented with novel noise patterns at test time. By inducing noise into the visual modality through face masking, here, the performance of different fusion baselines is analysed, in comparison with the COLD fusion. For this evaluation, the face masks are overlaid as external occlusions on the image sequences using the method proposed in MaskTheFace [Anwar and Raychowdhury, 2020][1]. MaskTheFace is applied to 50% of the randomly chosen consecutive frames of the AVEC 2019 CES validation set sequences, as shown in Figure 3.5. Note that all the fusion models evaluated here have not seen faces with masks during their training. As Table 3.4 shows, in this noise-induced evaluation set up, performance drop compared to the noise-free evaluation (Table 3.1) is considerably higher for all three fusion baselines (feature, prediction, and context) than for the COLD fusion. Furthermore, the relative performance difference between the COLD fusion and the best performing fusion baselines is increased from ∼6% in noise-free settings to ∼17% in this noise-induced case.

Figure 3.5 compares the COLD fusion predictions with the predictions from

[1]https://github.com/aqeelanwar/MaskTheFace

Model	Valence CCC ↑	Arousal CCC ↑	Avg. CCC ↑
AV Feature Fusion	0.378	0.351	0.364
AV Prediction Fusion	0.363	0.545	0.454
AV Context Fusion	0.385	0.508	0.445
AV COLD Fusion	**0.491**	**0.574**	**0.528**

Table 3.4: Impact of visual noise (external occlusions) on the AV fusion models: Dimensional emotion *regression* results *with 50% of randomly chosen face images masked during evaluation (see Fig. 3.5)* on the AVEC 2019 CES validation Set.

visual and audio branches, along with the inferred modality-wise fusion weight scores. We can clearly see that the visual fusion weights are much lower for the frames with masks compared to the frames without masks, and as a result, the final predictions rely more on the audio modality in the presence of visual noise. This result clearly demonstrates the ability of COLD fusion to dynamically adjust the importance of a specific modality according to how informative it is towards recognising the target emotions.

3.6.2 Uncertainty Calibration Performance Analysis

To measure the quality of uncertainty estimates, Expected Calibration Error (ECE) (see Section 3.5.2) values are computed for the unimodal and multimodal emotion classification models. Note that this calibration error metric applies only to the classification settings. By computing the ECE values before and after applying temperature scaling to the softmax distributions over the predictions of each model separately, the impact of explicit uncertainty calibration (temperature scaling) is investigated. An optimal temperature value is searched in the range of $1e-2$ to 1000 by doing a random search for 100 iterations. Similar to the technique fol-

Model	**Valence** CCC ↑	**Arousal** CCC ↑
With All three constraints ($\lambda_{CV} = \lambda_{CA} = \lambda_{CAV} = 1e-3$, $\lambda_R = 1e-4$)	**0.605**	**0.661**
Without Intramodal constraints ($\lambda_{CV} = \lambda_{CA=0}$)	0.573	0.615
Without Crossmodal constraint (λ_{CAV})	0.580	0.609
Without Regularisation constraint ($\lambda_R = 0$)	0.541	0.595
Without Any constraints ($\lambda_{CV} = \lambda_{CA} = \lambda_{CAV} = 0$, $\lambda_R = 0$)	0.517	0.578

Table 3.5: Ablation experiments on the proposed loss function (Eq. 3.6): Analysing the impact of different loss components in the COLD Fusion on the AVEC 2019 CES validation set (CCC-Concordance Correlation Coefficient).

Model	**Valence CCC ↑**	**Arousal CCC ↑**	**Avg. CCC ↑**
Transformer [Tsai et al., 2019]†	0.602	0.619	0.610
Proposed AV COLD Fusion	**0.611**	**0.661**	**0.636**

Table 3.6: Comparison with a pair-wise crossmodal self-attention based multimodal transformer [Tsai et al., 2019] († indicates in-house implementation for AV fusion): Regression results on the AVEC 2019 CES validation set.

lowed in [Mukhoti et al., 2020], the temperature value is chosen such that it achieves the lowest ECE value on the validation set.

It is important to consider that the COLD fusion models are trained to be implicitly calibrated (see Equation (3.6)) in terms of their temporal context variance values. Thus, even before applying explicit calibration, i.e., temperature scaling, we expect the predictive uncertainty values or class wise confidence scores of the COLD fusion models to have lower ECE values compared to the other fusion baselines.

Quantitative Results. Table 3.3 reports the ECE values for valence and

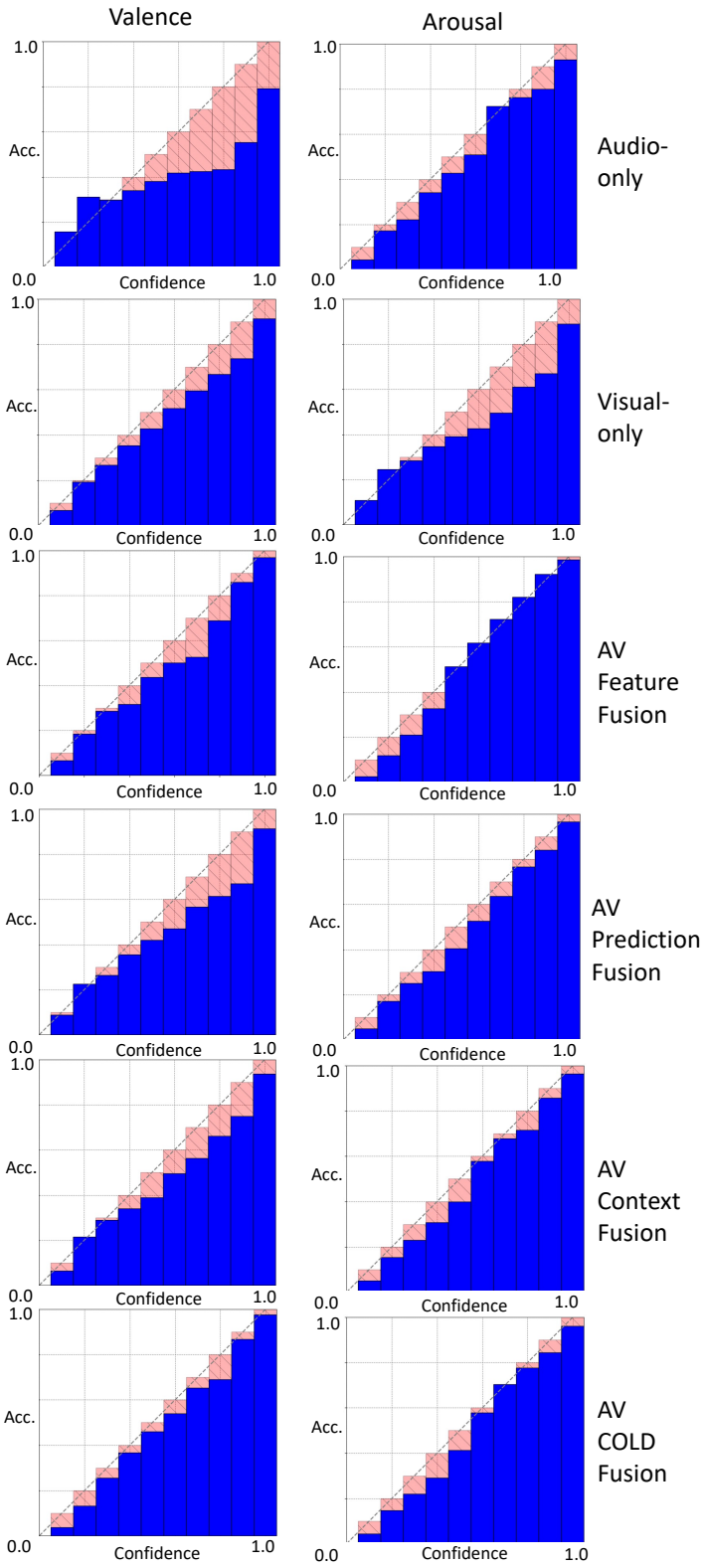


Figure 3.7: Reliability plots of unimodal and multimodal classification models evaluated on the AVEC 2019 validation set. A perfectly calibrated model should appear as a perfect right angled triangle, as marked by the diagonal lines and the red bars.

arousal attributes on the AVEC 2019 corpus. For both the attributes, before the application of temperature scaling, COLD fusion has the lowest calibration error when compared to the other models. After applying temperature scaling, it is obvious that the ECE values for all the models go down, and the COLD fusion still achieves the lowest error. Only in the case of valence, AV context fusion has a marginally lower ECE value compared to the COLD fusion. This minor discrepancy could be due to the random search of optimal temperature values and note that here, different models have different optimal temperature values that are tuned for valence and arousal, separately. Nevertheless, in all the remaining cases (both before and after temperature scaling), COLD fusion consistently demonstrates lower uncertainty calibration errors in relation to the other fusion models.

Reliability diagrams visually illustrate the uncertainty calibration performance of a model's predictions. As Figure 3.7 shows, when a model is perfectly calibrated, its confidence score vs the accuracy score histogram looks like a perfect right-angled triangle. The more the deviations are from the diagonal lines in them, the higher their ECE values are. Note that ECE is a scalar summary statistic of a reliability diagram, which computes the weighted average of such deviations over all the intervals in the reliability diagram. Though the ECE values reported for the AVEC 2019 corpus (Table 3.3) already validate the improved calibration results with COLD fusion. Here, as an example, Figure 3.7 compares the reliability plots of different models evaluated on the AVEC validation set. In Figure 3.7, we can see that compared to the unimodal cases and other fusion baselines, the COLD fusion reliability plot looks much closer to a perfect right-angled triangle. Among all the reliability plots illustrated, it can be observed that the audio branch for valence has the highest calibration error. This observation is in line with the poor performance achieved by the audio modality

Model Pair	Valence p-Value ↓	Arousal p-Value ↓
(Aud-branch, AV COLD Fusion)	6.4e-34	9.7e-23
(Vis-branch, AV COLD Fusion)	1.9e-15	3.3e-16
(AV Feature Fusion, AV COLD Fusion)	3.7e-14	1.1e-18
(AV Prediction Fusion, AV COLD Fusion)	5.5e-9	7.3e-3
(AV Context Fusion, AV COLD Fusion)	1.2e-4	2.0e-3
(AV COLD Fusion, AV COLD Fusion)	1.0e-0	1.0e-0

Table 3.7: Statistical significance testing (p <0.01): *Regression* t-test results on the AVEC 2019 CES validation set.

in terms of the valence prediction error (see Table 3.1 and Table 3.2) on the AVEC 2019 corpus.

Analysis of Audiovisual Fusion Weights. Figure 3.6 qualitatively illustrates modality-wise fusion weights estimated by the COLD fusion model on a validation sequence taken from the AVEC 2019 corpus. Note that these fusion weights are functions of the unimodal temporal context distributions (see Equation (3.2)). This illustration analyses the temporal patterns of fusion weights along with their corresponding unimodal and multimodal emotion predictions and their ground truth labels. This analysis clearly shows the well-calibrated nature of modality-wise fusion weights: when the predictions of one modality move closer to the ground truth compared to those of the other modality, the audiovisual weight values in the COLD fusion are found to be varying accordingly. From the transition points marked in Figure 3.6, we can see that the fusion weights are gradually inverted, as the predictions of one modality move closer to the ground truth while the other modality predictions move further. This result validates our main hypothesis of making unimodal latent distributions calibrated and ordinal for improved fusion performance.

Statistical Significance Analysis. As shown in Table 3.7, a paired t-test is performed on the validation set of the AVEC 2019 corpus, to verify the statistical significance of COLD fusion's performance improvements over the unimodal and remaining multimodal baselines. In line with the trends in regression, performance reported in Table 3.1, p-values of the student t-test indicate that the improvements achieved by the COLD fusion models are statistically quite significant compared to the baseline models.

Comparison with a Multimodal Transformer [Tsai et al., 2019]. In addition to the standard fusion baselines, a multimodal transformer model is implemented based on pair-wise crossmodal self-attention fusion proposed in [Tsai et al., 2019]. It is worth noting that the crossmodal self-attention fusion aims to cope with the problem of temporal misalignment between different modalities during fusion, similar to the temporal context fusion model we evaluated in this chapter. An audio-visual version of this multimodal transformer method is implemented by tailoring its original network architecture designed for the text, audio, and visual modalities[2]. A 3-layer self-attention network with 16 heads followed by an FC output layer, is used to implement this multimodal transformer baseline. As shown in Table 3.6, regression results on the AVEC 2019 CES corpus show that the COLD fusion clearly outperformed the transformer baseline, particularly in arousal prediction, by a large margin.

Ablation Studies. Table 3.5 presents the ablation results that quantify the contributions of calibration, ordinality, and context variance regularisation constraints to the performance gains achieved by COLD fusion. By individually nullifying the four optimisation hyperparameters of the COLD training objective (see Equation (3.8)), the emotion regression performance is measured on the AVEC 2019 validation set. Compared to the fully con-

[2]https://github.com/yaohungt/Multimodal-Transformer

strained COLD fusion model, different partially constrained and fully unconstrained models listed in Table 3.5, achieve considerably lower CCC scores. Most importantly, discarding the variance regularisation constraint results in more performance degradation than the remaining constraints. This observation indicates the importance of preventing the variance collapse problem by using the variance regularisation term, in line with the results reported in prior works [Sanchez et al., 2021, Chang et al., 2020].

3.7 Conclusion

This chapter introduced an uncertainty-aware multimodal fusion approach to dimensional emotion recognition from audiovisual data. To capture modality-wise uncertainty w.r.t. predicting valence and arousal dimensions, the proposed method probabilistically modelled the temporal context of faces and voices by learning unimodal latent distributions. For effective uncertainty-weighted audiovisual fusion, this method proposed to condition the unimodal latent distributions such that their temporal context variance norms are learnt to be *well-calibrated* and *well-ranked (ordinal)*. To jointly impose these two constraints on the latent distributions, it introduced a novel softmax distributional matching loss function that encourages the uncertainty scores to be well-calibrated and well-ranked. The proposed novel loss function for multimodal learning is applicable to both classification and regression settings.

On an in-the-wild spontaneous emotion recognition dataset, the proposed uncertainty-aware fusion model achieved significantly better recognition performance than the uncertainty-unaware model-agnostic fusion baselines, including a multimodal transformer [Tsai et al., 2019]. Validating the main

hypothesis of this chapter, extensive ablation studies showed that it is important to apply both calibration and ordinality constraints for improving the emotion recognition results of uncertainty-aware fusion models. Furthermore, the proposed COLD fusion models demonstrated noticeable improvements in terms of predictive uncertainty calibration errors of the emotion recognition models. It is important to note that the proposed calibration and ordinal ranking constraints can be easily applied to general model-fusion methods as well by quantifying model-wise predictive uncertainty values of emotion labels. In summary, this chapter showed the importance of uncertainty modelling for dynamic integration of emotional expression cues from multimodal signals.

Chapter 4

Affective Processes: Stochastic Modelling of Temporal Context for Audio-Visual Affect Recognition

> Probability theory is nothing but common sense reduced to calculation.
>
> *Pierre-Simon Laplace*

Chapter Summary. As demonstrated in the previous chapter, temporal context modelling is critical to recognise apparent emotions from face images and speech signals. Most existing temporal context models, similar to the ones used in the COLD fusion mechanism as discussed in Chapter 3, build on recurrent models or in the modelling of contextual dependencies at the feature level using self-attention. This chapter argues that such

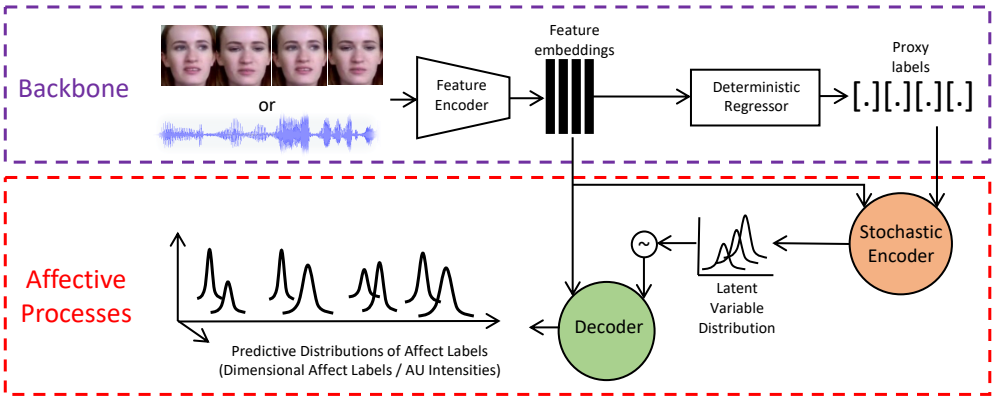


Figure 4.1: Affective Processes: stochastic temporal context modelling of affect labels from faces and voices. Given a sequence of feature embeddings and their proxy labels, a distribution over temporal functions is learned using a global latent variable.

canonical temporal models fail to effectively capture the long-term temporal dependencies that subtly occur at different levels of abstraction. To address this problem, this chapter introduces a novel uncertainty-aware temporal context modelling framework, Affective Processes. The proposed framework in this chapter aims to achieve superior affect recognition performance with little additional modelling complexity, by learning the temporal affect dynamics through a probabilistic global latent variable that captures context and induces dependencies in the outputs. In this chapter, first the formulation of Affective Processes is presented and then it is applied to visual-only, audio-only, and audio-visual affect recognition problems. Further, to improve the label efficiency of video-based affect recognition, an application of Affective Processes to Cooperative Machine Learning settings is proposed.

4.1 Introduction

Given the intrinsic temporal nature of emotion information expressed in faces and voices, the ability to capture long range global temporal context

is of paramount importance in building affect recognition models. In general, the problem of modelling temporal dynamics of facial expressions and affect has been studied over several years [Valstar, PhD Thesis, Imperial College London, 2008, Martinez et al., 2017, Kollias et al., 2020], however, effectively modelling the temporal context for in-the-wild affect recognition is still a challenging problem. Particularly, not much attention has been paid so far to the idea of probabilistic learning of affect-specific temporal context from audiovisual signals.

De facto temporal models used in the state-of-the-art methods [Toisoul et al., 2021, Ringeval et al., 2019, Zhao et al., 2019] are typically composed of canonical sequence learning models based on recurrent temporal models, or recently introduced self-attention mechanisms [Vaswani et al., 2017]. To model the temporal context of valence and arousal attributes in face and voice data, most existing works use canonical sequence learning models such as Time delay neural networks [Meng et al., 2015], vanilla recurrent neural networks and their variants (LSTM-RNNs and GRU-RNNs) in unidirectional or bidirectional manner [Kollias and Zafeiriou, 2020, Wang and Hsu, 2017, Deng et al., Tellamekala and Valstar, 2019, Zhao et al., 2019]. Recently, 1D Convolution networks have been shown as more efficient alternatives to RNNs, particularly in the case of audio dimensional affect recognition [Schmitt et al., 2019]. With the success of Transformers [Vaswani et al., 2017] in Natural Language Processing in recent years, multi-head attention has been applied to dimensional affect recognition problem as well [Huang et al., 2019].

Unlike the CNN+RNN based approaches that decouple spatial and temporal context processing, 3D CNNs [Zhang et al., Kuhnke et al.], 3D ConvLSTMs [Huang et al., 2018] and temporal hourglass networks [Du et al., 2019] jointly model the spatio-temporal contexts. While these temporal regres-

sion models have demonstrated superior performance in terms of temporal dynamics modelling in general, this chapter argues that they fall short in the case of temporal affect context modelling due to: (a). their inability to model the temporal context uncertainty and (b). their deterministic function learning nature that constrains their representation capacity to accommodate a wide range of temporal context variations. Note that here deterministic function learning nature means that at test time the aforementioned temporal regression models produce a fixed output sequence for a given input sequence. In spite of achieving reasonably good performance in practice, all the aforementioned models essentially learn a single temporal regression function by posing affect recognition as a deterministic regression problem, ignoring the inherently stochastic nature of the temporal affect estimation task.

This chapter argues that the aforementioned two temporal context modelling properties are essential to improve the generalisation performance of in-the-wild affect recognition models for the following reasons: (i). cues of affect are often sparsely and irregularly distributed over the temporal input signals (face image sequences and speech data) and (ii). affect label annotation is a highly subjective task, making *ground truth* temporal affect labels often ambiguous [Sethu et al., 2019] and not very consistent across different training sequences due to inter-rater disagreements.

4.2 Background: Stochastic Process Modeling

Gaussian Processes (GPs) [Rasmussen, 2003] is one of the widely used approximation methods for stochastic processes modeling [Williams

and Rasmussen, 2006, Wang et al., 2019, Trapp et al., 2020, Tresp, 2001]. Despite their flexibility, data-efficiency and probabilistic nature, GPs have two key limitations that constrain their applicability to large-scale regression problems: computationally intensive inference and the requirement of a hand-designed covariance function (kernel function) based on prior knowledge of the problem at hand.

Neural Processes (NPs) family [Garnelo et al., 2018a,b, Kim et al., 2018, Gordon et al., 2020, Singh et al., 2019, Lee et al., 2020] addresses both the limitations of GPs by adopting a data-driven context learning approach and by leveraging the computational efficiency of inference in deep neural networks. The unique function modeling features of NPs are very appealing especially for learning problems that involve stochastic function modelling; this chapter hypothesises that temporal affect recognition can benefit from stochastic function learning given the ambiguous nature of affect labels. However, extending NPs to large-scale temporal regression tasks is constrained by the condition that at test time NPs require ground truth labels for the context frames.

To address these challenges, this chapter introduces 'Affective Processes' (APs) as a more effiective alternative to learn the temporal context using a stochastic, global latent variable model. APs is the first method that demonstrates how to circumvent the need for ground truth context frame labels during inference by using noisy predictions of one or more deterministic regression models. As illustrated in Fig. 4.1, APs aims to model the stochastic process behind the temporal dynamics of emotions using Neural Processes [Garnelo et al., 2018b], which, contrary to the recurrent temporal models used in the COLD fusion, only assume the output distributions to be permutation-invariant. Neural Processes seek to model each training sequence as a realisation of some underlying stochastic pro-

cess (SP) with exchangeable joint finite distribution, and uses an encoder-decoder architecture with learnable parameters to model it. The encoder is a DeepSet [Zaheer et al., 2017] that captures the global context, from a provided set of input-label pairs, using a stochastic latent variable, which is used by the decoder to carry the function-specific predictions. To avoid the limitation of needing manually given input/output contextual information at test time, Affective Processes extend the Neural Processes family by using a pre-trained backbone that is tasked with delivering the input features and corresponding predictions, referred to as proxy labels, as well as with a context selection method that estimates the optimal frames for global context modelling. By posing the temporal affect recognition as a stochastic process regression problem, APs aim to render the flexibility of temporal context uncertainty modelling to affect recognition models. With an architecture that adds negligible complexity to the backbone, Affective Processes is designed to be a strong candidate for the temporal modelling of emotions, to advance the current state of the unidmodal and multimodal affect recognition models.

4.3 Method

In temporal affect modelling, given labelled sequences of face images or speech features $\{X_N = [x_1, x_2, ..., x_N]\}$ and their corresponding affect label sequences $\{Y_N = [y_1, y_2, ..., y_N]\}$, the objective is to learn a temporal function that maps the input frame sequence to the target label sequence $f : X_N \rightarrow Y_N$.

Assuming that a single deterministic temporal function f is inadequate to capture the wide range of intrinsic subjective variations of in-the-wild affect

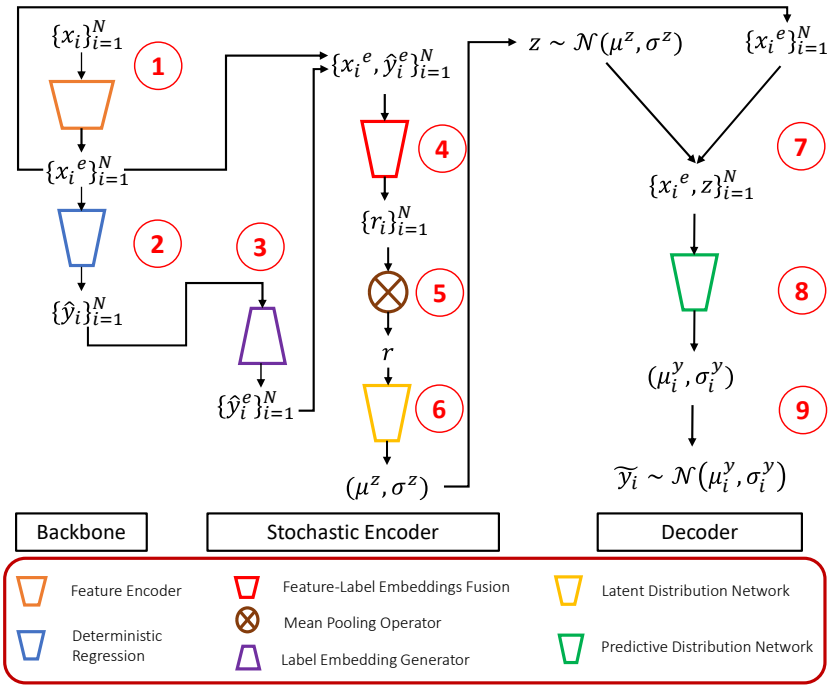


Figure 4.2: Building blocks of Affective Processes (APs) (see Section. 4.3.2 for a detailed description of each block above)

data and to model the temporal context uncertainty, this chapter instead proposes to learn a distribution over the temporal functions $\mathcal{D}(f)$ i.e. a stochastic process. At training time, given the labelled training sequences, the proposed method learns the function distribution $\mathcal{D}(f)$, and at test time a function f' is sampled from the distribution $\mathcal{D}(f)$ by conditioning it on a given test input sequence X'_N i.e. $f' \sim \mathcal{D}(f|X'_N)$.

4.3.1 Affective Processes

Affective Processes (APs) build on the recently proposed Neural Processes family [Garnelo et al., 2018b,a] to model exchangeable (i.e. permutation-invariant) stochastic processes, using an encoder-decoder architecture with a global latent variable. **Neural Processes (NPs)** are a class of deep neural latent variable models that combine the representation learning abilities of deep neural networks with the abilities of Gaussian Processes to learn

distribution of functions. Neural Processes use a (stochastic) global latent variable, captured by an encoder from a context set of input/output observations, which is used by a decoder in an individual basis to estimate the outputs for a set of target inputs. By conditioning the outputs on a global latent variable drawn from a latent function, Neural Processes model a particular family of Stochastic Processes: those whose finite marginals are modelled by an exchangeable distribution (i.e. permutation-invariant[1]). The learning of the Neural Process is inspired by that of Variational Auto Encoders [Kingma and Welling, 2013], where an encoder is used to perform amortized variational inference.

Let us assume we are given a training sequence composed of features X_N^e and annotated labels Y_N. The Neural Processes paradigm randomly chooses a subset of these points to act as *context*, that will be used to reconstruct the whole sequence labels $\tilde{Y}_N$. The goal is to maximise $P(\tilde{Y}_N)$. The **NP encoder** f_{enc} maps each of the context points, composed of input/output observations ($\{x_c^e, y_c\}$), into a latent representation r_c. The latent representations are then averaged yielding a global, fixed-dimension representation r_C of the context points[2]. Then, a small MLP transforms the global latent representation into the mean μ^z and variance σ^z vectors of a multivariate Gaussian distribution, from which the latent stochastic variable can be sampled, i.e. $Z \sim \mathcal{N}(\mu^z, \sigma^z)$. Thanks to the reparametrization trick [Kingma and Welling, 2013] learning the parameters of the encoder is enabled through the typical backpropagation techniques. A sample latent vector drawn from this distribution $Z \sim q(Z)$ is equivalent to a realisation of the underlying Stochastic Process F, that maps each of the points x^e

[1]It is out of the scope of this paper to show the formalism behind this notion, which builds on De Finetti's theorem [De Finetti, 1937]. Please refer to [Garnelo et al., 2018a] for further details

[2]Any permutation-invariant operation can be used to aggregate the individual representations

in the sequence to their corresponding labels y. The mapping F is done through the **NP decoder**, which is parameterised as a Neural Network that in addition to the input features x^e receives the sampled vector Z. As pointed out in [Garnelo et al., 2018b], the randomness in F is given by the latent variable Z[3]. Using this sampled latent vector Z as input, then the NP decoder outputs the predictive distribution for the target labels.

Neural Processes to Affective Processes. The direct application of NPs to temporal affect recognition would be constrained by the requirement of ground truth labels for the context points at test time. To circumvent the role of ground truth labels in the context inference, APs propose to make use of 'proxy labels' $\{\hat{y}_i\}_{i=1}^N$ produced by a deterministic regression model. Though the proxy labels are noisy and erroneous, it is assumed that they can still be informative about the underlying function space in order to guide the latent variable inference process.

4.3.2 Building Blocks of Affective Processes

i. Feature and Proxy Label Extraction. Given an input sequence of visual or audio frames $\{x_i\}_{i=1}^N$, a pre-trained frozen backbone is used to extract per-frame feature vectors $\{x_i^e\}_{i=1}^N$ and their corresponding proxy labels $\{\hat{y}_i\}_{i=1}^N$ (noisy label predictions, see Fig. 4.2). The extracted feature vectors are further transformed into a low-dimensional feature embedding $\{x_i^e\}_{i=1}^N$ using a feature embedding net. The proxy labels are also mapped to label embedding vectors $\{\hat{y}_i^e\}_{i=1}^N$ using a label embedding net. Note that the backbone model could be a static or temporal regression model depending on the input modality; it is independently trained and during training of the remaining AP modules it remains frozen.

[3]In technical terms, $P(F)$ would represent the pushforward measure of $P(Z)$

ii. Encoding Stochastic Latent Variable. Given the target frames' feature embeddings $\{x_i^e\}_{i=1}^N$ and their corresponding proxy label embeddings $\{\hat{y}_i^e\}_{i=1}^N$, first a set of context points $\{(x_i^e, \hat{y}_i^e)\}_{i=1}^C$ is randomly sampled. As Fig. 4.2 shows, these context points are mapped to per-frame deterministic context vectors $\{r_i\}_{i=1}^N$. Then a mean pooling operator is applied to the per-frame context vectors to derive a global deterministic context vector r, which is further mapped to the parameters (μ^z, σ^z) of a stochastic latent distribution.

iii. Decoding Predictive Distributions of Output Labels. The decoder of AP takes as input a latent vector Z sampled from the encoder output distribution $\mathcal{N}(\mu^z, \sigma^z)$, paired with feature embeddings of all frames in the input sequence $\{x_i^e\}_{i=1}^N$ (target frames). As Fig. 4.2 illustrates, the decoder outputs the predictive distributional parameters $\{(\mu_i^y, \sigma_i^y)\}_{i=1}^N$ for the output labels, by treating the target affect labels as random variables with univariate normal distribution. During inference, for the practical applications where only one prediction must be given, the predictive mean is used as the output, whereas the variance is used to represent the latent uncertainty on that estimation.

4.3.3 Training of Affective Processes

Loss Functions. Similar to VAEs [Kingma and Welling, 2013], existing implementations of NPs are trained to minimise the Evidence Lower Bound Objective (ELBO) function [Kingma and Welling, 2013],

$$L_{ELBO} = \mathbb{E}(-\log p(y_i|x_i, z)) + KL(q_c||q_t) \tag{4.1}$$

where $p(y_i|x_t, z)$ is the likelihood of ground truth label y_i with the AP output predictive distribution $\mathcal{N}(\mu_i^y,\sigma_i^y)$, q_c and q_t are the AP encoder output latent distributions $\mathcal{N}(\mu_c^z,\sigma_c^z)$, $\mathcal{N}(\mu_t^z,\sigma_t^z)$ for the context frames and target frames respectively. However, when the basic ELBO function is used for training APs — extension of NPs to large-scale noisy temporal regression tasks such as dimensional affect recognition — this work identifies that APs are vulnerable to two distribution collapse problems: encoder posterior collapse and decoder predictive variance collapse.

Encoder Posterior Collapse. Posterior collapse is a commonly encountered problem in generative latent variable models training [Lucas et al., 2019b,a], particularly when the training data contain high-dimensional noisy inputs, like that of temporal affect recognition. This collapse refers to a condition in which the decoder is encouraged to partially or completely ignore the noisy latent variable Z from the encoder in the early stages of model training, particularly when the decoder has high representation capacity.

This work finds that APs are also vulnerable to the posterior collapse problem—the decoder learning to give significantly less importance to the latent variable compared to the feature embeddings during training. This behaviour is expected given that the backbone features are already trained on the task at hand whereas the stochastic latent variable Z is learned from scratch, and it leads to an uninformative latent distribution. This work uses a commonly applied technique, Beta scheduling [Fu et al., 2019], in the conditional VAE literature to avoid the posterior collapse during APs training. In beta scheduling [Fu et al., 2019], the KL divergence regularization term in the ELBO objective is multiplied with a variable β as shown in Eq. 4.2. Here the value of beta is cyclically tuned in the range of 0 to 1 over the training iterations to improve the quality of learned latent

variable distributions [Fu et al., 2019].

Decoder Predictive Variance Collapse. The AP decoder tends to output smaller values for the variance σ of predictive distribution of output labels, in order to suppress unstable gradients during training, similar to the variance collapse problem reported in [Chang et al., 2020]. Due to this variance collapse, the decoder output predictive distributions end up with smaller variance values, degenerating the AP decoder to almost a deterministic model. Note that this variance collapse degrades the quality of latent variable distribution learned by the encoder. To address this problem, similar to the training of Variational Information Bottleneck [Chang et al., 2020, Alemi et al., 2016], this work applies a weighted regularisation term that forces the decoder output distribution to be closer to a normal distribution $\mathcal{N}(0,1)$. To achieve this, the model is trained to minimise the KL divergence between the AP output distribution $\mathcal{N}(\mu_i^y,\sigma_i^y)$ and the normal distribution $\mathcal{N}(0,1)$ in addition to the beta scheduled ELBO function. Thus, the complete AP loss function is reformulated by adding the above two regularization techniques, as follows:

$$\begin{aligned} L_{AP} = \mathbb{E}(-\log p(y_i|x_t,z)) + \beta * KL(q_c||q_t) + \\ \lambda * KL(\mathcal{N}(\mu_i^y,\sigma_i^y)||\mathcal{N}(0,1)), \end{aligned} \tag{4.2}$$

where β is cyclically tuned in the range of [0,1], $\mathcal{N}(\mu_i^y,\sigma_i^y)$ is the decoder output predictive distribution, λ is a hyper parameter, q_c and q_t are the encoder output latent distributions ($\mathcal{N}(\mu_c^z,\sigma_c^z)$ and $\mathcal{N}(\mu_t^z,\sigma_t^z)$) for the context and target point sets respectively.

4.3.4 Affective Processes for Audio-Visual Affect Fusion

By leveraging the ability of APs to capture stochastic global context of a temporal input, this chapter proposes a novel multimodal extension of APs for audio-visual affect recognition. The key idea of the proposed approach to audio-visual fusion is based on the intuition that global context fusion could be more effective than instance-level feature fusion for affect recognition task. The reasoning behind this intuition is as follows: fundamentally, affect information is expressed, perceived and processed at different frame rates in the audio-visual channels; this chapter argues that standard instance (feature) level fusion models are sub-optimal solutions due to the intrinsic frame-level temporal misalignment between the affect information in the audio-visual modalities. To address this problem, using APs, this chapter proposes to fuse the stochastic global latent variables of the audio and visual modalities.

Unlike the unimodal APs described earlier, audio-visual APs contains two separate (modality-specific) stochastic context encoders for the audio and visual inputs but a common predictive distribution decoder. First, the context encoders of the visual and audio modalities infer the modality-specific stochastic latent variable distributions q^v and q^a respectively. Then the latent vectors Z^v and Z^a from the distributions q^v and q^a are sampled and concatenated into a multimodal latent vector Z^{va}. The predictive distribution decoder receives the concatenated audio-visual features at frame level X^{va} paired with the multimodal latent vector Z^{va} as inputs. This chapter uses an audio-visual backbone model based on simple feature fusion for producing proxy labels for the audio-visual APs. To maximize the similarity between the global context information captured from the audio and

visual signals, additional regularisation terms are applied to the AP loss function defined in Eq. 4.2. The final loss function used for training the audio-visual APs is as follows:

$$\begin{aligned} L_{AV-AP} = \mathbb{E}(-\log p(y_i|x_t^{va}, z^{va})) + \beta * KL(q_c^v||q_t^v) \\ +\beta * KL(q_c^a||q_t^a) + \beta * KL(q_c^v||q_c^a) + \\ \beta * KL(q_t^v||q_t^a) + \lambda KL(\mathcal{N}(\mu_i^y, \sigma_i^y)||\mathcal{N}(0,1)) \end{aligned} \tag{4.3}$$

where q_c^v and q_t^v denote the visual latent distributions for the visual context and target frames respectively, and q_c^a and q_t^a denote the audio latent distributions for the audio context and target frames respectively. Additional regularisation terms $KL(q_c^v||q_c^a)$ and $KL(q_t^v||q_t^a)$ are for maximizing the similarity between audio and visual latent distributions for their corresponding context and target frames respectively.

4.4 Experiments

This section first describes the implementations of both unimodal and multimodal Affective Processes, and then it discusses their training and optimisation details. Appendix B describes (1). the datasets used for visual-only, audio-only, and audio-visual affect recognition tasks, (2). the backbone models, different temporal (GRUs and self-attention) and (3). some key multimodal fusion baselines.

Unimodal and Multimodal APs

As Fig. 4.2 shows, apart from the backbone, the key components of APs are: (a). Stochastic Context Encoder and (b). Predictive Distribution Decoder. The encoder is composed of two embedding nets (for features and proxy labels) followed by a context aggregation step and a latent distribution net. The decoder contains a predictive distribution net to output predictive distributions of target labels. Architectural details of the encoder and decoder modules of APs for visual, audio and audio-visual models are explained below.

Visual AP network architecture is shown in Fig. B.1. This model uses two 2-layer fully connected (FC) networks with 256 hidden units as the feature and label embedding nets in the encoder implementation. For context aggregation, mean pooling is applied to the concatenated feature and label embedding vectors to derive a deterministic context vector. This context vector is then fed into the latent distribution net, a 2-layer FC network. In the decoder, the predictive distribution net is implemented using a 3-layer FC network.

Audio AP network is same as that of visual AP but with two major differences (a). the mean pooling based context aggregation step in the encoder is replaced with a 1-layer GRU block with 256 hidden units and (b). a 3-layer GRU block with 256 hidden units followed by an FC output layer is used as the predictive distribution net in the decoder. Note that these two modifications are introduced to mitigate the effect of commonly encountered temporal misalignment between the audio features and their affect labels [Schmitt et al., 2019].

Audio-Visual AP network is similar to that of audio AP except that it

has two stochastic encoders, one for visual and one for audio. For context aggregation, encoders of both the modalities used two 1-layer GRUs with 256 hidden units. The predictive distribution net of the decoder is same as that of audio AP but with 512 hidden units. Additionally, audio-visual AP inference involves the multimodal fusion step. As a result, the decoder receives as input an audio-visual latent vector which is prepared by concatenating the global latent vectors that are sampled from the visual and audio context encoder output distributions individually.

4.4.1 Training of APs

Context Frame Selection. For training visual APs, the context frames are sampled from a given input sequence using uniform random sampling, in order to introduce variability into the stochastic latent distribution. Whereas for audio AP and audio-visual AP models training, consecutive frames are used for the context aggregation, but with the index of first frame in that consecutive sequence randomly sampled. The latent vector is randomly sampled from the context distribution during training phase, but for evaluation, only the mean vector of latent variable distribution is used as input to the decoder module.

The number of context frames (N_c) and number of target frames (N_t) are randomised from iteration to iteration to further increase the variability of temporal context in the training sequences, similar to [Le et al., 2018]. For training the visual APs, given that the sequence length is 70, N_t value is varied in the range [30, 70], with the N_c value range set to [3, N_t]. Whereas in the case of audio APs and audio-visual APs training, which use the sequence length of 200 frames, N_t and N_c values are varied in the ranges [50, 200] and [10, N_t] respectively. During inference, the values of

(N_c, N_t) are chosen as (30, 70) and (50, 200) for visual APs, and audio- as well as audio-visual APs respectively.

At test time, this work experimented three types of context frame selection techniques [Sanchez et al., 2021]: (a). uniform random sampling (b). frames with lowest context uncertainty and (c). frames with highest context uncertainty. Here the context uncertainty refers to L_2 norm the latent distribution's variance vector. For frame selection, this uncertainty is measured for each frame individually by passing its feature embedding and proxy label through the encoder module. Note that all the reported results with APs in this chapter use the lowest context uncertainty for frame selection.

Context Frame Labels. In APs, the labels of context frames play a key role in inferring accurate latent variable distributions, as demonstrated later in the ablation studies. During training, to introduce variability in the temporal functions learned by the stochastic latent variable, it is found that best results are achieved when the context frame labels are randomly drawn from the mixture of proxy labels (from the backbone) and the ground truth labels, with a probability of 0.5. During APs inference, only the proxy labels are used as the context frame labels, unlike in NPs which require the ground truth labels for context inference even at test time. Note that only when APs are used in the Cooperative Machine Learning settings, ground truth labels are utilised for the context frames sparsely at test time.

Optimisation Details. Unimodal and multimodal APs are trained using Adam optimizer[Kingma and Ba, 2014] to minimise the regularised ELBO loss functions in Eq. 4.2 and Eq. 4.3 respectively. The value of β is cyclically updated in the range [0, 1], with the cycle length fixed to 1 epoch. Learning rate and weight decay values are set to 1e-4 and 5e-4 respectively.

Model	**Valid. Set**	**Test Set**
ResNet-50 Backbone	(0.497, 0.440)	(0.630, 0.505)
ResNet-50+BiGRU	(0.570, 0.508)	(0.550, 0.552)
ResNet-50+Self-Atten	(0.558, 0.512)	(0.591, 0.564)
ResNet-50+APs	(0.577, 0.530)	(0.728, 0.583)
EmoFAN Backbone	(0.632, 0.609)	(0.690, 0.550)
EmoFAN+BiGRU	(0.687, 0.635)	(0.715, 0.568)
EmoFAN+Self-Atten	(0.664, 0.644)	(0.706, 0.580)
EmoFAN+APs	(**0.710**, **0.650**)	(0.739, **0.622**)
[Mitenkova et al., 2019]	-	(0.439, 0.392)
[Toisoul et al., 2021]	-	(0.650, 0.610)
[Kossaifi et al., 2020]	-	(**0.750**, 0.520)

Table 4.1: Visual-only affect recognition results (valence CCC ↑, arousal CCC ↑) **on the SEWA test set**

Here also, Cosine annealing in combination with the warm restarts is used for tuning the initial learning rate.

4.5 Results and Analysis

This section first analyses the dimensional affect recognition results of unimodal (visual-only and audio-only) and multimodal APs, in comparison with existing benchmarks and baselines on their respective datasets. Then it proceeds to discuss the ablation experiments of APs that verify the contribution of stochastic latent variable to the overall APs' performance under different configurations at test time.

4.5.1 Performance of Unimodal and Multimodal APs

Visual-only Dimensional Affect Recognition. As shown in Table 4.1, on SEWA, visual APs trained using both the backbones (ResNet-50 and EmoFAN) outperformed their corresponding deterministic temporal models (BiGRUs and self-attention) by significant margins. Table 4.1 also compares different models with existing benchmarks on SEWA. APs achieved better performance than the prior state-of-the-art [Kossaifi et al., 2020] for arousal, but slightly worse results on valence. The results of visual-only models on the AVEC'19 CES dataset, as Table 4.2 shows, follow the similar pattern–APs outperformed the deterministic regression baseline (the backbone model) as well as the AVEC'19 CES challenge winners [Zhao et al., 2019], particularly in the case of valence. However, in the case of arousal, APs performed slightly worse compared to the challenge winners' methodology [Zhao et al., 2019], which was based on an adversarial domain adaptation technique to cope with the cross-cultural variations in the affect data. In principle, APs can be complemented with such adaptation techniques to further boost their performance.

Overall, the results of visual-only models evaluated on both SEWA and AVEC'19 CES datasets validate the main hypothesis of APs that learning a distribution of temporal functions generalises much better than the deterministic function learning, implicitly accounting for the label ambiguity problem of affect annotations.

Audio-only Dimensional Affect Recognition. The results of audio-only models presented in Table 4.2 show that APs have superior generalisation performance, than the deterministic regression baseline (VGGish + BiGRU backbone) and the existing state-of-the-art model [Zhao et al., 2019]. Unlike in the visual-only case, APs considerably improved the af-

	Model	Valid. Set
	Aud-only AVEC Winners [Zhao et al., 2019]	(0.388, 0.518)
	Aud-only Backbone	(0.414, 0.546)
	Aud AP	(**0.458**, **0.592**)
$A\|\|V$		
	Vis-only AVEC Winners [Zhao et al., 2019]	(0.579, **0.594**)
	Vis-only Backbone	(0.527, 0.564)
	Vis AP	(**0.589**, 0.586)
	AVEC Winners [Zhao et al., 2019]	(0.614, 0.645)
	Uniformly Weighted Feature Fusion†	(0.597, 0.583)
	Globally Weighted Feature Fusion†	(0.598, 0.614)
$A\&V$	Locally Weighted Feature Fusion†	(0.583, 0.628)
	Crossmodal Self-Atten Fusion [Tsai et al., 2019]†	(0.602, 0.619)
	COLD Fusion	(0.611, **0.661**)
	AP-Uniformly Weighted Feature Fusion	(0.637, 0.623)
	AP-Global Context Fusion	(0.648, 0.631)
	AP-Global Context Fus.+$KL(q^v\|\|q^a)$	(**0.662**, 0.650)

Table 4.2: Audio-visual affect recognition results (valence CCC ↑, arousal CCC ↑) on AVEC'19 CES Validation Set. † denotes in-house implementations of different fusion baselines ($A||V$ denotes unimodal and $A\&V$ denotes multimodal).

fect recognition performance in terms of both the dimensions (valence and arousal). These results establish the modality-agnostic nature of APs' effectiveness in improving the emotion recognition models, through stochastic process regression.

Audio-Visual Dimensional Affect Recognition. As Table 4.2 shows, stochastic global context fusion based on audio-visual APs outperformed the standard feature fusion (both uniformly and non-uniformly weighted) baselines, crossmodal self-attention based fusion, and the COLD fusion models in terms of mean CCC values. Performance difference between APs based on the concatenated audio-visual features (uniformly weighted

feature fusion) and APs based on the global context fusion, validates the main hypothesis of multimodal APs that instance (feature) level fusion is a sub-optimal solution compared to the global context fusion for multimodal affect recognition. The results of global context fusion in APs improved when the constraint of audio-visual context similarity maximization ($KL(q^v||q^a)$) is applied (Eq. 4.3). This model achieved state-of-the-art results on the AVEC'19 CES corpus by outperforming the challenge winners [Zhao et al., 2019].

Similar to APs, crossmodal self-attention fusion [Tsai et al., 2019] captures long-range temporal context and addresses temporal misalignment between different modalities. However, as Table 4.2 shows, the performance of crossmodal self-attention is found to be inferior to that of AV-APs, which could be due to its fundamentally deterministic function learning nature, unlike the stochastic context modelling feature of APs.

Qualitative Analysis. Fig. 4.3 illustrates modality-wise temporal latent uncertainty patterns inferred using AV-APs on two AVEC'19 validation set examples. In AV-APs, the uncertainty measure for a modality refers to L2 norm of its corresponding latent distribution's variance vector. This analysis shows that when the valence is high, visual modality seems to have lower latent uncertainty than the audio; in the arousal case this pattern seems to be almost inverse. This observation is in line with the findings reported in prior works [Ringeval et al., 2019] regarding the informativeness of visual and audio modalities w.r.t valence and arousal inference.

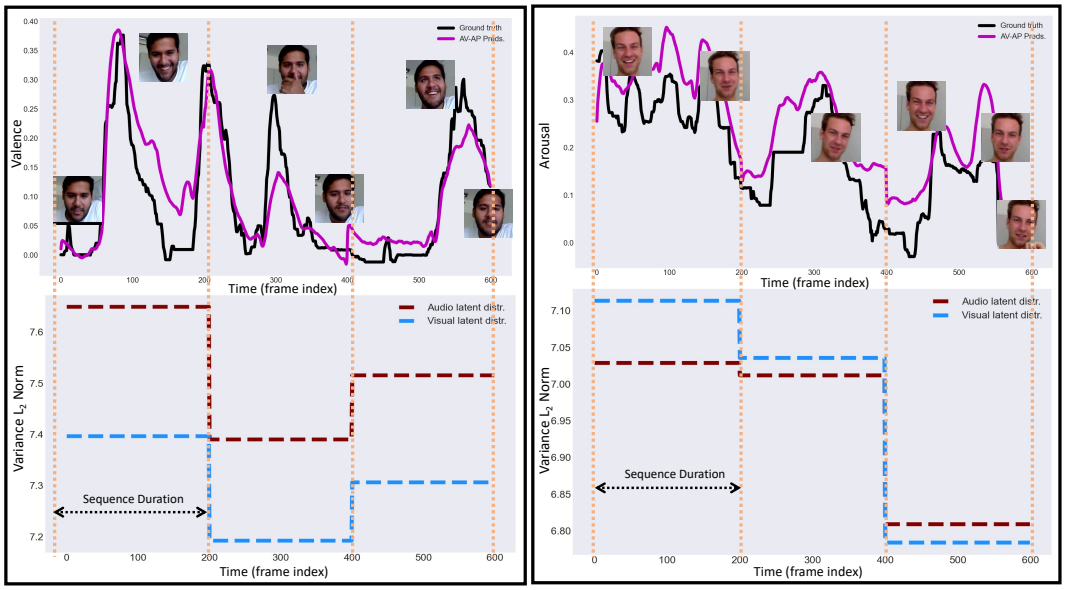


Figure 4.3: Latent uncertainty patterns in audio-visual affect (valence and arousal) recognition using AV-APs on AVEC'19 validation set: For the visual and audio latent distributions inferred in AV-APs, this work computed their variance vectors' L_2 norm values and consider them as modality-wise uncertainty measures. Here, all the frames in an input sequence segment (marked as "Sequence Duration" above) have a global uncertainty value due to the underlying global latent distribution modelling in APs. In this example, when the valence is high the visual modality has lower latent uncertainty than the audio modality, and it is almost vice-versa in the case of arousal – matching with similar observations mentioned in [Ringeval et al., 2019].

AP Objective Function	(Valence, Arousal)
Unregularised ELBO	(0.685, 0.626)
ELBO+Beta Scheduling	(0.691, 0.638)
ELBO+KL Divergence Reg.	(0.697, 0.644)
ELBO+Beta Sched.+KL Div. Reg.	(**0.710**, **0.650**)

Table 4.3: Visual AP results (CCC ↑) on SEWA validation set with different loss functions (using EmoFAN backbone).

4.5.2 Ablation Studies on Affective Processes

To delineate the impact of some important design choices involved in APs' training and inference, the following ablation experiments are conducted. For this experimental analysis, the visual APs based on EmoFAN backbone are evaluated under different training and inference conditions, on SEWA validation dataset.

APs Loss Function. Table 4.3 compares the performance of visual AP models trained with different loss functions. Unregularised ELBO loss function (Eq 4.2) clearly exhibited poorer performance than the remaining regularised AP loss functions. Improved generalisation performance with Beta scheduling and KL divergence regularisation (Eq 4.2) validates the hypothesis of this chapter that training of APs using unregularised training objective is vulnerable to encoder posterior collapse and decoder predictive variance collapse problems.

Number of Context Points and Context Frame Sampling Methods. Fig. 4.5 presents visual APs performance for different number of context points with the number of target points set to a fixed value (70), at test time. Further, it compares the performance of three different context frame sampling techniques: (a). uniform random sampling, (b). lowest encoder-sigma criteria and (c). highest encoder-sigma criteria. The encoder-sigma

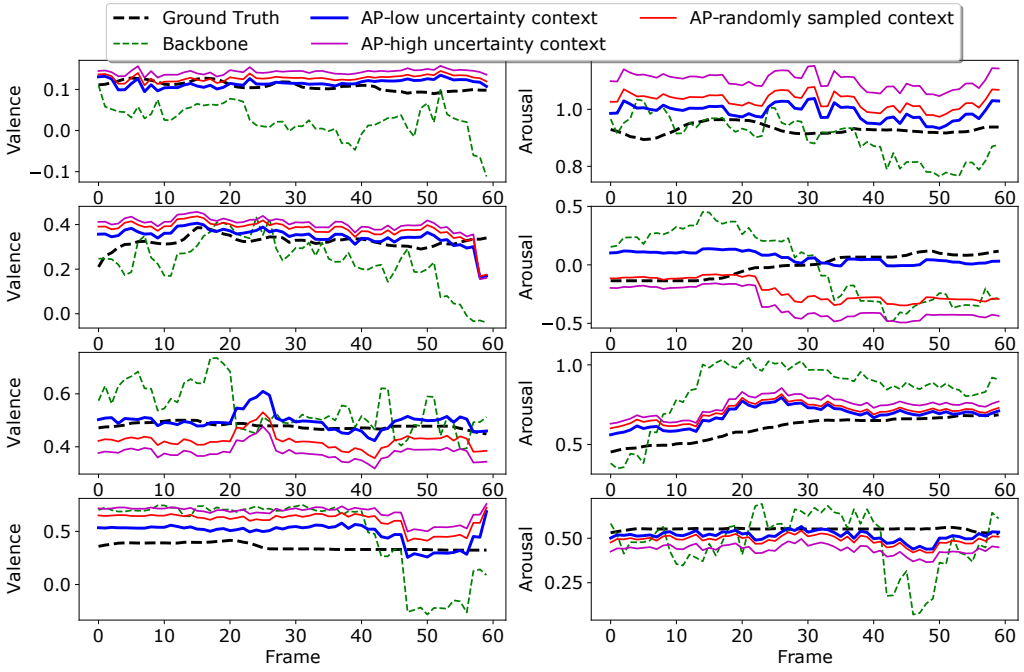


Figure 4.4: Qualitative results of visual APs on SEWA validation set, with *different context frame selection methods applied.*

here refers to L2 norm of the variance vector of AP encoder's output latent distribution. For each frame in the input sequence, the encoder-sigma value is computed by passing that frame's feature embedding and its proxy label through the AP encoder. The results presented in Fig. 4.5 show that the lowest encoder-sigma criteria achieved much better performance than the uniform random sampling and highest encoder-sigma criteria. The qualitative results presented in Fig. 4.4 also indicate the same trends in both valence and arousal cases. Note that the highest encoder-sigma criteria results are worse than the backbone results when the number of context points is small, which indicates that the AP encoder is capable of characterising each input frame and its proxy label in terms of the temporal context uncertainty associated with it.

Context based priors Vs Random-valued Priors. APs essentially infer data-driven priors over the temporal functions through a stochastic latent variable, from the context points. The impact of latent prior quality

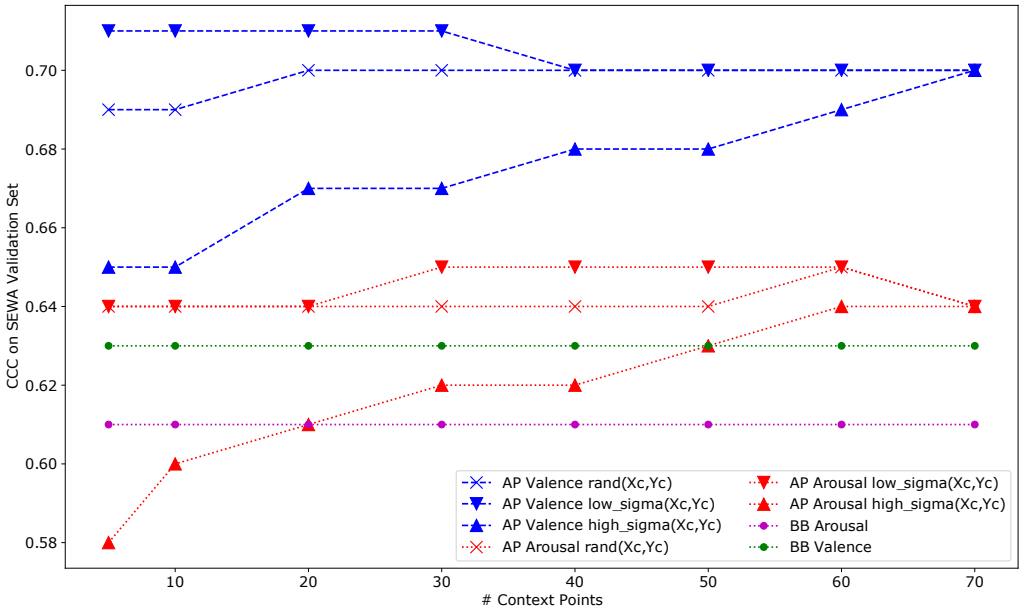


Figure 4.5: Visual AP results (CCC ↑) on SEWA validation set using *different number of context points (N_c)* (with the number of target points (N_t) fixed to 70) with different context frame selection techniques: **rand**(X_c,Y_c) — uniform random sampling, **low sigma**(X_c,Y_c) and **high sigma**(X_c,Y_c) — frame selection based on the lowest and highest AP encoder variance L_2 norm values criteria respectively.

	Z^{rand}	$Z(X_c, Y_c^{rand})$	$Z(X_c, Y_c^{bb})$	$Z(X_c, Y_c^{gt})$
AP	(0.29, 0.16)	(0.54, 0.44)	(0.71, 0.65)	(0.92, 0.93)

Table 4.4: Visual AP results (valence CCC ↑, arousal CCC ↑) on SEWA validation set using different priors during inference: 1.random-valued latent vector Z^{rand}, latent vector Z as a function input features X_c and 2. random-valued context labels Y_c^{rand}, 3. proxy context labels Y_c^{bb}, and 4. ground truth context labels Y_c^{gt}.

on APs performance is investigated in Table 4.4. By setting the number of context and target points to 30 and 70 respectively, visual APs are evaluated using different priors over latent functions at test time: 1. a random-valued latent vector Z^{rand}, and the latent vector as a function input features X_c and 2. proxy context labels Y_c^{bb}—$Z(X_c, Y_c^{rand})$, 3. random-valued context labels Y_c^{rand}—$Z(X_c, Y_c^{bb})$, and 4. ground truth context labels Y_c^{gt}—$Z(X_c, Y_c^{gt})$. Here, a 'random-valued' vector refers to a vector drawn from a multi-variate normal distribution with zero mean and unit variance. As Table 4.4 shows, APs performance heavily relies on the quality of stochastic latent variable, and the highest quality is obtained when ground truth labels are used as the context frame labels.

4.6 Application: Cooperative Machine Learning for Label-Efficient Affect Recognition

Towards addressing the label scarcity problem of affect recognition tasks, this chapter proposes a novel use case of Affective Processes to propagate sparse human supervision to unlabelled data points, based on the following observation: APs can easily tune their temporal regression function at test time without requiring model retraining or fine tuning. Note that in APs the function distribution inferred in the form of stochastic latent variable is a function of the context frame features (X_c) and their labels (Y_c). Thus, in the form of context frame labels (Y_c) APs can easily accommodate human supervision at test time. The more accurate the context frame labels are, the closer the inferred function distribution to the true underlying function. These properties enable APs to improve the accuracy of their predictions by

using very little human supervision, most importantly without requiring the costly model retraining step as is the case with the standard deterministic models. Building on these unique properties of APs, this chapter proposes to use them as Cooperative Machine Learning models for the dimensional affect recognition task.

Cooperative Machine Learning (CML) aims at effectively combining sparse human supervision with an already trained model's predictions to annotate unlabelled data. CML is applied to affect recognition tasks in [Wagner et al., 2018] in which active learning is used to identify the frames to be passed to human annotators and a costly model retraining step to update the pretrained model weights. The method proposed in this work simplifies this approach to CML by circumventing the active learning and model retraining steps, by leveraging the flexibility of APs.

To use APs as CML models, the method here uses ground truth labels in the place of proxy labels from the backbone, as the context frames labels (Y_c), during both training and inference. However, at inference time, APs need a very few context frames to infer the function distribution that produces accurate predictions for the target labels as showed later in the experimental results. Thus, APs as CML models require very sparse human supervision for randomly sampled context frames at test time, and they circumvent the active learning and model retraining steps, but still achieve superior performance on the unlabelled frames.

Fig. 4.6 illustrates the performance of visual APs on SEWA validation set, as a Cooperative Machine Learning model. Here the goal is to verify how well APs can utilise sparse human supervision in the form of ground truth labels for the context frames at test time. It is import to note that here *only one randomly sampled context frame per input sequence is used* and

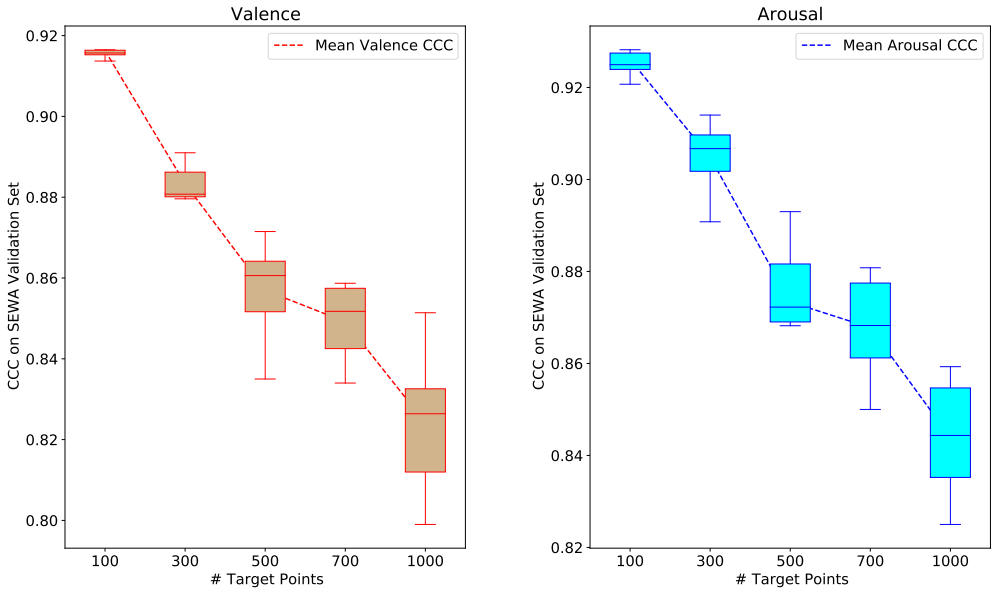


Figure 4.6: Visual AP results (CCC ↑) on SEWA validation set with *ground truth labels* used as context frame labels, with ***only one context frame*** (N_c) for different number of target frames (points) (N_t), evaluated over 20 runs with 20 different random seeds (the same set of random seeds is used for all the evaluations with different N_t values).

the length of input sequences (number of target points) is varied in the range of 100 to 1000. It shows the mean and variance values of visual APs collected over 20 different runs with each run using a different random seed. When the input sequence length is 100, APs achieved the best performance overall (0.915 mean valence CCC and 0.932 mean arousal CCC) *with just one randomly sampled frame manually annotated*. Even when the input sequence length is increased to 1000, APs still achieved reasonably good performance, 0.825 mean CCC for valence and 0.844 mean CCC for arousal, with just one frame out of 1000 provided with human supervision. These results show the potential of APs as Cooperative Machine Learning models that do not need active learning and costly model retraining steps, to propagate sparse human supervision to unlabelled data points.

4.7 Conclusion

This chapter presented Affective Processes, a novel stochastic temporal context modelling framework designed for affect recognition tasks. The experimental results showed that Affective Processes are capable of addressing some fundamental challenges encountered in affect recognition using deterministic temporal function learning, which fails to: a. account for the inherently stochastic nature of affect expression and perception processes, b. cope with modality-specific stochasticity when integrating affect information from multiple modalities, and c. make the manual affect annotation process less laborious by effectively leveraging sparse human supervision. The solutions proposed in this chapter to the aforementioned challenges using Affective Processes demonstrated consistent performance gains and promising results on in-the-wild challenging datasets of different unimodal and multimodal affect recognition.

In summary, towards the objective of overcoming deterministic function learning models' limitations, the main technical contributions of this chapter are three fold. 1. Firstly, it is identified that the training of Neural Processes is prone to distributional collapse problems, which are addressed by including additional regularisation functions in the training objective of Affective Processes. 2. Then, with some architectural changes, the modality-agnostic nature of Affective Processes is showed by applying them to both audio and audio-visual affect recognition tasks. By building on the ability of Affective Processes to capture the global stochastic temporal context, this chapter proposed a novel audio-visual fusion technique for multimodal affect recognition. Most importantly, compared to the COLD fusion technique proposed in Chapter 3, the global context fusion technique implemented in the audio-visual APs has shown superior

generalisation performance. 3. Finally, this chapter demonstrated a novel application of Affective Processes to the label propagation task in Cooperative Learning models, with the potential to significantly speed up the laborious task of affect label annotation, with minimal human intervention. Thus, the solutions proposed in this chapter attempted to address the two fundamental affect recognition challenges that this thesis is concerned with: label ambiguity and label scarcity.

Chapter 5

A Holistic Uncertainty Model of Temporal Affect and Its Application to Personality Recognition

> Information is the resolution of uncertainty
>
> *Claude Shannon*

Chapter Summary. Temporal uncertainty modelling methods such as the ones proposed in Chapter 3 and Chapter 4, are based on introducing non-deterministic function learning properties into a specific intermediate layer or variable of the temporal networks. In the COLD approach proposed in Chapter 3, this intermediate layer is chosen as the final hidden state vector of an RNN model. Similarly in the Affective Processes introduced in Chapter 4, the output of the encoder module, referred to as the global

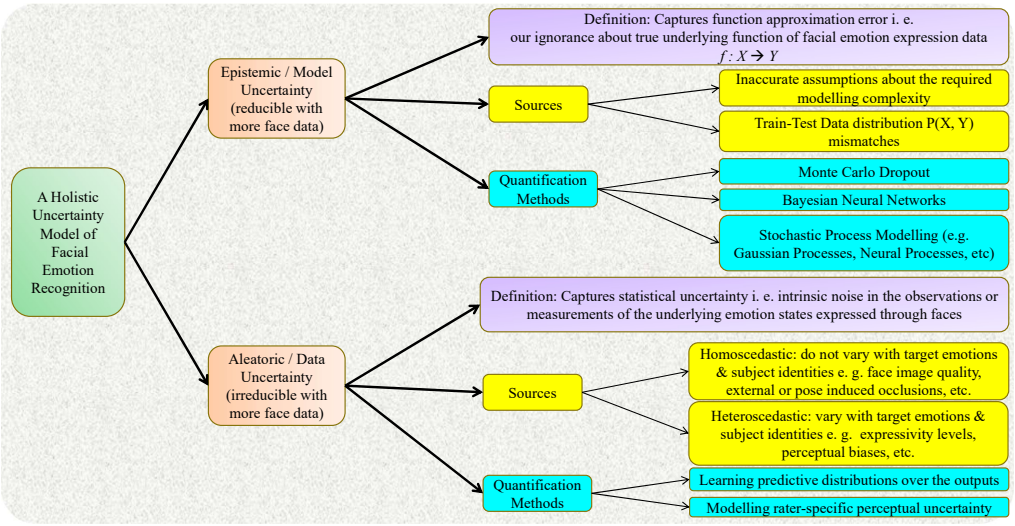


Figure 5.1: Modelling holistic uncertainty of dimensional emotion recognition from face images, using Epistemic and Aleatoric categorisation (X – a face image sequence, Y – its corresponding ground truth emotion label sequence, f – true underlying mapping function between X and Y, and $P(X, Y)$ – joint probability distribution of X and Y)

stochastic latent variable, is used as a proxy for capturing the temporal uncertainty. While both these approaches demonstrated good performance gains in affect recognition in both unimodal and multimodal settings, they are not designed to capture the temporal uncertainty in a holistic manner. Here, the concept of 'holistic temporal uncertainty modelling' refers to learning the uncertainty information based on all the network parameters in a temporal model, as well as the uncertainty induced by input data into the model. This chapter introduces an alternative temporal uncertainty model that aims to holistically capture the uncertainty of temporal functions, without confining the uncertainty modelling to any particular intermediate latent layer or latent variable of a temporal model.

This chapter aims to learn uncertainty modelling for equipping video-based

expressive behaviour recognition models with the abilities to (a). quantify the predictive uncertainty estimates as completely as possible and (b). propagate those estimates to the benefit of downstream behavioural analysis tasks. Towards this objective, this chapter first quantifies uncertainties in dimensional emotion recognition from face videos, by adopting the framework of epistemic (model) and aleatoric (data) uncertainty categorisation. Then for evaluating the practical utility of uncertainty-aware emotion predictions, we propose to introduce emotion uncertainty estimates in learning an important downstream task, apparent personality recognition.

5.1 Introduction

Creation of large-scale emotion labelled datasets [Kossaifi et al., 2019, Kollias and Zafeiriou, 2018a, Kollias et al., 2020] and rapid progress in Deep Neural Networks for processing face data [Toisoul et al., 2021, Kossaifi et al., 2020, Kollias et al., 2019] in recent years, enabled facial emotion recognition models to achieve impressive generalisation performance. However, to effectively utilise these advancements in emotion recognition to benefit downstream behavioural analysis tasks, it is important to know how confident the models are about their predicted emotions. Hence, it is important to equip the facial emotion recognition models with the ability to quantify the uncertainties associated with their output point estimates. This chapter asks two important questions: (1). how to holistically quantify uncertainties in facial emotion recognition? and (2). how to evaluate the quantified uncertainties by introducing them in learning a downstream affective computing task?

Here, the objective is to quantify uncertainties in time- and value-continuous

dimensional emotion (valence and arousal) recognition from face videos. Typical sources of uncertainty in machine learning models [Hüllermeier and Waegeman, 2021] include data distribution mismatches from training to deployment, approximation errors in mapping functions etc. In addition to such general sources, a wide range of ambiguous elements are involved in learning emotion recognition models: subjects' emotional expressiveness levels, their personality type, the nature of emotion inducing stimuli, human annotators' perceptual biases due to socio-cultural differences [Kossaifi et al., 2019, Han et al., 2017, 2021], etc. Therefore, the proposed method argues that in the case of emotion recognition it is impractical to identify all possible uncertainty sources and quantify them individually, due to the intrinsic subjective nature of apparent emotions. To cope with this problem, this chapter adopts a holistic uncertainty modelling framework [Hüllermeier and Waegeman, 2021, Kendall and Gal, 2017a] with two broad categories: epistemic (model) uncertainty and aleatoric (data) uncertainty, as illustrated in Fig. 5.1. Epistemic uncertainty captures function modelling errors which are due to our ignorance of the true underlying mapping function between facial images and their emotion labels, and it is reducible with more training data. Whereas aleatoric uncertainty is irreducible with more data, as it captures the statistical noise inherent to the labelled data collection, for instance, ambiguities in facial emotion perception by human raters [Ghandeharioun et al., 2019].

The method introduced in this chapter quantifies epistemic and aleatoric uncertainties in canonical CNN+GRU models trained for video-based emotion recognition, using Monte Carlo dropout [Gal and Ghahramani, 2016] and predictive distribution modelling [Kendall and Gal, 2017a] techniques respectively. Having quantified the emotion recognition uncertainties, the focus then shifted to the second question: how to evaluate the usefulness

of predicted uncertainty estimates of emotions? To this end, this chapter proposes an evaluation protocol based on how well the uncertainty estimates represent the reliability of predicted emotions in learning a downstream task that heavily relies on emotion information [Zhang et al., 2019]. Thus, the practical utility of emotion predictions' uncertainty estimates is evaluated by using them as input features in an important video-based behavioural analysis task, apparent personality traits estimation or personality recognition from face videos.

For estimating the personality traits from a face video, this chapter proposes to use a conditional latent variable model (CLVM) that builds on a recently proposed global context aggregation method based on neural latent variable models [Garnelo et al., 2018a,b, Sanchez et al., 2021, Tellamekala et al., 2021], to derive personality-related information from a sequence of image embeddings and their corresponding dimensional emotions (valence and arousal). Here, it is assumed that apparent personality traits inferred from a face video can be viewed as temporal aggregate functions of per-frame emotional expressions, motivated by the emotion-to-apparent-personality relationship discussed in [Zhang et al., 2019]. Based on this assumption, this chapter hypothesises that it is more effective to learn a global latent variable that summarises the personality information from face image features and dimensional emotion predictions. Based on this premise, the quality of uncertainty-aware dimensional emotion predictions is assessed by training and evaluating different CLVMs with and without uncertainty estimates as inputs.

This chapter presents the results of extensive experiments on two large-scale in-the-wild databases; SEWA [Kossaifi et al., 2019] for dimensional emotion recognition and ChaLearn [Ponce-López et al., 2016, Escalera et al., 2017] for personality recognition. First, SEWA dataset is used for training

uncertainty-unaware and uncertainty-aware (both epistemic and aleatoric) emotion recognition models. Then on the ChaLearn corpus, the quality of emotion predictions and their uncertainty estimates is evaluated in terms of their impact on personality recognition performance. The proposed CLVM, even without using uncertainty estimates, achieved state-of-the-art results, outperforming existing personality recognition methods. When emotion uncertainty estimates are included as additional inputs, personality recognition performance significantly improved further, compared to the models based on point estimates of emotion. Particularly, fusing epistemic and aleatoric uncertainties achieved best results with a substantial performance improvement, $\sim$42% better performance (in terms of mean Pearson's correlation coefficient) than the existing state-of-the-art method [Song et al., 2021] on ChaLearn.

In summary, the contributions made in this chapter are:

- The proposed method quantifies predictive uncertainties in dimensional emotion recognition from face videos, using epistemic and aleatoric uncertainty categorisation.
- An evaluation protocol is introduced for assessing the quantified uncertainties of emotions by propagating them to a downstream behavioural analysis task, apparent personality traits estimation from face videos.
- In personality recognition, to leverage the already predicted uncertainty-aware dimensional emotions as additional inputs, this chapter proposes to use a global latent variable model that builds on [Garnelo et al., 2018a].
- The proposed method achieves new state-of-the-art results on in-the-

wild personality traits estimation, outperforming the existing methods by significant margins.

5.2 Related Work

This section presents a brief survey of existing works on uncertainty modelling in various Computer Vision and Affective Computing tasks. For a general introduction to uncertainty modelling, the reader can refer to Eyke et al. [Hüllermeier and Waegeman, 2021] which offers a thorough treatment of techniques for measuring and evaluating the uncertainties. For a detailed account of ambiguities or uncertainty sources that are inherent to models of emotion representation models (discrete, continuous, ordinal, etc), which are not covered in this chapter, the reader is recommended to refer to [Sethu et al., 2019]. As the main focus here is on holistic uncertainty modelling in affect recognition, this section does not discuss the literature of apparent personality recognition from visual information in detail; refer to [Escalante et al., 2020] for the most recent comprehensive review on this topic.

5.2.1 Epistemic and Aleatoric Uncertainty Modelling in Computer Vision

As Deep Neural Networks have been pushing prediction accuracies towards near-perfect levels in Computer Vision, robustness and reliability aspects of these models started receiving wide spread attention recently. Most notably, Kendall et al. [Kendall and Gal, 2017a] systematically deconstruct the holistic uncertainty in Computer Vision tasks by adopting epistemic

and aleatoric uncertainty categorisation framework, which inspired the work presented in this chapter. Similar to [Kendall and Gal, 2017a], several other works focused on investigating various task-specific sources of uncertainties in some general tasks like image classification [Peterson et al., 2019, Khan et al., 2019], object detection [He et al., 2019], multi-view representation learning [Geng et al., 2021], semantic segmentation [Hu et al., 2020], depth estimation [Eldesokey et al., 2020] etc.

Task-Specific Uncertainty. In image classification, Peterson et al. [Peterson et al., 2019] utilise human perceptual ambiguities in the form of label distributions to model the uncertainty. Sample and class uncertainties in image classification tasks are modelled to quantify the sample-rarity and class imbalance properties in [Khan et al., 2019]. To capture the label uncertainty in object detection tasks, uncertainty of bounding box regression is modelled in [He et al., 2019] by factoring in the ambiguities involved in bounding box annotation process. Dynamic uncertainty aware networks are proposed in [Geng et al., 2021] for multi-view image representation learning for uncertainty-aware fusion of information present in different views. A novel Bayesian uncertainty estimation method proposed in [Hu et al., 2020] is applied to semantic segmentation task, which models the network outputs with Gaussian and Laplacian distributions. In [Eldesokey et al., 2020] a framework is proposed to learn the uncertainties present in depth completion tasks. By combining the CNN models with Gaussian Processes (GPs) [Rasmussen, 2003], scalable models for aleatoric uncertainty quantification are explored in [Carvalho et al., 2020].

5.2.2 Epistemic and Aleatoric Uncertainty Modelling in Affective Computing

Data Uncertainty. [Kim and Kim, 2018] use a multi-label representation learning of discrete emotions from audio-visual data, to model human like labelling errors due to perceptual biases. Joint learning of hard and soft discrete emotion labels is used in [Chou and Lee, 2019] to capture the label uncertainty and rater-specific biases explicitly. Dang et al. [Dang et al., 2018] model continuous-time emotion uncertainty from speech by combining multi-rater Gaussian mixture regression with Kalman filters that capture temporal dependencies in the annotation signals. Similarly, inter-rater variability or disagreement scores of the dimensional emotions are used to define explicit targets for perceptual uncertainty modelling in [Han et al., 2017, 2021].

Model Uncertainty. Bayesian Neural Networks [Ebrahimi et al., 2019, Sun et al., 2017, Kwon et al., 2020] are used in [Rizos and Schuller, 2019, 2020] to quantify sample informativeness in dimensional emotion recognition model training, so that it is possible to regularise the impact of less informative samples on the model predictions. [Sridhar and Busso, 2020] employ a teacher-student ensemble model for uncertainty-aware speech emotion recognition, in which the teacher model passes a probabilistic embedding (generated using Monte Carlo dropout [Gal and Ghahramani, 2016]) as a guiding signal for the ensemble of student models. In [She et al., 2021], ambiguities in the facial expression recognition task are modelled using (a). latent distribution mining and (b). pairwise uncertainty estimation. The former technique derives a latent distribution in the label space using multi-branch learning and the latter exploits the semantic feature similarity between pairs of instances.

In discrete facial expression recognition, Ghandeharioun et al. [Ghandeharioun et al., 2019] model both the data and model uncertainties in facial expressions in order to (a). use the predicted uncertainties as proxies for model calibration and (b). disentangle the biases of training data distribution and label annotation. These uncertainties are quantified for studying the model calibration and interpretability aspects of facial expression recognition. An interesting observation reported in [Ghandeharioun et al., 2019] is that inter-rater disagreement scores of discrete facial expressions are highly correlated with the data uncertainties, with a correlation coefficient of 0.3. This result is one of the motivations for this work to model aleatoric uncertainty in continuous dimensional emotion recognition which may also implicitly capture the dimensional emotion labelling uncertainty. Further, it is worth noting that the other uncertainty modelling methods introduced in this thesis, COLD and APs, completely ignore the aleatoric component of the temporal uncertainty.

In contrast to the aforementioned efforts, the work presented in this chapter provides a more holistic perspective of uncertainties in video-based facial affect recognition, by quantifying and fusing the epistemic and aleatoric components of valence and arousal. Furthermore, an alternative is proposed to the existing evaluation protocols for uncertainty estimates (improved performance on the emotion prediction itself or model calibration or interpretability etc) by showing how to effectively propagate the emotion predictions to downstream behavioural analysis tasks.

5.3 Method

In video-based dimensional emotion recognition, given a sequence of face images $X : [x_1, x_2, ..., x_T] \in \mathcal{X}$ as input, the objective is to predict a sequence of continuous emotion labels $Y : [y_1, y_2, ..., y_T] \in \mathcal{Y}$ where y_t is composed of two dimensions: valence (degree of pleasantness) and arousal (degree of activeness). Considering that the underlying mapping function $f^* : \mathcal{X} \longrightarrow \mathcal{Y}$, and its complexity $\gamma(f^*)$ are not known, assuming some modelling complexity $\gamma(f^o)$, the aim is to learn an approximate function f^o from the given training data $\{X^i, Y^i\}_{i=1}^N$.

General Sources of Uncertainty. Two primary sources of uncertainty in data-driven model learning are based on: how well the training data $\{X^i, Y^i\}_{i=1}^N$ represents the entire sample space $(\mathcal{X}, \mathcal{Y})$ and how close the approximated function f^o is to the true underlying function f^*. The former uncertainty source impacts the model's performance when out-of-distribution samples are encountered at test time, whereas the latter uncertainty source is due to the gap between assumed function complexity $\gamma(f^o)$ and true underlying function complexity $\gamma(f^*)$.

This chapter argues that it is not practical to identify and explicitly quantify all possible sources of uncertainty involved in facial emotion recognition model training. As an alternative, it adopts a holistic uncertainty modelling approach in which two broad categories are defined: epistemic (model) uncertainty and aleatoric (data) uncertainty [Hüllermeier and Waegeman, 2021, Kendall and Gal, 2017a,b]. Fig. 5.1 shows a detailed decomposition of the epistemic and aleatoric uncertainties specific to facial emotion recognition.

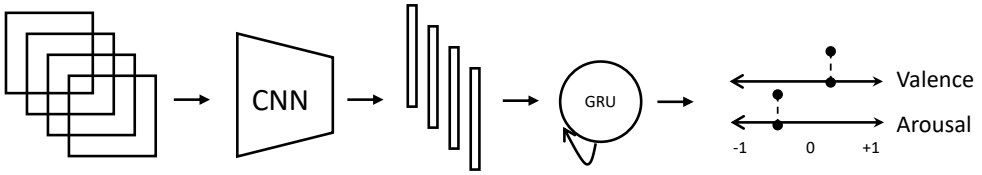


Figure 5.2: CNN+GRU baseline: Dimensional affect recognition

5.3.1 Epistemic and Aleatoric Categorisation of Uncertainties in Dimensional Emotion Recognition

Epistemic uncertainty entails the uncertainties that stem from our assumptions about the underlying function complexity, and data-distribution mismatches from train time to test time, same as the aforementioned general sources of uncertainty. By providing more training data, epistemic uncertainty can be reduced to null, at least in theory. However, in practice, given the high degree of variability in in-the-wild facial emotion data, it may not be possible to fully avoid the data-distribution mismatches and accurately capture the true underlying mapping function. To minimise the function approximation error and quantify epistemic uncertainty, rather than learning a single mapping functions, an ensemble of functions (e.g. [Lakshminarayanan et al., 2017]) or a distribution over functions (e.g. [Mitros and Mac Namee, 2019]) can be learned from the training data.

Monte Carlo (MC) Dropout [Gal and Ghahramani, 2016], or stochastic forward passes, is one of the simple yet effective techniques to model the epistemic uncertainty in large-scale DNNs [Kendall and Gal, 2017a]. It performs dropout based variational inference by leveraging the idea that using dropout layers in a DNN is equivalent to learning an ensemble of models. Thus, it performs dropout operations at test time as well in order to draw multiple predictions for a single input. The reader is recommended to refer to [Gal and Ghahramani, 2015] for formal details of dropout vari-

ational inference. Other approaches to epistemic uncertainty modelling include Bayesian Neural Networks (BNNs) [Ebrahimi et al., 2019, Sun et al., 2017, Kwon et al., 2020] that learn distributions over weights (function parameters) explicitly, and Gaussian Processes (GPs) [Rasmussen, 2003] that instead learn distributions over the function space directly.

The proposed method chooses MC dropout method [Gal and Ghahramani, 2016] to quantify epistemic uncertainties of a dimensional emotion recognition baseline model, a CNN (feature extraction) + GRU (temporal regression) network, as Fig. 5.3 shows. Running multiple forward passes through the entire CNN+GRU network during inference phase is computationally intensive, particularly for video inputs, so the implementation used in this chapter chooses to use dropout layers in the temporal regression model alone for epistemic uncertainty modelling. Thus, the computationally expensive feature extraction step is run only once, whereas the inexpensive temporal regression step is run multiple times at test time with dropout configurations varying from run to run.

Aleatoric uncertainty captures statistical noisy factors inherently present in the labelled data generation processes (e.g. ambiguities in the emotion expression and perception [Sethu et al., 2019]). This chapter argues that intrinsically subjective nature of emotion recognition task significantly contributes to the aleatoric component of the total uncertainty [Ghandeharioun et al., 2019], unlike in other Computer Vision tasks such as image classification and object detection that mostly deal with objective labels. Aleatoric uncertainties can be further categorised into homoscedastic and heteroscedastic sources (see Fig. 5.1). The former category contains uncertainties due to factors like image quality, occlusions induced by head pose variations, etc, that are invariant to the underlying emotion information. Heteroscedastic sources entail the ambiguous factors that vary for different

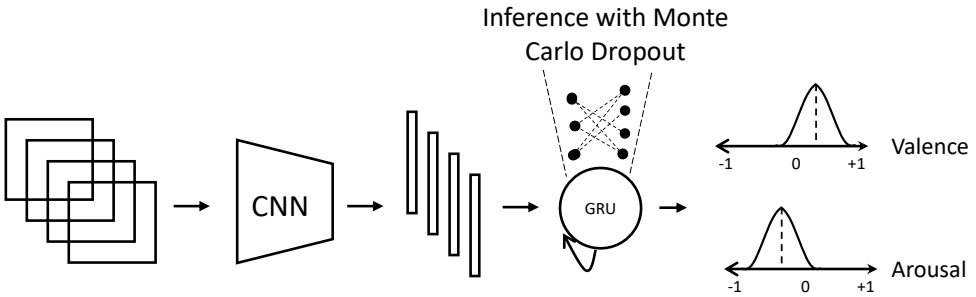


Figure 5.3: Epistemic uncertainty modelling of dimensional emotion recognition using Monte Carlo dropout [Gal and Ghahramani, 2016] inference

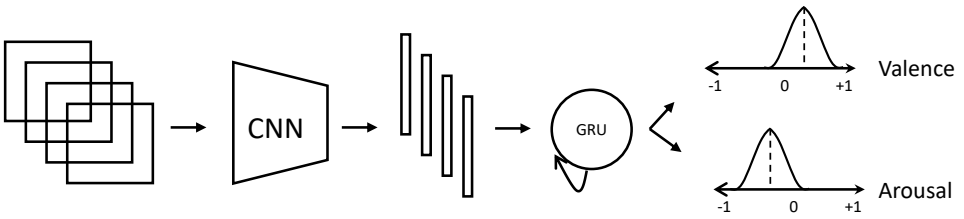


Figure 5.4: Aleatoric uncertainty modelling of dimensional emotion recognition using predictive distribution learning [Kendall and Gal, 2017a]

emotions, for instance personality types and socio-cultural backgrounds of different groups of subjects and raters, etc. This chapter focuses on capturing the heteroscedastic component of emotion aleatoric uncertainty.

Predictive Distribution Modelling [Kendall and Gal, 2017a] A prominent approach to quantify aleatoric uncertainty, in particular its heteroscedastic component, is based on training models that directly predict the distribution parameters (mean and variance) of target labels (valence and arousal), rather than predicting the point estimates. Another potential approach to estimate one specific component of aleatoric uncertainty could be based on perceptual uncertainty modelling using inter-rater disagreement labels [Han et al., 2017, 2021]. However, such methods need additional target labels and the quality of their uncertainty estimates relies heavily on the size and composition of the pool of human raters.

Note that training the predictive distribution models does not necessarily

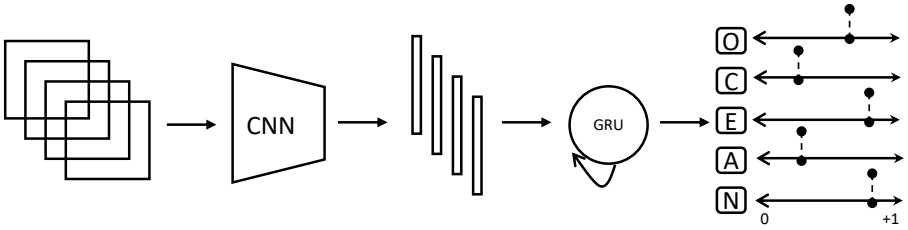


Figure 5.5: CNN+GRU baseline: Personality traits estimation

require the true distributions of ground truth labels as targets. As demonstrated in [Kendall and Gal, 2017a], negative log likelihood loss (Eq. 5.1) can be used to train the predictive distribution models. Here, as Fig. 5.4 shows, point estimates of the ground truth valence and arousal labels are used to train an emotion recognition model that can predict the parameterised distributions.

$$Loss_{AU} = -log(p(V^*|\mathcal{N}(\mu_V, \sigma_V))) - log(p(A^*|\mathcal{N}(\mu_A, \sigma_A))) \tag{5.1}$$

where V^* and A^* denote the ground truth labels (point values) of valence and arousal respectively, and $\mathcal{N}(\mu_V, \sigma_V)$ and $\mathcal{N}(\mu_A, \sigma_A)$ are the predictive distributions of valence and arousal respectively. It is worth noting that this approach to quantify the aleatoric uncertainty does not involve drawing multiple predictions at test time, unlike in the MC dropout method used for epistemic uncertainty modeling.

5.3.2 Evaluating Uncertainty-Aware Emotion Predictions

Since the ground truth emotion labels' true distributions are not known, it is not possible to directly assess the quality of emotion uncertainties. Here an evaluation protocol is proposed for predictive uncertainties based

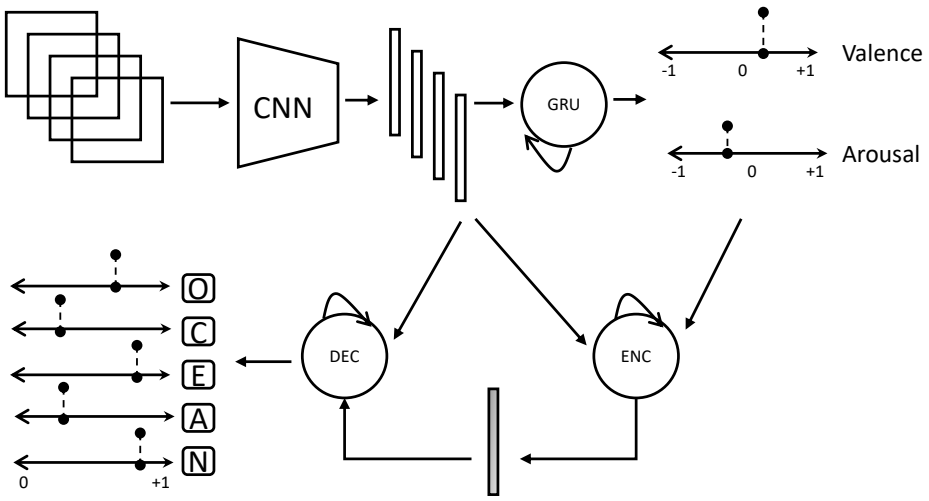


Figure 5.6: CLVM: $PT|X,(VA)$: Uncertainty-**unaware** CLVM for personality recognition using *point estimates of predicted valence and arousal* as inputs (ENC – encoder, DEC – decoder)

on the premise that the notion of modelling uncertainty in the predictions essentially refers to measuring the reliability of those predictions. By using uncertainty-aware predictions as inputs to a downstream task, this method can indirectly measure the quality of uncertainties based on how positively or negatively they impact the downstream task performance. Thus, this chapter evaluates the emotion prediction uncertainties by using them as input features in an important downstream behavioural analysis task, apparent personality traits estimation or simply personality recognition from face videos.

Apparent Personality Recognition refers to predicting continuous valued scores of Big Five (also known as Five Factor Model) personality traits: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN), from visual information [Escalante et al., 2020]. Unlike the dimensional emotions that are local (per-frame) behavioural attributes, video-based personality traits are typically global (per-sequence) behavioural constructs that are sequential cumulative functions of local

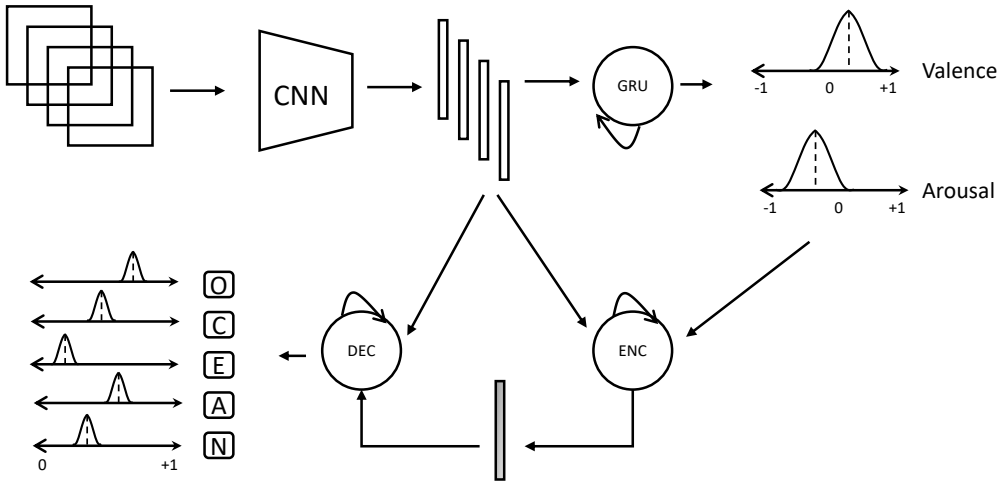


Figure 5.7: CLVM-A: $PT|X, (VA_{\mu}^{EU+AU}, VA_{\sigma}^{EU+AU})$: Uncertainty-**aware** CLVM for personality recognition using *distributions of predicted valence and arousal* as inputs (ENC – encoder, DEC – decoder)

emotional behaviour. As Fig. 5.5 shows, given a face video input, a personality recognition model estimates sequence-level predictions of the OCEAN traits.

As shown in prior works [Zhang et al., 2019], the performance of a personality recognition model heavily relies on how effectively it can aggregate local (per-frame) emotional expressive behavioural (e.g. valence and arousal) and map it to the personality traits. Hence, here the evaluation protocol of emotion uncertainty estimates is based on measuring the performance difference between personality recognition models that are trained with and without uncertainty estimates (epistemic-only, aleatoric-only, epistemic + aleatoric) of per-frame valence and arousal.

Conditional Latent Variable Models (CLVM) for Apparent Personality Traits Estimation

To utilise the already predicted dimensional emotions (valence and arousal) and their uncertainty estimates as inputs to a personality recognition model,

in addition to per-frame face embeddings provided by a CNN model, this chapter proposes a conditional latent variable model (CLVM). The design of CLVM is based on the premise that personality traits in a video are global behavioural primitives that are functions of local apparent emotions expressed in each frame. This chapter hypothesises that canonical sequence learning models (RNNs) are less effective in aggregating the personality-related global context spread over a typically long-term temporal scale. Inspired by different global context aggregation schemes proposed in [Garnelo et al., 2018a,b, Sanchez et al., 2021, Tellamekala et al., 2021], a CLVM is designed here not only to effectively aggregate the global context of personality-relevant information, but also to combine the uncertainty-aware emotion predictions with their corresponding facial embeddings in a principled approach.

Building Blocks of CLVM. As shown in Fig. 5.6 and Fig. 5.7, the proposed CLVM architecture shares the design principles originally introduced in Conditional Neural Processes (CNP) [Garnelo et al., 2018a]. CNP is an encoder-decoder composition in which the encoder (ENC) module infers a global latent variable based on temporal context aggregation, and the decoder (DEC) module is conditioned on that global latent variable to predict the target labels. Here, the latent variable is expected to inform the decoder about global temporal context, which is learned as a function of both feature space (X) and label space (Y) of a subset of frames in the input sequence. Wheras in the proposed CLVM, the encoder-decoder combination summarises the whole sequence of input features and emotion labels (image embeddings X_i concatenated with per-frame emotion predictions E_i) into a global latent variable which is then used to estimate the personality traits, $(X_i, E_i)_{i=1}^{N} \longrightarrow OCEAN$. Unlike in CNP [Garnelo et al., 2018a], here the encoder and decoder modules are implemented using GRUs, following

[Tellamekala et al., 2021], where the encoder outputs a global latent vector which the decoder uses, along with image embeddings X_i, to predict the distributions (mean and variance values) of personality traits. Similar to the aleatoric models, CLVMs are trained using the negative log likelihood loss (Eq. 5.1). More details of CLVM implementations can be found in Section. 5.4.2. To measure the quality of emotion uncertainties, the proposed method trains and evaluates both uncertainty-unaware (point estimates of emotion as inputs, as Fig. 5.6 shows) and uncertainty-aware (emotion distributions as inputs, as Fig. 5.7 shows) CLVMs for personality recognition.

Fusion of epistemic and aleatoric uncertainties

To introduce the predicted emotions and their uncertainties into the learning of personality recognition task, the valence and arousal distributions (their mean and variance vectors) are fed to the encoder module of CLVM (see Fig. 5.7). In the cases of epistemic-only and aleatoric-only, unimodal distributions parameters of the valence and arousal $[\mu_v, \sigma_v, \mu_a, \sigma_a]$ are passed as input features, along with per-frame image embeddings X provided by a CNN model. To inform the personality recognition model about the total uncertainty of predicted emotions, the epistemic and aleatoric distributions of valence and arousal are fused. To this end, the following two simple fusion techniques are considered:

Gaussian Mixture Model (GMM) Fusion assumes that the total uncertainty is a multimodal distribution composed of epistemic and aleatoric modes. With the mixture coefficients set to 1.0, this fusion leads to a a simple concatenation of the epistemic (EU) and aleatoric (AU) distribution

parameters separately for valence and arousal.

$$\mu_v = [\mu_v^{EU}, \mu_v^{AU}], \sigma_v = [\sigma_v^{EU}, \sigma_v^{AU}] \tag{5.2}$$

$$\mu_a = [\mu_a^{EU}, \mu_a^{AU}], \sigma_a = [\sigma_a^{EU}, \sigma_a^{AU}] \tag{5.3}$$

Sum Fusion method directly adds the parameters of both the distributions, assuming that the epistemic and aletoric components are two independent random variables carrying complementary information about the total uncertainty,

$$\mu_v = \mu_v^{EU} + \mu_v^{AU}, \sigma_v = \sigma_v^{EU} + \sigma_v^{AU} \tag{5.4}$$

$$\mu_a = \mu_a^{EU} + \mu_a^{AU}, \sigma_a = \sigma_a^{EU} + \sigma_a^{AU} \tag{5.5}$$

5.4 Experiments

Appendix C describes the datasets, evaluation metrics, and backbone CNN models used for evaluating the holistic uncertainty models of dimensional affect evaluated in this chapter.

5.4.1 Dimensional Emotion Recognition Models

Uncertainty-*Unaware* Emotion Recognition

As shown in Fig. 5.2, a temporal model on top of the backbone CNN was learned to predict the point estimates of valence and arousal on SEWA. This temporal model is composed of a 2-layer bidirectional GRU-RNN with 128 hidden units followed by one fully connected (FC) output layer with 2 units (for valence and arousal). It contains three dropout layers with a probability of 0.5. Note that at test time all the dropout layers are disabled in this uncertainty-unaware baseline. Inverse CCC (1.0 - CCC) + Mean Squared Error (MSE) loss function [Toisoul et al., 2021] was used for training this baseline by setting the sequence length to 100 frames.

Uncertainty-*Aware* Emotion Recognition

Epistemic Uncertainty Model of emotion recognition was implemented by applying Monte Carlo dropout [Gal and Ghahramani, 2016] to the temporal model at test time. This model was same as the above discussed uncertainty-unaware baseline (5.4.1) but with the dropout layers enabled during inference. Enabling dropout layers at test time alters the network configuration from run to run, as a result, the model produces different predictions for the same input. Thus, the generated ensemble of predictions were used to compute the mean and standard deviation output values, which represent the model or epistemic uncertainty. With 200 forward passes per frame, it is observed that the best personality recognition performance is achieved on the ChaLearn validation set.

Aleatoric Uncertainty Model is also same as the uncertainty-unaware baseline (5.4.1), but with its output FC layer (with 2 units) replaced with

a FC layer of 4 units to predict the mean and standard deviation values of valence and arousal: $(\mu_v, \sigma_v, \mu_a, \sigma_a)$. This model was trained separately using the negative log likelihood loss (Eq. 5.1).

Optimisation Details. All the above emotion recognition models are trained using Adam optimiser [Kingma and Ba, 2014] for 150 epochs with initial learning rate and weight decay values set to 1e-4 and 1e-5 respectively. The learning rate value is tuned using Cosine annealing method [Loshchilov and Hutter, 2016] (with warm restarts, multiplication factor set to 2 and first restart applied at epoch 1). Each mini batch has 6 input image sequences, with each sequence containing 100 face frames.

5.4.2 Personality Recognition Models

CNN+GRU Baseline of personality recognition model was composed of the backbone CNN followed by a temporal model – a 2-layer bidirectional GRU-RNN with 256 hidden units + one FC output layer with 5 units – to jointly predict per-video point estimates of the five personality traits, as shown in Fig. 5.5. Here also, the temporal model has 3 dropout layers with a probability of 0.5. Mean Absolute Error (MAE) is used to train this model.

CLVM Implementations

Emotion Uncertainty-*Unaware* CLVM. As shown in Fig. 5.6, similar to [Tellamekala et al., 2021], in the proposed CLVM the encoder module – a 1-layer bidirectional GRU-RNN with 128 hidden units + one FC layer with 256 output units – receives a face image embedding sequence along with its corresponding predicted emotions (point estimates) concatenated.

Here, the facial features learned for emotion recognition are fine tuned for the personality recognition task as well. The encoder module's FC layer outputs a global (256 dimensional) latent vector, which aims to capture the personality related global temporal context from the image embeddings and emotions. The decoder module – again a 1-layer bidirectional GRU-RNN with 128 hidden units + one FC layer with 5 nodes – takes as inputs the image embedding sequence and the global latent vector to predict the point estimates of personality traits. Similar to the CNN+GRU baseline training, this model was also trained to minimise the MAE objective.

Emotion Uncertainty-*Aware* CLVM. To make the CLVM aware of input emotion uncertainties, the proposed method introduced three changes (Fig. 5.7): (a). the point estimate emotion inputs to the encoder are replaced with mean and standard deviation values predicted using the epistemic-only or aleatoric-only or epistemic+aleatoric uncertainty-aware emotion recognition models (number of input units in the encoder's GRU changed accordingly) (b). the output FC layer with 5 output units in the decoder is replaced with a new FC layer containing 10 output units (5 for the mean vector and 5 for the standard deviation vector) and (c). for training the negative log likelihood loss (Eq. 5.1) is used instead of MAE.

Optimisation Details. All the personality recognition models evaluated in this chapter are trained using Adam optimiser [Kingma and Ba, 2014] for 100 epochs. Here the mini batch size is 3 sequences, each with 200 randomly sampled consecutive frames from the input video that originally contains 450 frames. At test time, non-overlapping windows of 200 frames are extracted and their predictions are averaged to produce the final outputs. Note that the emotion prediction model remains frozen during the training of CLVM encoder and decoder modules. The initial learning rate and weight decay values are set to 5e-5 and 1e-4 respectively. Similar to the

emotion recognition models, the learning rate value is tuned using Cosine annealing with warm restarts [Loshchilov and Hutter, 2016].

5.5 Results and Discussion

Model	Valence CCC ↑	Arousal CCC ↑	Avg. CCC ↑
[Mitenkova et al., 2019]	0.469	0.392	0.415
[Toisoul et al., 2021]	0.650	0.610	0.630
[Kossaifi et al., 2020]	**0.750**	0.520	0.635
Visual Affective Processes	0.739	**0.622**	**0.680**
VA: CNN+GRU	0.712	0.618	0.665
EU: CNN+GRU	0.717	0.614	0.665
AU: CNN+GRU	0.714	0.619	0.667
EU+AU: CNN+GRU	0.719	0.620	0.670

Table 5.1: Results on the test set of SEWA (VA – uncertainty-unaware baseline, EU and AU – Epistemic and Aleatoric Uncertainty-Aware models (see Section. 5.4.1))

5.5.1 Dimensional Affect and Personality Recognition

Emotion Recognition With vs. Without Uncertainty Modelling. Table. 5.1 compares the results of uncertainty-unaware (VA) and uncertainty-aware (EU, AU, and EU+AU) dimensional emotion recognition models on SEWA. Here, different uncertainty-aware models are learned for the epistemic (EU) and aleatoric (AU) categories separately, and their predictions (both mean and variance values) are averaged for computing the EU+AU results. For the evaluation purpose, only the mean values of emotion pre-

Metric	Model	Extr.	Agree.	Consc.	Neuro.	Open.	Avg.
PCC ↑	Histogram [Jaiswal et al., 2019]	0.30	0.05	0.22	0.22	0.20	0.20
	DCC [Güçlütürk et al., 2016]	0.36	0.12	0.20	0.25	0.25	0.24
	Spectral [Song et al., 2020]	0.37	0.30	0.34	0.36	0.32	0.34
	NJU-LAMDA [Wei et al., 2017]	0.43	0.37	0.45	0.34	0.36	0.39
	SSL [Song et al., 2021]	0.52	0.31	0.45	0.45	0.44	0.45
	CNN+GRU: $PT\|(VA)$	0.49	0.20	0.16	0.38	0.37	0.32
	CNN+GRU: $PT\|X$	0.59	0.33	0.45	0.48	0.50	0.47
	CNN+GRU: $PT\|X,(VA)$	0.61	0.36	0.48	0.52	0.50	0.49
	CLVM: $PT\|(VA)$	0.53	0.27	0.22	0.39	0.39	0.36
	CLVM: $PT\|X$	0.64	0.39	0.48	0.55	0.52	0.52
	CLVM: $PT\|X,(VA)$	0.66	0.42	0.51	0.58	0.56	0.54
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU},(VA)_{\sigma}^{EU})$	0.69	0.50	0.59	0.63	0.61	0.61
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{AU},(VA)_{\sigma}^{AU})$	0.69	0.49	0.61	0.62	0.61	0.60
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -GMM Fusion	0.68	0.50	0.56	0.62	0.60	0.59
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -Sum Fusion	**0.71**	**0.54**	**0.62**	**0.66**	**0.63**	**0.64**
RMSE ↓	Histogram [Jaiswal et al., 2019]	0.170	0.150	0.170	0.170	0.160	0.160
	DCC [Güçlütürk et al., 2016]	0.150	0.140	0.150	0.150	0.140	0.150
	Spectral [Song et al., 2020]	0.150	0.130	0.140	0.140	0.140	0.140
	NJU-LAMDA [Wei et al., 2017]	0.140	0.120	0.130	0.140	0.130	0.130
	SSL [Song et al., 2021]	0.120	0.100	0.130	0.120	0.110	0.120
	CNN+GRU: $PT\|(VA)$	0.133	0.131	0.155	0.143	0.134	0.139
	CNN+GRU: $PT\|X$	0.121	0.124	0.137	0.132	0.120	0.126
	CNN+GRU: $PT\|X,(VA)$	0.118	0.121	0.135	0.127	0.121	0.124
	CLVM: $PT\|(VA)$	0.127	0.128	0.149	0.141	0.133	0.135
	CLVM: $PT\|X$	0.111	0.120	0.136	0.122	0.120	0.121
	CLVM: $PT\|X,(VA)$	0.109	0.118	0.130	0.121	0.118	0.119
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU},(VA)_{\sigma}^{EU})$	0.100	0.106	0.112	0.109	0.107	0.109
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{AU},(VA)_{\sigma}^{AU})$	0.102	0.104	0.115	0.111	0.109	0.108
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -GMM Fusion	0.108	0.105	0.119	0.112	0.110	0.110
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -Sum Fusion	**0.096**	**0.103**	**0.110**	**0.109**	**0.105**	**0.104**
Accuracy ↑	Histogram [Jaiswal et al., 2019]	0.8949	0.8970	0.9001	0.8913	0.8975	0.8962
	DCC [Güçlütürk et al., 2016]	0.9088	0.9097	0.9109	0.9085	0.9092	0.9109
	Spectral [Song et al., 2020]	0.9165	0.9099	0.9178	0.9109	0.9117	0.9134
	NJU-LAMDA [Wei et al., 2017]	0.9112	0.9135	0.9128	0.9098	0.9105	0.9116
	SSL [Song et al., 2021]	0.9183	0.9262	0.9082	0.9133	0.9180	0.9168
	PML [Bekhouche et al., 2017]	0.9155	0.9103	0.9137	0.9082	0.9100	0.9115
	PersEmoN [Zhang et al., 2019]	0.9200	0.9140	0.9210	0.9140	0.9150	0.9170
	CR-Net [Li et al., 2020b]	0.9200	0.9176	0.9218	0.9150	0.9191	0.9187
	CNN+GRU: $PT\|(VA)$	0.8930	0.8944	0.8765	0.8850	0.8920	0.8881
	CNN+GRU: $PT\|X$	0.9040	0.8991	0.8833	0.8918	0.8980	0.8953
	CNN+GRU: $PT\|X,(VA)$	0.9062	0.8999	0.8871	0.8963	0.8978	0.8974
	CLVM: $PT\|(VA)$	0.8988	0.8977	0.8789	0.8878	0.8934	0.8913
	CLVM: $PT\|X$	0.9076	0.9002	0.8872	0.8998	0.9027	0.8995
	CLVM: $PT\|X,(VA)$	0.9100	0.9009	0.8935	0.9005	0.9048	0.9019
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU},(VA)_{\sigma}^{EU})$	0.9217	0.9178	0.9133	0.9156	0.9168	0.9172
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{AU},(VA)_{\sigma}^{AU})$	0.9205	0.9197	0.9106	0.9136	0.9151	0.9160
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -GMM Fusion	0.9199	0.9175	0.9092	0.9130	0.9146	0.9148
	CLVM-A: $PT\|X$, $((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$ -Sum Fusion	**0.9263**	**0.9192**	**0.9148**	**0.9157**	**0.9197**	**0.9191**

Table 5.2: Personality recognition on ChaLearn (CLVM:..-Uncertainty-*unaware*, CLVM-A:..-uncertainty-*aware*, PT-personality traits, VA-valence & arousal, X-image features, EU-epistemic uncert., AU-aleatoric uncert.)

dictions are used from the uncertainty-aware models.

As it is clearly evident from the results in Table. 5.1, making the emotion recognition models aware of their predictive uncertainty did not improve the performance compared to the uncertainty-unaware baseline. Note that the primary objective of uncertainty-aware learning in this chapter is to reliably estimate the standard deviation intervals around a model's predictions, and they may not necessarily make the mean values of predictions more accurate. Such a relationship between the model accuracy and its predictive uncertainty was discussed in more detail in several prior works (e.g. [Krishnan and Tickoo, 2020]).

However, when the uncertainty estimates are applied to a downstream task i.e. personality recognition, it is expected to observe noticeable performance gains with the uncertainty-aware models over uncertainty-unaware baselines. Table 5.2 shows the personality recognition results of a wide range of models on ChaLearn test set. This section presents a detailed analysis and discussion of these results below.

Personality Recognition With Uncertainty-Aware Vs. Uncertainty-Unaware Dimensional Emotion Predictions Emotion uncertainty-aware CLVMs (CLVM-A models in Table 5.2) that utilised epistemic-only or aleatoric-only or epistemic+aleatoric uncertainties of valence and arousal predictions, outperformed the baseline model CLVM: $PT|X,(VA)$ that is trained using the point estimates of valence and arousal. This result validates that the uncertainty-aware emotion recognition models are able to quantify their confidence in their predicted emotions such that the downstream task can effectively utilise those uncertainty estimates as reliability indicators for the emotion information. Among the uncertainty-aware models (CLVM-A in Table 5.2), the model that combines both epistemic

and aleatoric uncertainty, CLVM-A: $PT|X, ((VA)_{\mu}^{EU+AU}, (VA)_{\sigma}^{EU+AU})$ with sum fusion, outperformed the epistemic-only and aleatoric-only CLVM-A models. While the performance gap between epistemic-only and aleatoric-only CLVMs is marginal, significant performance improvement with epistemic+aleatoric fusion confirms the complementary nature of both the uncertainty components. Also, it achieved state-of-the-art personality recognition results by outperforming the existing ChaLearn benchmarks in all three metrics.

Fig. 5.9 qualitatively illustrates trait-wise predictions' correlation patterns for all examples in the ChaLearn test set. As evident from this correlation analysis, uncertainty-aware CLVM exhibited highest correlation performance for the extroversion trait, and the lowest correlation performance in the case of agreeableness trait. Furthermore, as Fig. 5.8 illustrates, on an example from the ChaLearn test set this section qualitatively compared the personality trait scores predicted by the uncertainty-unaware (CLVM) and different uncertainty-aware (CLVM-A) models, and their corresponding input valence and arousal ratings. Here, the confidence intervals of emotion predictions from different uncertainty models indicate 3 $\times$ the standard deviation values around the mean predictions. Note that ChaLearn is not annotated with emotion labels, so it is not possible here to evaluate the quality of emotion predictions. However, as evident from Fig. 5.8, in terms of the personality recognition results, the scores predicted by the epistemic and aleatoric uncertainty-aware model (EU+AU) are much closer to the ground truth personality scores, compared to the scores predicted by the uncertainty-unaware emotion predictions (VA). Regarding the uncertainty patterns in emotion predictions from different models, it is interesting to observe that the confidence intervals are slightly smoother in the case of aleatoric uncertainty than in the epistemic case.

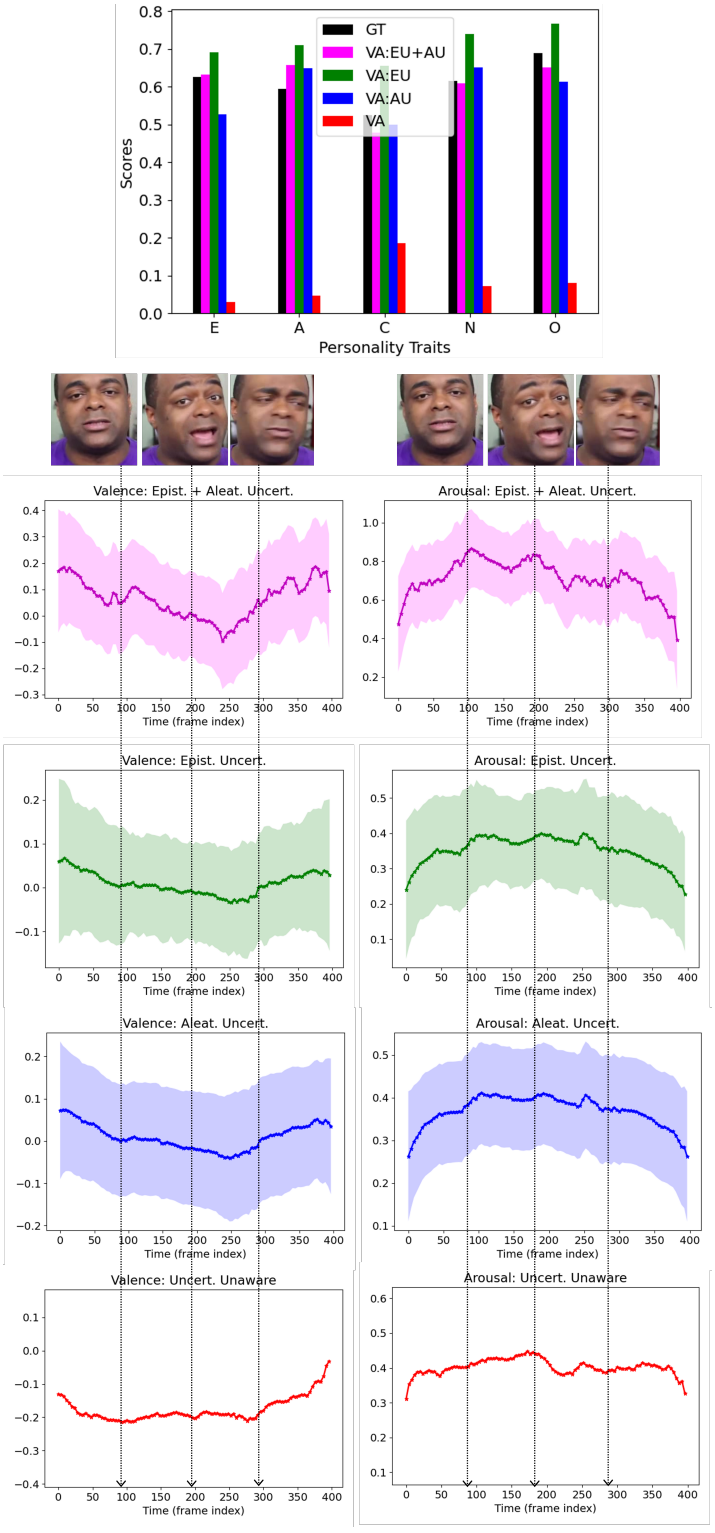


Figure 5.8: Comparison of different CLVMs' predictions on an example from the ChaLearn test set: Trait-wise ground truth scores (GT) are compared with the predictions made by emotion (valence and arousal) uncertainty-unaware (VA) model, and different uncertainty-aware models (EU-Epistemic, AU-Aleatoric, and EU+AU). Confidence intervals of the valence and arousal predictions depict three times the standard deviation values predicted their corresponding uncertainty models.

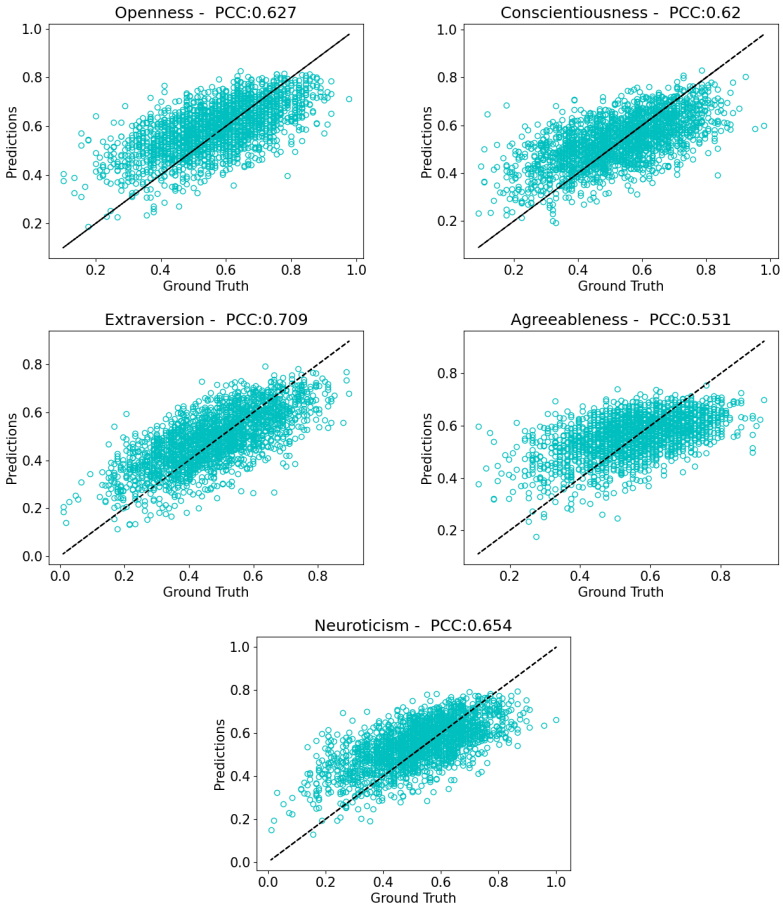


Figure 5.9: Correlation Analysis: Trait-wise predictions from epistemic+aleatoric uncertainty-aware CLVM on ChaLearn test set

CLVM Vs. CNN+RNN for Personality Recognition. All three CLVM baseline models, $PT|(VA)$, $PT|X$, and $PT|X,(VA)$, achieved significant performance gains over the canonical CNN+GRU baselines, ($PT|(VA)$, $PT|X$, and $PT|X,(VA)$) on all five traits. These gains validate the effectiveness of learning global latent variable models for summarising the temporal context of personality traits from video data. Moreover, this result is in line with the observations reported in [Sanchez et al., 2021, Tellamekala et al., 2021] regarding the effectiveness of global context models compared to RNNs. Furthermore, it is important to note that the results with emotion predictions (VA) used as input features in CNN+GRU models are less significant compared to the performance improvements achieved by their CLVM counterparts. This performance gap shows that our proposed CLVM is more effective than the canonical models based on CNN+GRUs in combining the personality-related information embedded in the emotion predictions and face image features. Thus, by including an intermediate step in the form of global latent variable modelling, the proposed CLVM framework offers the flexibility to effectively combine different low level behavioural attributes (e.g. emotions) as prior variables in inferring high level behavioural attributes (e.g. personality traits).

5.5.2 Ablation Study

Fusion of Epistemic and Aleatoric Uncertainty. Table 5.2 also compares the performance of uncertainty fusion techniques evaluated in this chapter: sum fusion and GMM fusion. Sum fusion, the fusion technique based on linear combination of random variables with normal distributions, consistently outperformed the GMM fusion technique in all three metrics. This result implies that epistemic and aleatoric models capture the inde-

pendent and complementary components of the total uncertainty.

CLVM With vs. Without Emotion Predictions as Input Features. To understand the contribution of emotion prediction inputs and global latent variable modelling to the CLVM performance, this work trained three baseline models of CLVM, $PT|X$, $PT|(VA)$ and $PT|X,(VA)$, as shown in Table 5.2. The performance gaps between these three models indicate that both the proposed ideas, the global latent variable modelling as well as the emotion prediction inputs, are critical to the performance gains that the CLVM demonstrated.

CNN+GRU With vs Without Emotion Predictions as Input Features. By concatenating the emotion predictions (VA) to the image features X, this work evaluated the CNN+GRU models, $PT|X$, $PT|(VA)$ and $PT|X,(VA)$ for personality recognition in Table 5.2. Unlike in the case of CLVM baselines, performance gains are less significant when the emotion predictions are used as additional input features in CNN+GRU models. This result demonstrates that the canonical sequence learning models based on CNN+GRUs are less effective compared to our CLVM in terms exploiting the already known behavioural attributes such as per-frame emotion predictions in the downstream tasks. For more ablation experiments of the CLVM and the significance of its performance gains, refer to the supplementary material.

5.5.3 Statistical Significance Analysis

Note that the personality recognition performance difference between the emotion uncertainty-unaware (CLVM: $PT|X,(VA)$) and uncertainty-aware (CLVM-A: $PT|X,((VA)_{\mu}^{EU+AU},(VA)_{\sigma}^{EU+AU})$-Sum Fusion) models in Table. 5.2

appears more significant in the PCC metric but less significant in the RMSE and accuracy metrics. To verify the significance of overall improved recognition results with emotion uncertainties, we conducted paired Student's t-test on both the models (CLVM and CLVM-A) for each trait separately. As shown in Table. 5.3, across all five traits the p-values are noticeably low, validating that the the performance gains achieved by the CLVM-A model are statistically significant.

Model Pair	Extr.	Agree.	Consc.	Neuro.	Open.
(CLVM, CLVM-A)	3.6e-17	4.1e-49	7.5e-40	4.6e-24	3.8e-15

Table 5.3: Statistical significance (p <0.01) analysis results on the ChaLearn test set: Paired Student's t-test between the emotion uncertainty-unaware (CLVM) and uncertainty-aware (CLVM-A) predictions of all five traits separately.

5.5.4 Application of Affective Processes' Emotion Predictions and their Uncertainty Estimates to Personality Recognition

As shown in Table. 5.1, Affective Processes (APs) has better emotion recognition performance than all three CNN+GRU models that are trained and evaluated in this chapter. Here, we investigated the possibility that personality recognition models may perform better using APs' emotion predictions, than the epistemic and aleatoric uncertainty-aware emotion predictions from our CNN+GRU models. For this purpose, in the CLVM model training we replaced our CNN+GRU emotion predictions with APs' emotion predictions. It is important to note that in APs the decoder provides both mean and variance values over the emotion predictions.

As Table. 5.4 shows, in the uncertainty-unaware (CLVM) case, APs per-

form better than our CNN+GRU models (compared to CLVM: $PT|X - AP, (VA) - AP$ in Table. 5.2). However, CLVM-A model i.e. the uncertainty-aware counter part of APs has exhibited poorer performance compared to our emotion uncertainty predictions (CLVM-A: $PT|X, ((VA)_{\mu}^{EU+AU}, (VA)_{\sigma}^{EU+AU})$-Sum Fusion in Table. 5.2). This result shows that although more accurate emotion predictions could improve the personality recognition performance slightly, making them uncertainty-aware in a holistic manner (combining epistemic and aleatoric components) achieves significantly better results on the downstream task.

Metric	Model	Extr.	Agree.	Consc.	Neuro.	Open.	Avg.
PCC↑	CLVM	0.67	0.45	**0.57**	0.61	**0.60**	0.57
	CLVM-A	**0.69**	**0.47**	0.56	0.61	0.58	**0.58**
RMSE↓	CLVM	0.110	**0.110**	0.124	0.123	**0.108**	0.115
	CLVM-A	**0.106**	0.111	**0.120**	**0.116**	0.112	**0.113**
Acc.↑	CLVM	0.9047	**0.9052**	0.8941	0.895	**0.9054**	0.9008
	CLVM-A	**0.9082**	0.9038	**0.8963**	**0.9004**	0.9029	**0.9023**

Table 5.4: Personality recognition results on the ChaLearn test set using **Affective Processes (APs)** emotion predictions and their uncertainty estimates: Here, the uncertainty-aware (CLVM-A) is comparable to $PT|X, ((VA)_{\mu}^{EU+AU}, (VA)_{\sigma}^{EU+AU})$ in Table. 5.2 and the uncertainty-unaware model (CLVM) is equivalent to $PT|X, (VA)$ in Table. 5.2.

5.5.5 Emotion Predictions Directly Fed Into The CLVM Decoder

To delineate the influence of latent vector on the decoder module in the proposed CLVM, we modified the architecture as follows: the latent variable to the decoder is replaced with the raw emotion predictions in both uncertainty-unaware and uncertainty-aware configurations. As a result, in this modified architecture, the decoder has access to only the local (per-

frame) emotional behaviour, unlike in the original CLVM model where the global context is fed into the decoder in the form of latent variable input. Validating our hypothesis about the importance of global emotion context for personality trait analysis, the results shown in Table. 5.5 confirm that the performance of CLVM drops significantly in the absence of global latent variable input to the decoder.

Metric	Model	Extr.	Agree.	Consc.	Neuro.	Open.	Avg.
PCC↑	CLVM	0.60	0.38	**0.49**	0.53	0.52	0.50
	CLVM-A	**0.62**	**0.40**	0.48	**0.54**	**0.53**	**0.51**
RMSE↓	CLVM	0.117	0.122	**0.130**	0.128	0.122	0.124
	CLVM-A	**0.115**	**0.120**	0.131	**0.127**	**0.120**	**0.122**
Acc.↑	CLVM	0.9057	0.9009	**0.8945**	0.8970	0.9005	0.8997
	CLVM-A	**0.9072**	**0.9030**	0.8937	**0.8977**	**0.9029**	**0.9009**

Table 5.5: ChaLearn test set results with **emotion predictions directly fed to the CLVM decoder**. Note that here the latent variable input to the decoder is replaced with uncertainty-unaware (CLVM) and uncertainty-aware (CLVM-A) emotion predictions directly.

5.6 Conclusion

Towards capturing the holistic temporal uncertainty of temporal affect, this chapter presented a systematic decomposition of uncertainties in dimensional emotion recognition from face videos. The methodology proposed in this chapter first quantified epistemic (model) and aleatoric (data) uncertainty components of the emotion recognition, without requiring any additional information. Then it evaluated the quality of emotion uncertainties by using them as additional input features in apparent personality recognition task. To this end, this chapter proposed to use a conditional global latent variable model to effectively summarise temporal context of the personality traits from uncertainty-aware emotion predictions and face image

features, which achieved state-of-the-art results on in-the-wild personality recognition. Most importantly, the proposed emotion uncertainty-aware personality recognition models achieved substantial performance gains over their uncertainty-unaware counterparts, validating the quality of the emotion uncertainty estimates quantified in this chapter. Further, compared to the predictive uncertainty estimates of previously proposed methods such as Affective Processes, the holistic uncertainty models based on epistemic and aleatoric uncertainty estimates of affect recognition, demonstrated considerable performance improvements in the downstream behavioural learning task.

Chapter 6

Label-Efficient Affect Recognition using CHeF: Clustering Hand-Engineered Emotion Features for Self-Supervised Pre-Training

6.1 Introduction

This chapter tackles the problem of learning emotion recognition models from face and voice data *with minimal human supervision*. State-of-the-art emotion recognition methods heavily rely on end-to-end representation learning models, which are highly inefficient in terms of the amounts of labelled training data they require. With the objective of making emotion recognition models label-efficient, a novel self-supervised representation learning method is proposed in this chapter.

The proposed method aims to exploit emotion-related cues embedded in the standard hand-engineered features of affect as *free supervision signals* for pre-training the feature encoders. Compared to the end-to-end representation learning models, models learned from hand-crafted features tend to have poor predictive performance in general. However, it is worth noting that hand-crafted features are designed for a particular task at hand, and they essentially capture the task-specific intuitive knowledge, often using a compact set of descriptors. Thus, hand-engineered features could summarise, at least partially, the key latent characteristics of a learning task, often in a lower-dimensional space than that of the raw input data. Most importantly, it does not require any manual supervision to extract the standard hand-engineered emotion features from large amounts of unlabelled audiovisual data.

To give an example, in the case of voice emotion recognition, compact low-level descriptors such as the extended Geneva Minimal Acoustic Parameter Set (eGeMAPs) [Eyben et al., 2015] capture expert-level understanding of emotion-related cues in vocal expressions. Similarly, in face emotion recognition models, the predictions of facial action unit (AU) intensities are commonly used as hand-crafted features of facial affect expressions (e.g. [Tarnowski et al., 2017] and [Senechal et al., 2014]), and they aim to objectively describe changes in the facial muscle movements caused by emotional expressions. In both vocal and facial emotion recognition tasks, their corresponding hand-engineered features perform reasonably well in general, but not as well as the end-to-end learning models [Kollias et al., 2020]. However, these hand-engineered features can be highly valuable in guiding the end-to-end representation learning models, given that they capture emotion cues embedded in the raw audio and visual data. Guided by this intuition, this chapter proposes to leverage the standard hand-

engineered emotion features as *task-specific representation learning priors* in the self-supervised pre-training of the audio and visual CNN feature encoders.

Given a large unlabelled dataset, a self-supervised learning (SSL) model is trained by learning a pretext or proxy task, for which labels can be automatically generated based on the intrinsic structure of the dataset. Effectively formulating the proxy task is a critical factor in learning an SSL model, as the proxy task defines how well the feature encoder can capture the rich semantics embedded in the high-dimensional unlabelled data. The proposed proxy task for SSL pre-training in this chapter, is based on a novel clustering paradigm, Max-Margin Deep Temporal Clustering, applied to the facial and vocal hand-crated features widely used for emotion recognition. The cluster indices derived in this process are used as the target class labels for pre-training the visual and audio CNN feature encoders.

This chapter evaluates the efficacy of the CNN encoders pre-trained using the proposed method, by analysing their unimodal emotion recognition performance on the benchmark face and voice datasets. The experimental analysis shows that the proposed SSL pre-training considerably outperforms the standard transfer learning methods and performs almost on par with the pre-trained representations on other closely related emotion recognition datasets. Thus, this chapter demonstrates how to improve the label-efficiency of existing facial and speech emotion recognition models, by using self-supervision for combining the best of both worlds – rich task-specific information in the hand-engineered emotion features with the superior generalisation performance of the end-to-end representation learning models.

6.2 Related Work

SSL in Face and Speech Emotion Recognition. Due to the limited number of training examples in the existing emotion-labelled datasets, several recent works (e.g [Shukla et al., 2021, Roy and Etemad, 2021, Morais et al., 2022]) have turned to the idea of leveraging abundantly available unlabelled data, particularly through self-supervised pre-training, for improving emotion recognition performance. Most existing SSL methods applied to face emotion recognition [Roy and Etemad, 2021] and speech [Neumann and Vu, 2019, Morais et al., 2022] emotion recognition tasks exploit only unimodal information in the unlabelled data. Considering the intrinsically multimodal nature of emotional expressions, the focus in recent works [Khare et al., 2021, Shukla et al., 2021, 2020] is shifted to leveraging multimodal unlabelled data for SSL method applied to the emotion recognition tasks.

Although SSL pre-trained representations demonstrated promising results in all the early works, the proxy tasks used in their SSL models are largely evaluated on generic target tasks such as image classification, speech recognition, etc. Considering that emotion is often a weak signal embedded in high-dimensional input space, this work argues that using such generic proxy tasks can limit the potential of SSL methods when applied to the emotion recognition problem. To address this limitation, a downstream-specific proxy task is proposed in this chapter to improve the SSL pre-trained representations' quality for improved emotion recognition performance. Further, unlike the experimental analysis presented in this chapter, the existing works that aim to leverage SSL models for emotion recognition tasks mostly ignore the analysis of label-efficiency advantages.

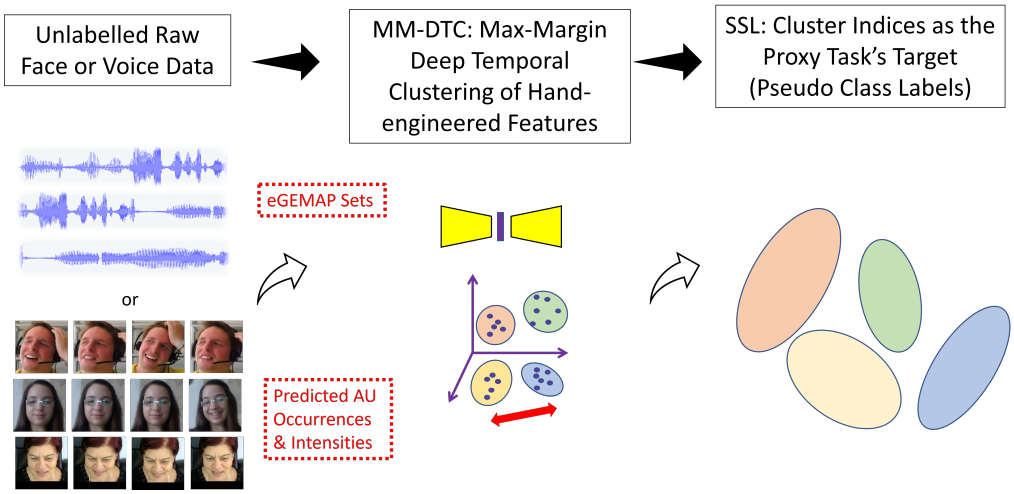


Figure 6.1: Illustration of the proposed proxy task, CHeF: Clustering Hand-Engineered Features, for self-supervised pre-training: First, the hand-engineered features of the downstream learning task (emotion recognition) are clustered in a low-dimensional latent space, guided by the proposed Max-Margin Deep Temporal Clustering technique. Then, the cluster indices are posed as *pseudo* class labels to be used as targets in learning the proxy task.

6.3 Method

A General Overview of Self-Supervised Pre-Training. The main objective of SSL pre-training methods is to encourage the feature encoder model to capture the general semantics embedded in unlabelled data, most importantly, without requiring any manual supervision. Given a large collection of unlabelled data points, in the SSL pre-training phase of a model the goal is to learn the parameters of its feature encoder module, which maps a high dimensional input into a low-dimensional feature embedding. As there is no explicit supervision signal available for learning the features encoder here, a pretext or proxy task is designed such that its labels can be automatically generated from the intrinsic structure of the unlabelled dataset. This structure is often defined in terms of generic representation learning priors [Bengio et al., 2013] such as predictability, redundancy, spatial or temporal coherency, invariance to different views of the same image,

etc. These priors are expected to force the feature encoder to learn representations of general semantics from the unlabelled data, and it is implicitly assumed that such general semantic representations transfer well to all relevant downstream supervised learning tasks. Thus, the effectiveness of a proxy task is determined based on how well its resultant feature encoder performs on the downstream tasks.

Note that in a typical SSL pre-training method the proxy task's design choices are mostly agnostic to the characteristics of the downstream tasks. In general, this downstream-agnostic SSL approach is found to be effective in the case of standard learning tasks [Jing and Tian, 2020] such as image classification, object detection etc, in the existing works. However, considering that apparent emotion is often a weak and noisy semantic factor embedded in a high-dimensional input space, this chapter argues that the existing downstream-agnostic proxy tasks are less effective in learning emotion-related feature representations. Motivated by this argument, this chapter hypothesises that the closer the SSL pre-trained features get to the downstream task characteristics, the lower the labelled examples requirement becomes when learning a target downstream task.

6.3.1 Hand-engineered Features as Priors for Self-Supervised Representation Learning

To account for the target downstream task's characteristics in the process of designing a proxy task, this chapter proposes to exploit the downstream task-specific hand-engineered features as representation learning priors. It is worth noting that such hand-engineered features are typically compact low-dimensional descriptors of the target task's properties, at least partially. Given a raw high-dimensional input, most importantly, its hand-

crafted feature representations for a specific target learning task are *freely* available, as it does not require any manual supervision to extract them from unlabelled data.

For example, pitch variation is an important property that is commonly captured in the hand-crafted acoustic features designed for speech emotion recognition tasks. When the pitch variation is used as a representation learning prior in an SSL proxy task, it is more likely that the feature encoder can capture emotion-related properties much better than the most existing SSL proxy tasks that are guided by generic representation learning priors. Based on this premise, this chapter proposes a novel proxy task that leverages hand-engineered features as the target-specific priors to guide the SSL pre-training. As discussed below, the proposed proxy task builds on the idea of grouping unlabelled data points in a low-dimensional embedding space composed of their hand-engineered features' summary.

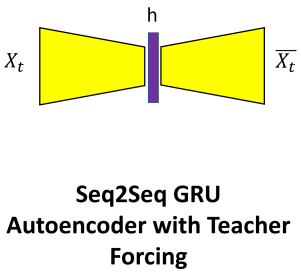
6.3.2 Proxy Task: Clustering Hand-Engineered Features to Derive *Pseudo* Class Labels

Building on the existing SSL proxy tasks that apply the idea of clustering unlabelled data in a low-dimensional embedding space [Caron et al., 2018], this chapter proposes a novel proxy task, which takes as inputs sequences of hand-crafted features. As illustrated in Fig. 6.1, given a set of unlabelled sequences of face images or speech segments as inputs, first a k-means clustering model is learned. This model is intended to group the input sequences into k different clusters, according to the Euclidean distances among their hand-crafted feature sequences. Then, the cluster indices provided by the resultant clustering model for a particular unlabelled example is considered as its *pseudo* class label. These labels are

used as the supervision signals for pre-training the feature encoder module by posing the proxy task as a standard classification problem.

It is important to account for the sequential structure of unlabelled data points while learning the clustering model. Note that the temporal variations in per-frame hand-engineered features are important for learning the downstream task of interest here i.e. apparent emotion recognition. Directly clustering the raw sequences of hand-crafted features can be difficult, as it scales up the problem dimensionality. For this reason, the proposed proxy task first runs a dimensionality reduction step by compressing a sequence of features into a single temporal summary or context vector.

To make sure that the resultant lower dimensional space supports the clustering operations well, both the operations, dimensionality reduction and k-means clustering, are jointly learned in the proposed proxy task. Simple clustering methods based on the standard k-means may fall short in learning an unified model that can implement all the aforementioned ideas: temporal clustering of hand-crafted features through simultaneous dimensionality reduction and k-means clustering with max-margin constraint. Ideally, while solving a classification problem it is desirable to have the target class labels that are as highly discriminative as possible. To make the pseudo labels i.e. cluster indices as discriminative as possible here, this chapter proposes to constrain the clustering model such that the learned cluster centroids are as far apart as possible from each other. This notion of maximising the centroid-to-centroid distance can be implemented by applying a maximum-margin regularisation constraint to the training objective of the clustering model. Thus, this chapter proposes a novel clustering methodology, Max-Margin Deep Temporal Clustering, to implement the idea of CHeF by combining all the three ideas into a single model that



$$Training\ Objective = \lambda * RE + \beta * C2P - \gamma * C2C$$

$$argmin - Reconstruction\ Error\ (RE) = C_i || X_t - \overline{X_t} ||$$

$$argmin - Centroid\ to\ Point\ Distance\ (C2H) = || C_i - h ||$$

$$argmax - Centroid\ to\ Centroid\ Distance\ (C2C) = || C_i - C_j ||$$

Figure 6.2: Implementation of the proposed Max-Margin Deep Temporal Clustering model using a sequence-to-sequence (Seq2Seq) autoencoder composed of the GRU-RNN encoder and decoder modules, and its training objective composed of the standard reconstruction loss coupled with clustering-specific loss components.

is trained by optimising a joint loss function, as discussed in detail below.

6.3.3 MM-DTC: Max-Margin Deep Temporal Clustering

To prepare the **pseudo** class labels for self-supervised pre-training, the proposed MM-DTC framework performs the following three operations:

a. Dimensionality Reduction: As illustrated in Fig. 6.2, given a sequence of hand-engineered features, here the goal is to encode its temporal summary into a lower dimensional 1D vector. For this purpose, a sequence-to-sequence autoencoder (Seq2Seq-AE) model is adopted here, which is widely used in the literature of audio representation learning (e.g. [Amiriparian et al., 2017]). Here, the autoencoder module, composed of encoder and decoder blocks, is tasked with the reconstruction of input sequences. Here, the encoder and decoder blocks are implemented using standard gated recurrent neural networks.

First the encoder compresses an input sequence into a 1D temporal context vector h, which is fed into the decoder module to predict the input sequence

step-by-step. At any given time step t, the decoder receives its previous prediction at the time step $\bar{\mathbf{x}}_{t-1}$ and its corresponding hidden state vector h_{t-1}. To stabilise the reconstruction performance of the AE model, especially during the early training stages, Teacher-Forcing technique [Williams and Zipser, 1989] is applied, in which the decoder input $\bar{\mathbf{x}}_{t-1}$ is randomly replaced with the original input $\mathbf{x}_{t-1}$, with the probability of 0.25. The parameters or weights of the encoder and decoder modules in this Seq2Seq-AE model are optimised by minimising the following sequence reconstruction loss:

$$L_{RE} = \sum_{t=1}^{T} (\mathbf{x}_t - \bar{\mathbf{x}}_t) \tag{6.1}$$

Note that the AE model is first pre-trained using the reconstruction loss alone for 10 epochs before the clustering operation is included in the model training. After the pre-training phase, to initialise the cluster centroids, a simple k-means clustering model is applied to the hidden state vectors ($\{h_{i=1}^{N}\}$) of all the N unlabelled data points in the dataset, as shown in Fig. 6.1

b. Temporal Context Clustering is applied to the output hidden state data h from the encoder module of the AE model. During this phase of AE model training, given a mini batch composed of B hand-crafted feature vector sequences, the clustering operation is applied to their corresponding hidden vector set $\{h_{i=1}^{B}\}$. For each hidden vector in this set, first its closest centroid among the k centroids is chosen by computing its Eucledian distances w.r.t all k centroids. To improve the clustering performance, the AE model is optimised to minimise the following loss term, in addition to the reconstruction objective.

$$L_{C2H} = \|C_k - h_i\|_2 \tag{6.2}$$

where C_k denotes the centroid that is closest to the current hidden state vector h_i. This loss function is intended to improve the clustering performance by minimising the distance values between the cluster centroids and their corresponding hidden states, thus encouraging the clusters to become as compact as possible. Note that locations of k centroids are smoothly updated every iteration based on the hidden state vectors computed for the current mini batch.

c. Max-Margin Constraint is applied as a regularisation condition, alongside the aforementioned reconstruction and clustering loss components. For learning the cluster centroids that are as discriminative as possible, this chapter proposes to apply a maximum-margin regularisation constraint by learning the hidden state vectors that can maximise the pair-wise centroid-to-centroid distances,

$$L_{C2C} = -\|C_j - C_k\|_2, \forall j, k, j \neq k \tag{6.3}$$

Thus, the complete training objective used for the proposed MM-DTC framework is composed of all the three aforementioned loss components:

$$L_{total} = \lambda * L_{RE} + \beta * L_{C2H} - \gamma * L_{C2C} \tag{6.4}$$

where λ, β, and γ are the hyper-paramters that are tuned to maximise the target downstream task's performance on its corresponding validation set.

6.4 Experiments

6.4.1 CHeF SSL Pre-Training of CNN Feature Encoders

Dataset. A large-scale audiovisual dataset, VoxCeleb2, is used for the purpose of SSL pre-training in this chapter. This dataset contains over one million utterances that are derived from YouTube videos recorded in in-the-wild conditions. Considering the high computational costs associated with clustering such large amounts of data, we chose to use a small partition of it containing over 36,000 utterances. The same utterances are used for training both the visual-only and audio-only SSL models on their corresponding face image sequences and voice signals respectively.

Network architectures of the visual and audio CNN feature encoder implementations are described in Appendix D.

Hand-Engineered Features for *Vocal* Emotion Recognition. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [Eyben et al., 2015] is a standard set of vocal features that are widely used for training speech emotion recognition models. These parameters capture emotion-related speech characteristics by computing various low-level descriptors such as frequency (pitch, jitter, etc), energy (shimmer, loudness, etc), and spectral balance and dynamics (harmonic differences, spectral slope, etc). Following prior works like [Ringeval et al., 2018, 2019, Mallol-Ragolta et al., 2020, Schmitt et al., 2019], first and second order functionals of both feature sets are computed here using a sliding window of 4 seconds with a stride of 100 ms. For a given input audio segment, a 46 dimensional feature vector is extracted.

Hand-Engineered Features for *Facial* Emotion Recognition. The presence of different facial Action Units (AUs) and their intensities are considered as the hand-crafted features for facial emotion recognition here. AUs capture atomic changes in the facial muscle movements to objectively describe facial expressions. Given that the annotation of AUs for large-scale in-the-wild datasets such as VoxCeleb2, is a prohibitively expensive process, this work makes use of a standard AU prediction model provided by the OpenFace 2.0 toolkit. This model predicts the presence or absence of 18 AUs and their intensity values in the range 0 to 5, which are normalised to the range [0, 1] here. Thus, for a given face image input, the binary predictions of AUs' presence and their continuous-valued intensity predictions are concatenated into a single feature vector.

Seq2Seq Auto Encoder Network is implemented by stacking a set of encoder and decoder modules implemented using bidirectional Gated Recurrent Unit Recurrent Neural Networks (BiGRU-RNNs). The encoder module maps the input sequence of hand-crafted features into a 32 dimensional embedding, using which the decoder aims to reconstruct the input sequence. Two different 2-layer BiGRU-RNNs with 32 hidden units are used as the encoder and decoder modules. Note that the input dimensionality for the encoder module is same as the output dimensionality of the decoder, and input dimensionality is different for the audio and visual models.

SSL Pre-Training using CHeF as Proxy Task. By considering the cluster indices provided by the AE model as the target class labels, the audio and visual CNN feature encoders are separately pre-trained to predict the cluster indices. For this purpose, the CNN feature encoder coupled with a 2-layer BiGRU-RNN (with 128 hidden units) and a fully connected output layer, is trained in an end-to-end fashion as a standard k-class classification

model using the general cross entropy loss. Given an unlabelled face image or Mel-spectrogram sequence as input, first its hand-engineered feature sequence is fed into the already trained AE model to predict its cluster index. Thus, the audio and visual CNN feature encoders are separately trained to predict the cluster indices provided by their corresponding CHeF models.

Dense Predictive Coding [Han et al., 2019] Baseline – A Generic Prior Guided SSL Method. Unlike the downstream task-aware SSL pre-training method proposed in this chapter, most existing SSL methods rely on generic representation learning priors. Predictability of sequential data is one such prior that has been widely exploited in the literature [Han et al., 2019, Lu et al., 2020, Shukla et al., 2020]. In this chapter, an SSL model based on predictive coding is implemented as a representative baseline for the generic prior guided SSL. Particularly, a state-of-the-art formulation of it based on Dense Predictive Coding (DPC) [Han et al., 2019] is trained and evaluated for both visual and audio data separately. A DPC [Han et al., 2019] model is composed of a CNN feature encoder and a BiGRU-RNN for predicting the last of half of input image sequence, given the first half of the sequence as input. Our implementation follows the same training methodology proposed in the original DPC framework. To ensure a fair comparison, we implement both visual and audio DPC models that are comparable with the CHeF models in terms of the total number of network weights. For more information on the DPC implementation and training details, the reader is referred to the original DPC implementation [1].

Optimisation Details Adam optimiser [Kingma and Ba, 2014] with the weight decay value set to 1e-4, is used for training all the models evaluated in this chapter. First, in the case of AE model, only during its pre-training

[1] https://github.com/TengdaHan/DPC

phase, the teacher forcing factor is set to 0.4 for the first 5 epochs, and 0.2 for the later 5 epochs. When training the AE model with the total loss function (see Eq. 6.4, the key hyper-parameters (k – number of clusters, α – reconstruction loss weight, β – clustering loss weight, γ – max-margin loss weight) are tuned based on the validation set performance on the downstream evaluation task i.e. emotion recognition. When the value of k is set to 6, the best results are achieved in both audio and visual emotion recognition cases. The dropout values in the GRU layers of the AE model are set to 0.3. The initial learning rate value is 5e-4, and it is tuned using a cosine annealing based scheduler with warm restarts enabled [Loshchilov and Hutter, 2016]. The batch size is set to 1024 sequences, with each sequence containing 100 frames, in the case of both the visual and audio AE models.

Details of the downstream task evaluation of the SSL pre-trained CNN encoders can be found in Appendix D.

Label-Efficiency Evaluation Protocol

Measuring the overall emotion predictive performance alone may fall short in comprehensively illustrating the advantages of a particular pre-training method. As the main objective of this chapter is to improve the label-efficiency of representation learning step in the emotion recognition models, an additional evaluation protocol is adopted in order to measure and compare the label-efficiency of different pre-training methods. In this evaluation protocol, only 10% of subject-wise randomly sampled labelled examples from SEWA (for visual models) and from AVEC'19 (for audio models) are used for training different emotion recognition models in which the CNNs' weights are initialised using different pre-training techniques.

Thus, a particular pre-training method that achieves the best predictive performance using just 10% of the training data can be considered as the most label-efficient method. The main hypothesis of this chapter is that CHeF-SSL pre-trained models should capture emotion-related representations more effectively than the SSL methods based on generic representation learning priors such as predictability. Hence, the performance of CHeF-SSL proxy task is expected to be significantly better than that of the Dense Predictive Coding baseline. As additional baselines, CNNs that are pre-trained directly on the emotion recognition task itself but on a different emotion-labelled dataset are included in this evaluation.

6.5 Results and Discussion

This section presents unimodal emotion recognition results of the proposed CHeF-SSL method, in comparison with different pre-training methods and the existing state-of-the-art benchmarks of emotion recognition from face and voice data. The experimental results presented here include the overall predictive performance as well as the label-efficiency analysis of different CNN pre-training methods on the tasks of facial and speech emotion recognition, as discussed below.

Q1. – ***Overall Emotion Predictive Performance Analysis*** – *How does the proposed SSL pre-training method perform when* ***fine tuned with all the emotion-labelled examples*** *available in the training set?*

Face Emotion Recognition. Table 6.1 presents valence and arousal recognition results of different CNN pre-training methods and the current state-of-the-art models on the SEWA test set. Among all the pre-training methods evaluated in Table 6.1, the proposed SSL framework, CHeF-AU-

Table 6.1: Face Emotion Recognition: SEWA Test Set Results

Model	**Valence CCC ↑**	**Arousal CCC ↑**	**Avg. CCC ↑**
SOTA: Aff. Proc. [Tellamekala et al., 2022]	0.739	**0.622**	**0.680**
SOTA: [Kossaifi et al., 2020]	**0.750**	0.520	0.635
Rand-init	0.508	0.461	0.485
2D Face Alignment-init	0.635	0.508	0.572
SSL-DPC-init [Han et al., 2019]	0.601	0.494	0.548
SSL-CHeF-AU-init	*0.720*	*0.615*	*0.668*
AffectNet-init	0.715	0.568	0.641

init, achieved the best results in terms of both valence CCC and arousal CCC scores. When compared to the SSL proxy task based on DPC-init, which relies on a standard generic representation learning prior i.e. temporal predictability, the proposed CHeF proxy task exhibited noticeably better emotion recognition performance. This trend clearly validates the main hypothesis of this chapter: making an SSL proxy task aware of the target downstream task's characteristics, is more effective than a proxy task that is completely downstream-agnostic. Most importnatly, considerable predictive performance gains achieved with the SSL-CHeF pre-trained model strongly indicate the importance of exploring downstream-specific representation learning priors in SSL.

Among the other pre-training baselines listed in Table 6.1, AffectNet-init model achieved the second best mean CCC score, which is obvious considering that the emotion information in static face images is explicitly used in AffectNet pre-trained models. It is interesting to note that the CHeF-init model, which implicitly uses the emotion information, showed slightly better recognition performance than the AffectNet-init model. This trend could be due to the fact that the temporal dynamics of affect are leveraged

Table 6.2: Speech Emotion Recognition: AVEC'19 Results

Model	**Valence CCC ↑**	**Arousal CCC ↑**	**Avg. CCC ↑**
SOTA: SPR-NPs [Mani Kumar et al., 2021]	**0.441**	**0.618**	**0.530**
Rand-init	0.366	0.515	0.440
AudioSet-init	0.428	0.559	0.493
SSL-DPC-init[Han et al., 2019]	0.381	0.534	0.457
SSL-CHeF-eGEMAPs-init	0.413	0.578	0.496
IEMOCAP-init	*0.419*	*0.585*	*0.502*

in the CHeF proxy task, unlike in the case of AffectNet-init model.

In terms of the mean CCC score, compared to the current state-of-the-art model on the SEWA test set, Affective Processes [Tellamekala et al., 2022], the proposed CHeF-init model has slightly poor recognition performance. Note that the stochastic temporal context modelling of emotion data used in Affective Processes is complementary to the improvements in the CNN pre-training methods. Thus, the performance of the proposed SSL pre-training method guided by the hand-crafted emotion features, can be significantly improved further by coupling it with the advanced temporal models such as Affective Processes. As the main focus of this chapter is on improving the SSL pre-trained representations for emotion recognition, integration of the proposed CHeF-init CNN models with Affective Processes is left for the future work.

Speech Emotion Recognition. Table 6.2 presents the results of emotion recognition models based on different pre-trained audio representation learning methods on the AVEC'19 validation set. Unlike in the case of face emotion recognition, in terms of mean CCC, the performance of SSL-CHeF-init model is found to be slightly worse than the IEMOCAP-init model, and

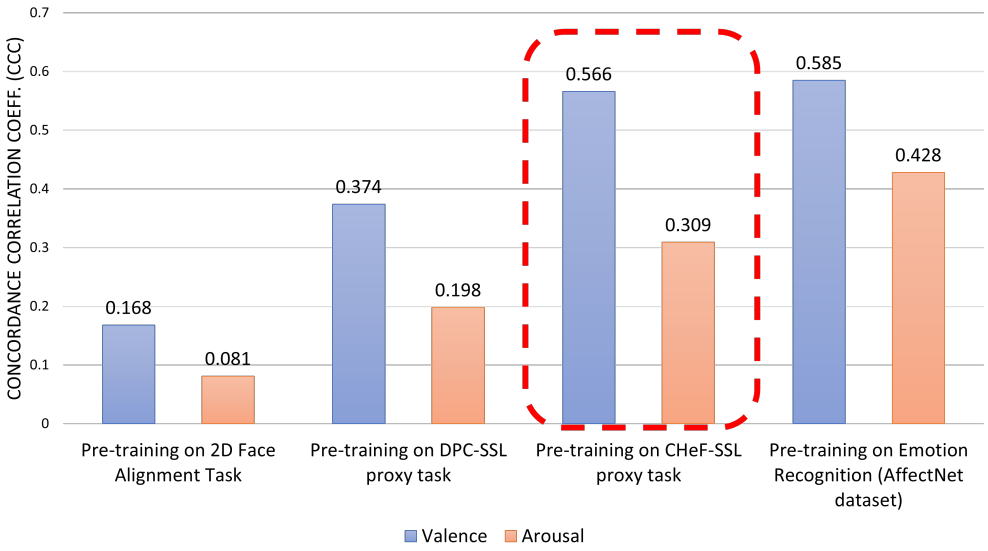


Figure 6.3: Label-efficiency results of the **visual-CHeF** models on the **SEWA validation set**: Emotion recognition performance of different CNN pre-training methods when finetuned using only *10%* of the total labelled data in the SEWA training set.

slightly better than the AudioSet-init models. But, in the case of arousal prediction, the CHeF-init model showed noticeably better results than the remaining models, except for the IEMOCAP-init model which makes use of the temporal emotion labels for the CNN pre-training. Most importantly, compared to the results of the SSL-DPC-init model which is the only other baseline trained on the unlabelled audio data, CHeF-init model exhibited considerably better results, validating the main hypothesis of this chapter.

Q2. – ***Label-Efficiency Analysis*** – *How does the proposed SSL pre-training method perform when* ***fine tuned with only a fraction of the emotion-labelled examples*** *available in the training set?*

Fig. 6.3 compares the performance of four different pre-training methods applied to the task of face emotion recognition on the SEWA corpus, using as few as 10% of the total number of labelled examples in the original training set. In this setting also, the model pre-trained on the proposed CHeF proxy task performed almost on par with the AffectNet-init pre-

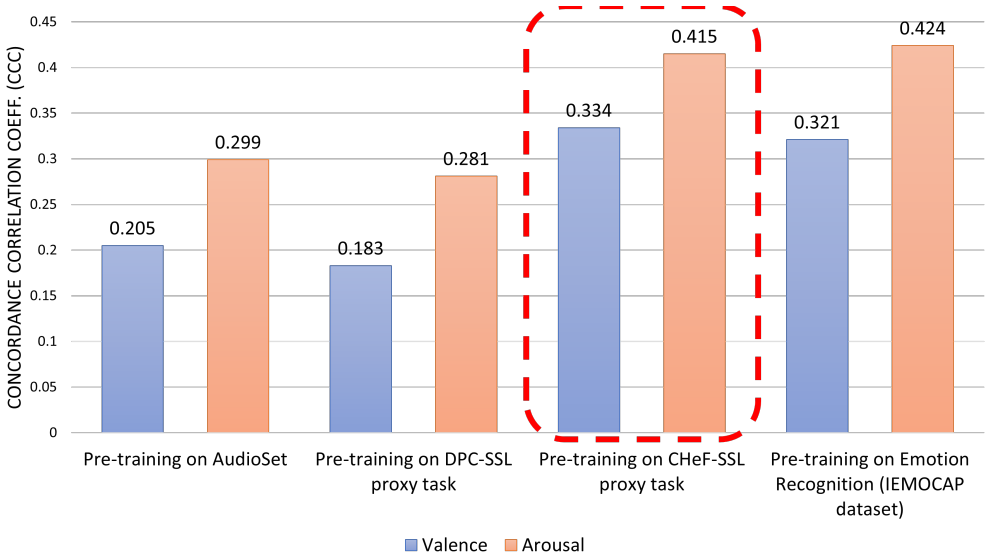


Figure 6.4: Label-efficiency results of the **audio-CHeF** models on the **AVEC'19** validation set: Emotion recognition performance of different CNN pre-training methods when finetuned using only *10%* of the total labelled data in the AVEC'19 training set.

trained baseline, which requires a large amount of emotion-labelled examples to achieve this level of recognition performance. Whereas the CHeF pre-trained model purely relies on the unlabelled video data to capture and induce emotion-related representations into the CNN feature encoder. Further, compared to the model pre-trained on the DPC proxy task, the proposed CHeF-init model performed noticeably better, particularly in the case of valence prediction. This trend clearly indicates that we can achieve substantial performance gains using SSL pre-training when its proxy task is made aware of the downstream task's characteristics, through *freely* available task-specific hand-engineered features. Thus, it demonstrates the potential of the proposed CHeF proxy task towards learning label-efficient emotion recognition by effectively making use of the large-scale unlabelled data. Note that the labelled examples used in this experiment cover only 10% of the subjects in the SEWA corpus, in order to mimic the conditions with few labelled data points from a small number of subjects.

Similarly, Fig. 6.4 presents the label-efficienct speech emotion recognition

results of different pre-training models on the AVEC'19 validation set. Akin to the results shown in its visual counter part, the CHeF pre-trained audio CNN model achieved superior emotion recognition performance, almost on par with the prediction results of IEMOCAP-init model. Further, in the case of valence prediction, the CHeF-init model showed the best results compared to the rest of the pre-training baselines. Here also, the DPC-init model's performance is considerably lower than the CHeF-init model, illustrating the insufficiency of generic representation learning priors like temporal predictability in effectively capturing emotion-related features directly from the unlabelled data. Overall, the results shown in Fig. 6.4 validate the modality-agnostic nature of the proposed SSL proxy task, CHeF, in learning the target downstream task using as few labelled examples as possible.

6.6 Conclusion

This chapter demonstrated that hand-crafted features in Machine Learning can be viewed as task-specific representation priors, given that they systematically encode domain expertise developed towards a specific learning task. With the success of end-to-end feature learning directly from raw data, hand-engineered features have been largely ignored in recent years, primarily due to their inferior generalisation performance compared to the deep representation learning. However, not much attention has been paid in the literature to the the idea of exploiting hand-engineered features to address the limitations of data-driven representation learning such as label-inefficiency. The self-supervised learning method proposed in this chapter, to the best of our knowledge, for the first time illustrated that we can exploit the rich task-specific information encoded in hand-crafted features to

the benefit of end-to-end representation learning.

With the objective of improving the label-efficiency of existing emotion recognition models, this chapter proposed a novel solution based on the idea that the hand-crafted features of emotion recognition can be leveraged as weak-supervision signals to pre-train the visual and audio feature encoders. This chapter illustrated a novel use case of emotion-related hand-engineered features in self-supervised pre-training of the feature encoder models. Particularly, in the absence of large sets of emotion labelled examples, the proposed solution based on SSL pre-training achieved good performance gains in facial and speech emotion recognition tasks, with minimal human supervision.

Chapter 7

Conclusions

7.1 Summary

This chapter summarises the key contributions of this thesis towards improving the current state of automatic affect recognition from face and voice data. First, this thesis identified two important challenges faced by the existing apparent affect recognition methods: *label ambiguity* and *label scarcity*. The former challenge arises due to the inherently ambiguous nature of manual affect annotations. Whereas the latter is caused by the prohibitively expensive nature of manual affect annotation process. To address these two challenges, this thesis proposed to build uncertainty-aware and label-efficient affect recognition models respectively. In particular, the solutions proposed in this thesis explored (a). non-deterministic function learning models for probabilistic temporal context modelling to cope with the one-to-many mapping nature of the affect labels and (b). label-efficient representation learning through self-supervised pre-training to minimise the requirement of manual affect annotations.

7.1.1 On Uncertainty-Aware Affect Recognition

Through probabilistic modelling of temporal affect context in face and voice data, different uncertainty modelling methods were presented in Chapter 3, 4 and 5. These methods demonstrated superior emotion recognition performance than the existing temporal models such as RNNs and self-attention, which are largely based on deterministic function learning models that completely ignore the label ambiguity problem. Additionally, the three key uncertainty-aware affect recognition methods proposed in this work: Calibrated and Ordinal Latent Distributions (COLD in Chapter 3), Affective Processes (APs in Chapter 4), and Epistemic-Aleatoric Uncertainty (EAU in Chapter 5), led to some promising applications. First, the COLD fusion showed that the audiovisual affect information fusion can be made more robust to the visual occlusions, most importantly without requiring any additional computational complexity. Second, APs demonstrated a novel application in cooperative machine learning which holds the potential to not only accelerate the affect annotation process but also improve the quality of the affect labels. Finally, the EAU model showed how to holistically capture the affect predictive uncertainty and its significance in improving the performance of downstream behavioural analysis tasks such as apparent personality recognition.

7.1.2 On Label-Efficient Affect Recognition

Considering that the existing affect recognition models based on deep representation learning require large amounts of affect labelled data, a novel self-supervised pre-training method is proposed in this thesis. The proposed pre-training method demonstrated how to leverage large amounts of unlabelled data in order to reduce the label requirement for learning affect

recognition models. In contrast to the existing self-supervised pre-training methods which heavily rely on generic representation learning priors, the method introduced in Chapter 6 proposed to apply affect-specific representation learning priors. To this end, hand-crafted emotion features of face and voice data were exploited as task-specific representation learning priors and a novel clustering-based self-supervised proxy task was introduced, dubbed as CHeF - Clustering Hand-engineered Emotion Features. Affect recognition models based on the proposed CHeF pre-training method, demonstrated considerably better label-efficiency than the commonly used pre-training techniques in the existing affect recognition methods. Most importantly, using as few as 10% of the labelled training data, the proposed CHeF framework showed promising emotion recognition results, on par with the pre-training baselines that require large amounts of labelled data from different emotion recognition corpora.

In summary, this thesis presented novel uncertainty-aware and label-efficient machine learning approaches that (a). account for the label-ambiguity problem of affect recognition tasks and (b). require fewer labelled examples for training respectively. A potential future direction is to integrate both these advancements, label-efficient feature encoders and probabilistic temporal models, into a single affect recognition model for improved reliability and recognition performance.

Bibliography

S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv preprint:1612.00410*, 2016.

S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller. Sequence to sequence autoencoders for unsupervised representation learning from audio. In *DCASE*, pages 17–21, 2017.

C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

K. Anderson and P. W. McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(1):96–105, 2006.

A. Anwar and A. Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

M. Atcheson, V. Sethu, and J. Epps. Gaussian process regression for continuous emotion recognition with global temporal invariance. In *IJCAI*

2017 Workshop on Artificial Intelligence in Affective Computing, pages 34–44. PMLR, 2017.

Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016.

A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

T. Baltrušaitis, N. Banda, and P. 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.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

P. Barros, D. Jirak, C. Weber, and S. Wermter. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, 72:140–151, 2015.

M. S. Bartlett, G. Littlewort, T. Sejnowski, and J. Movellan. A prototype for automatic recognition of spontaneous facial actions. *Advances in neural information processing systems*, 15, 2002.

M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. 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.

M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 223–230. IEEE, 2006.

Bekhouche et al. Personality traits and job candidate screening via analyzing facial videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 10–13, 2017.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

S. Bruch, X. Wang, M. Bendersky, and M. Najork. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 75–78, 2019.

A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030, 2017.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N.

Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

E. D. Carvalho, R. Clark, A. Nicastro, and P. H. Kelly. Scalable uncertainty for computer vision with functional variational inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12003–12013, 2020.

P. Chandran, D. Bradley, M. Gross, and T. Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *CVPR*, pages 5861–5870, 2020.

J. Chang, Z. Lan, C. Cheng, and Y. Wei. Data uncertainty learning in face recognition. In *CVPR*, pages 5710–5719, 2020.

J.-R. Chang, Y.-S. Chen, and W.-C. Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021.

Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli. Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In *AVEC*, pages 19–26, 2019.

L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion/expression recognition. In *Proceedings Third IEEE Interna-*

tional Conference on Automatic Face and Gesture Recognition, pages 366–371. IEEE, 1998.

S. Chen, X. Li, Q. Jin, S. Zhang, and Y. Qin. Video emotion recognition in the wild based on fusion of multimodal features. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 494–500, 2016.

S. Chen, Q. Jin, J. Zhao, and S. Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *AVEC*, pages 19–26, 2017.

H.-C. Chou and C.-C. Lee. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890. IEEE, 2019.

L. Christ, S. Amiriparian, A. Baird, P. Tzirakis, A. Kathan, N. Müller, L. Stappen, E.-M. Meßner, A. König, A. Cowen, et al. The muse 2022 multimodal sentiment analysis challenge: Humor, emotional reactions, and stress. 2022.

C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez. Addressing failure prediction by learning model confidence. *arXiv preprint arXiv:1910.04851*, 2019.

R. Cowie and R. R. Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1-2):5–32, 2003.

R. Cowie, E. Douglas-Cowie, S. 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*, 2000.

R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1): 1–17, 2012.

T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 27–35, 2017.

T. Dang, V. Sethu, and E. Ambikairajah. Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4929–4933. IEEE, 2018.

C. Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 1948.

B. De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

J. J. Del Coz, J. Díez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10(10), 2009.

F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1970–1973. IEEE, 1996.

D. Deng, Z. Chen, and B. E. Shi. Multitask emotion recognition with incomplete labels. In *IEEE FG*, pages 828–835.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A

large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Deng, Z. Zhang, E. Marchi, and B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

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

S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3): 1–36, 2015.

E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *Ninth European conference on speech communication and technology*, 2005.

S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Trans. on Affect. Comput.*, 2019.

S. Ebrahimi, M. Elhoseiny, T. Darrell, and M. Rohrbach. Uncertainty-guided continual learning with bayesian neural networks. *arXiv preprint arXiv:1906.02425*, 2019.

S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.

P. Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60): 16, 1999.

M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

A. Eldesokey, M. Felsberg, K. Holmquist, and M. Persson. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12014–12023, 2020.

H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güç, U. Güçlü, X. Baró, I. Guyon, J. C. Jacques, M. Madadi, et al. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 2020.

S. Escalera, X. Baró, H. J. Escalante, and I. Guyon. Chalearn looking at people: A review of events and resources. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1594–1601. IEEE, 2017.

G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and

fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

F. Eyben, M. Wöllmer, and B. Schuller. Openear—introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–6. IEEE, 2009.

F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE TAC*, 7(2):190–202, 2015.

B. Fasel. Head-pose invariant facial expression recognition using convolutional neural networks. In *Proceedings. Fourth IEEE international conference on multimodal interfaces*, pages 529–534. IEEE, 2002a.

B. Fasel. Robust face analysis using convolutional neural networks. In *Object recognition supported by user interaction for service robots*, volume 2, pages 40–43. IEEE, 2002b.

H. M. Fayek, M. Lech, and L. Cavedon. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 international joint conference on neural networks (IJCNN)*, pages 566–570. IEEE, 2016.

T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, and M. Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multi-

task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, 2021.

M. Fortunato, C. Blundell, and O. Vinyals. Bayesian recurrent neural networks. *arXiv preprint:1704.02798*, 2017.

N. M. Foteinopoulou, C. Tzelepis, and I. Patras. Estimating continuous affect with label uncertainty. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint:1903.10145*, 2019.

Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.

M. J. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami. Conditional neural processes. In *ICML*, pages 1704–1713, 2018a.

M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. W. Teh. Neural processes. In *ICML Workshop on*

Theoretical Foundations and Applications of Deep Generative Models, 2018b.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Y. Geng, Z. Han, C. Zhang, and Q. Hu. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7545–7553, 2021.

M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 2021.

A. Ghandeharioun, B. Eoff, B. Jou, and R. Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206. IEEE, 2019.

J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost. Progressive neural networks for transfer learning in emotion recognition. *arXiv preprint arXiv:1706.03256*, 2017.

J. Gordon, W. P. Bruinsma, A. Y. Foong, J. Requeima, Y. Dubois, and R. E. Turner. Convolutional conditional neural processes. In *International Conference on Learning Representations*, 2020.

R. L. Gregory. The medawar lecture 2001 knowledge for vision: Vision for knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1458):1231–1251, 2005.

M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 381–385. IEEE, 2005.

P. W. Große, H. Holzapfel, and A. Waibel. Confidence based multimodal fusion for person identification. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 885–888, 2008.

Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European conference on computer vision*, pages 349–358. Springer, 2016.

H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 827–834. IEEE, 2011.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

G. Guo and C. R. Dyer. Learning from examples in the small sample case: face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):477–488, 2005.

J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the

perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897, 2017.

J. Han, Z. Zhang, Z. Ren, and B. Schuller. Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, pages 1–10, 2020.

J. Han, Z. Zhang, Z. Ren, and B. Schuller. Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, 13:231–240, 2021.

T. Han, W. Xie, and A. Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshop*, 2019.

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

Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019.

O. J. Hénaff, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE, 2017.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

P. Hu, S. Sclaroff, and K. Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. In *NeurIPS*, 2020.

J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi. End-to-end continuous emotion recognition from video using 3d convlstm networks. In *ICASSP*, pages 6837–6841. IEEE, 2018.

J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu. Efficient modeling of long temporal contexts for continuous emotion recognition. In *ACII*, pages 185–191. IEEE, 2019.

Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *WACV*, pages 1–8. IEEE, 2016.

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

N. Jaitly and G. Hinton. Learning a better representation of speech soundwaves using restricted boltzmann machines. In *2011 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5884–5887. IEEE, 2011.

L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017a.

A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, pages 5574–5584. Curran Associates, Inc., 2017b. URL http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-comput pdf.

S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019.

A. Khare, S. Parthasarathy, and S. Sundaram. Self-supervised learning with cross-modal transformers for emotion recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 381–388. IEEE, 2021.

H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum,

O. Vinyals, and Y. W. Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2018.

Y. Kim and J. Kim. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5104–5108. IEEE, 2018.

Y. Kim and E. M. Provost. Leveraging inter-rater agreement for audio-visual emotion recognition. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 553–559. IEEE, 2015.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

H. Kobayashi and F. Hara. The recognition of basic facial expressions by neural network. *Transactions of the society of instrument and control engineers*, 29(1):112–118, 1993.

D. Kollias and S. Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint:1811.07770*, 2018a.

D. Kollias and S. Zafeiriou. A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. *arXiv preprint:1805.01452*, 2018b.

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

D. Kollias and S. P. Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Trans. on Affect. Comput.*, 2020.

D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou. Recognition of affect in the wild using deep neural networks. In *CVPR Worksh.*, pages 1972–1979. IEEE, 2017.

D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vis.*, pages 1–23, 2019.

D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. *arXiv preprint:2001.11409*, 2020.

J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3):1022–1040, 2019.

J. Kossaifi, A. Toisoul, A. Bulat, Y. Panagakis, T. M. Hospedales, and M. Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *CVPR*, pages 6060–6069, June 2020.

R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69, 2017.

I. Kotsia and I. Pitas. Facial expression recognition in image sequences

using geometric deformation features and support vector machines. *IEEE transactions on image processing*, 16(1):172–187, 2006.

R. Krishnan and O. Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *arXiv preprint arXiv:2012.07923*, 2020.

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

F. Kuhnke, L. Rumberg, and J. Ostermann. Two-stream aural-visual affect analysis in the wild. In *IEEE FG*, pages 366–371.

V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.

A. Kumar, S. Eslami, D. J. Rezende, M. Garnelo, F. Viola, E. Lockhart, and M. Shanahan. Consistent generative query networks. *arXiv preprint:1807.02033*, 2018a.

A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018b.

Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017.

G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6874–6883, 2017.

I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

T. A. Le, H. Kim, M. Garnelo, D. Rosenbaum, J. Schwarz, and Y. W. Teh. Empirical evaluation of neural process objectives. In *NeurIPS Worksh.*, 2018.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

J. E. LeDoux and S. G. Hofmann. The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, 19:67–72, 2018.

J. Lee, Y. Lee, J. Kim, E. Yang, S. J. Hwang, and Y. W. Teh. Bootstrapping neural processes. *NeurIPS*, 33, 2020.

L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 312–317. IEEE, 2013.

S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2021.

Y. Li, J. Zeng, and S. Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):302–317, 2020a.

Y. Li et al. Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, pages 1–18, 2020b.

R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

W. Lim, D. Jang, and T. Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE, 2016.

F. Lingenfelser, J. Wagner, J. Deng, R. Brueckner, B. Schuller, and E. André. Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches. *IEEE Transactions on Affective Computing*, 9(4):410–423, 2016.

J. Liscombe, J. Venditti, and J. B. Hirschberg. Classifying subject ratings of emotional speech using acoustic features. 2003.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint:1608.03983*, 2016.

L. Lu, L. Tavabi, and M. Soleymani. Self-supervised learning for facial action unit recognition through temporal consistency. In *BMVC*, 2020.

J. Lucas, G. Tucker, R. Grosse, and M. Norouzi. Don't blame the elbo! a

linear vae perspective on posterior collapse. *arXiv preprint:1911.02469*, 2019a.

J. Lucas, G. Tucker, R. Grosse, and M. Norouzi. Understanding posterior collapse in generative latent variable models. *ICLR Workshop*, 2019b.

S. Lucey, A. B. Ashraf, and J. F. Cohn. *Investigating spontaneous facial action recognition through aam representations of the face*, volume 2. INTECH Open Access Publisher, 2007.

A. Mallol-Ragolta, N. Cummins, and B. W. Schuller. An investigation of cross-cultural semi-supervised learning for continuous affect recognition. *Proc. Interspeech 2020*, pages 511–515, 2020.

T. Mani Kumar, E. Sanchez, G. Tzimiropoulos, T. Giesbrecht, and M. Valstar. Stochastic process regression for cross-cultural speech emotion recognition. *Proc. Interspeech 2021*, pages 3390–3394, 2021.

B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Trans. on Affect. Comput.*, 2017.

K. Mase. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483, 1991.

G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1079–1084. IEEE, 2010.

A. Mehrabian. Communication without words. In *Psychology Today*, volume 2, pages 53–56, 1968.

H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas. Time-delay neural network for continuous emotional dimension pre-

diction from facial expression sequences. *IEEE transactions on cybernetics*, 46(4):916–929, 2015.

M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic. Valence and arousal estimation in-the-wild with tensor methods. In *IEEE FG*, pages 1–7. IEEE, 2019.

J. Mitros and B. Mac Namee. On the validity of bayesian neural networks for uncertainty estimation. *arXiv preprint arXiv:1912.01530*, 2019.

A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

J. Moon, J. Kim, Y. Shin, and S. Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR, 2020.

E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE, 2022.

E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009.

J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

M. Neumann and N. T. Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE, 2019.

H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

M. A. Nicolaou, H. Gunes, and M. 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.

F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2017.

I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, and Y. Tzimiropoulos. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Transactions on Affective Computing*, 2021.

M. Pantic and M. S. Bartlett. *Machine analysis of facial expressions*. INTECH Open Access Publisher, 2007.

M. Pantic and L. J. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

M. Pantic and L. J. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, 2004.

G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

F. I. Parke. *A parametric model for human faces.* The University of Utah, 1974.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

E. Pei, Y. Zhao, M. C. Oveneke, D. Jiang, and H. Sahli. A bayesian filtering framework for continuous affect recognition from facial images. *IEEE Transactions on Multimedia*, 2022.

J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

R. W. Picard. Affective computing mit press. *Cambridge, Massachsusetts*, page 2, 1997.

S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 536–543, 2017.

J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

R. Plutchik and H. Kellerman. *Theories of emotion*, volume 1. Academic Press, 2013.

V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision*, pages 400–418. Springer, 2016.

S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li. Global ranking using continuous conditional random fields. *Advances in neural information processing systems*, 21, 2008.

L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

B. Reeves and C. Nass. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10:236605, 1996.

F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.

F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.

F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *AVEC*, pages 3–13, 2018.

F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019.

G. Rizos and B. Schuller. Modelling sample informativeness for deep affective computing. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3482–3486. IEEE, 2019.

G. Rizos and B. W. Schuller. Average jane, where art thou?–recent avenues in efficient machine learning under subjectivity uncertainty. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 42–55. Springer, 2020.

R. Roady, T. L. Hayes, R. Kemker, A. Gonzales, and C. Kanan. Are out-of-distribution detection methods effective on large-scale datasets? *arXiv preprint arXiv:1910.14034*, 2019.

P. V. Rouast, M. Adam, and R. Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 2019.

S. Roy and A. Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 253–257, 2021.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

P. Salovey and J. D. Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.

E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos. Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *CVPR*, 2021.

E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

M. Schmitt, N. Cummins, and B. W. Schuller. Continuous emotion recognition in speech—do we need recurrence? *Proc. Interspeech 2019*, pages 2808–2812, 2019.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

L. Schoneveld, A. Othmani, and H. Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021.

A. Schörgendorfer and W. Elmenreich. *Extended confidence-weighted averaging in sensor fusion.* na, 2006.

B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The inter-

speech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

T. Senechal, K. Bailly, and L. Prevost. Impact of action unit detection in automatic emotion recognition. *Pattern Analysis and Applications*, 17 (1):51–67, 2014.

V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*, 2019.

J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021.

A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Visually guided self supervised learning of speech representations. *ICASSP*, 2020.

A. Shukla, S. Petridis, and M. Pantic. Does visual self-supervision improve learning of speech representations for emotion recognition. *IEEE Transactions on Affective Computing*, 2021.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

G. Singh, J. Yoon, Y. Son, and S. Ahn. Sequential neural processes. In *NeurIPS*, pages 10254–10264, 2019.

H. Song, T. Diethe, M. Kull, and P. Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906, 2019.

S. Song, S. Jaiswal, L. Shen, and M. Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 2020.

S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar. Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *IEEE Transactions on Affective Computing*, 2021.

K. Sridhar and C. Busso. Ensemble of students taught by probabilistic teachers to improve speech emotion recognition. In *INTERSPEECH*, pages 516–520, 2020.

L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, et al. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 35–44, 2020.

L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller. The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 5–14. 2021.

M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, and J. Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian

variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2019.

S. Sun, C. Chen, and L. Carin. Learning structured weight uncertainty in bayesian neural networks. In *AISTATS*, pages 1283–1292. PMLR, 2017.

X. Sun, J. Zeng, and S. 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.

C. Sutton, A. McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

Suwa. A preliminary note on pattern recognition of human emotional expression. *Proc. of The 4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013a.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013b.

P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017.

M. K. Tellamekala and M. 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.

M. K. Tellamekala, E. Sanchez, M. Valstar, and G. Tzimiropoulos. Stochastic process regression for cross-cultural speech emotion recognition. In *INTERSPEECH*, pages 3390–3394, 2021.

M. K. Tellamekala, T. Giesbrecht, and M. Valstar. Modelling stochastic context of audio-visual expressive behaviour with affective processes. *IEEE Transactions on Affective Computing*, (01):1–1, 2022.

J. Tian, W. Cheung, N. Glaser, Y.-C. Liu, and Z. Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE, 2020.

A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

M. Trapp, R. Peharz, F. Pernkopf, and C. E. Rasmussen. Deep structured mixtures of gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2251–2261. PMLR, 2020.

V. Tresp. Mixtures of gaussian processes. In *NeurIPS*, pages 654–660, 2001.

G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP*, pages 5200–5204. IEEE, 2016.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal

language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8): 1301–1309, 2017.

S. Utpala and P. Rai. Quantile regularization: Towards implicit calibration of regression models. *arXiv preprint arXiv:2002.12860*, 2020.

M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 1, pages 635–640. IEEE, 2004.

M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

M. F. Valstar. Timing is everything: A spatio-temporal approach to the analysis of facial actions. PhD Thesis, Imperial College London, 2008.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,

Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

J. Wagner, E. Andre, F. Lingenfelser, and J. Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.

J. Wagner, T. Baur, Y. Zhang, M. F. Valstar, B. Schuller, and E. André. Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora. *arXiv preprint:1802.02565*, 2018.

J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *arXiv preprint arXiv:2203.07378*, 2022.

D.-B. Wang, L. Feng, and M.-L. Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34, 2021.

K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson. Exact gaussian processes on a million data points. In *NeurIPS*, pages 14648–14659, 2019.

K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020.

S.-H. Wang and C.-T. Hsu. Ast-net: An attribute-based siamese temporal network for real-time emotion recognition. In *BMVC*, 2017.

X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315, 2017.

J. Whitehill and C. W. Omlin. Haar features for facs au recognition. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 5–pp. IEEE, 2006.

C. E. Williams and K. N. Stevens. Emotions and speech: Some acoustical correlates. *The journal of the acoustical society of America*, 52(4B): 1238–1250, 1972.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

C.-H. Wu, Y.-M. Huang, and J.-P. Hwang. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology*, 47(6):1304–1323, 2016.

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

E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin. Affective computing in education: A systematic review and future research. *Computers & Education*, 142:103649, 2019.

J. Yang, A. Bulat, and G. Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 34, pages 12621–12628, 2020.

X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5455, 2017.

G. N. Yannakakis. Enhancing health care via affective computing. 2018.

G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 12(1):16–35, 2018.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola. Deep sets. In *NIPS*, pages 3394–3404, 2017.

Z. Zeng, J. Tu, M. Liu, and T. S. Huang. Multi-stream confidence analysis for audio-visual affect recognition. In *International Conference on Affective Computing and Intelligent Interaction*, pages 964–971. Springer, 2005.

Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson. Audio-visual affect recognition. *IEEE Transactions on multimedia*, 9 (2):424–428, 2007.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31 (1):39–58, 2008.

L. Zhang, S. Peng, and S. Winkler. Persemon: a deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing*, 2019.

S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28 (10):3030–3043, 2017a.

Y. Zhang, R. Huang, J. Zeng, and S. Shan. M3f: Multi-modal continuous valence-arousal estimation in the wild. In *IEEE FG*, pages 617–621.

Y. Zhang, Y. Liu, F. Weninger, and B. Schuller. Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4990–4994. IEEE, 2017b.

Y. Zhang, C. Wang, and W. Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34, 2021.

J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin. Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 37–45, 2019.

Appendices

Appendix A

COLD Fusion: Network Architectures and Optimisation Details

A.0.1 Network Architectures

Visual CNN Backbone. EmoFAN [Yang et al., 2020], a 2D CNN proposed recently for facial feature extraction, is proved highly efficient by building on hour-glass-based network architectures. This CNN backbone, as illustrated in Figure B.2, when pretrained on 2D face alignment task, has been found very efficient for transfer learning tasks [Toisoul et al., 2021, Ntinou et al., 2021]. This work uses its pretrained model[1] on image-based emotion recognition on the AffectNet dataset [Mollahosseini et al., 2017]. Using this backbone, a 512D feature vector is extracted per frame.

Audio CNN Backbone. For speech feature extraction a 2D CNN back-

[1]Pretrained models of Toisoul et al. [Toisoul et al., 2021] are available at https://github.com/face-analysis/emonet

bone is used, which was originally proposed in [Chen et al., 2019] for audio features in an end-to-end fashion. This backbone based on a VGGish [Hershey et al., 2017] pre-trained module is applied to 2D Mel-spectrograms that are derived by setting the hop size and window length values to 0.1 s and 1 s respectively. Similar to [Chen et al., 2019], only the last two fully connected layers of this VGGish module are fine-tuned. To differentiate the interlocutor's information from that of the target speaker, the feature dimensionality-doubling technique [Chen et al., 2017] is adopted.

Temporal Networks are stacked on top of the unimodal CNN backbones to model the temporal dynamics and integrate the multimodal affect information. Note that all the fusion models evaluated in this work follow different temporal network implementations. However, all the temporal networks have the following GRU block in common: a 2-layer bidirectional GRU module followed by a fully connected (FC) output layer. This GRU block contains 256 hidden units with the dropout value set to 0.5. The number of GRU blocks and their input-output dimensionality vary across different fusion models, as discussed below.

In feature fusion, a single GRU+FC block is used to process the input feature sequence that is prepared via frame-wise concatenation of the unimodal embeddings, whereas, in the prediction fusion, different unimodal temporal models (GRU+FC) are applied separately, and their output softmax label distributions are aggregated into the final predictions. The context fusion implementation has two different GRU blocks, but a common FC layer. As shown in Figure 3.2, COLD fusion is similar to the context fusion, but with the GRU block's output layer modified to predict the mean and variance vectors. Note that the unimodal output branches are trained simultaneously along with the fusion branch in all the multimodal models (see Figure 3.2).

Data Augmentation. Strong data augmentation techniques are applied to the audiovisual inputs to minimise the overfitting problem. It is important to note that under heavy overfitting, the COLD loss function (Equation (3.6)) may collapse, since the calibration and ordinality constraints rely on the prediction errors of the training instances. For face image data augmentation, the following techniques are used: horizontal flipping with the probability set to 0.5, random scaling by a factor of 0.25, random translation by +/- 30 pixels, and random rotation by 30°. In the audio case, SpecAugment [Park et al., 2019] is applied, which directly augments the 2D spectrogram itself, instead of its original 1D waveform. Here, the standard SpecAugment operations are applied: time warping, frequency masking and time masking, with their order defined arbitrarily. The parameters[2] of time warping (ω), frequency masking (f), and time masking (t) are chosen from different uniform distributions in the range [0, 50], [0,27], and [0,40] respectively.

A.0.2 Optimisation Details

Input sequences of 30 seconds duration with per-frame targets are used. The visual and audio backbones and all the fusion models are trained using the Adam optimiser [Kingma and Ba, 2014] by jointly minimising the CCC loss [Kossaifi et al., 2020] and mean squared error for the regression task and class-weighted cross-entropy loss for the classification task. The batch size, learning rate, and weight decay values chosen for training all these models are 4, 5e-3, and 1e-4, respectively. For tuning the learning rate, Cosine annealing coupled with warm restarts [Loshchilov and Hutter, 2016] is used with the number of epochs for the first restart set to 1 and

[2] ω – warping length, f – number of consecutive mel frequency channels masked, t – number of consecutive time steps masked

the multiplication factor set to 2. The hyper-parameter values in the loss function (Eq. 3.6) are tuned on the logarithmic scale in the range [1e-5, 1e+5] using RayTune [Liaw et al., 2018]. Based on the best validation set performance, the following values are found to be optimal: 1e-3 for λ_{CO_V}, λ_{CO_A} and $\lambda_{CO_{AV}}$, and 1e-4 for λ_R.

Appendix B

Affective Processes: Datasets, Network Architectures, Backbones and Baselines

B.0.1 Datasets

Datasets To evaluate APs on visual dimensional affect recognition task, in addition to the AVEC 2019 CES [Ringeval et al., 2019] corpus, this work used an in-the-wild video dataset annotated with valence and arousal dimensions: basic SEWA [Kossaifi et al., 2019]. For audio-only and audio-visual affect recognition tasks, the AVEC 2019 CES dataset is used, similar to the COLD fusion.

SEWA [Kossaifi et al., 2019] basic dataset[1] contains 538 short (10s-30s) videos collected from 398 subjects of six different cultures. All the recordings are annotated with valence and arousal ratings by five different raters at 50 frames per second. This work used the same training, validation and

[1]https://db.sewaproject.eu/

test sets, containing 431, 53, and 53 videos respectively, [2] that were used in [Kossaifi et al., 2020].

AVEC'19 CES Corpus [Ringeval et al., 2019] is used for the audio-only and audio-visual emotion recognition experiments. This dataset is designed for cross-cultural in-the-wild affect recognition tasks by capturing audio-visual recordings of interactions between pairs of individuals from German, Hungarian and Chinese cultures. It provides 64 videos for training and 32 audio-visual recordings for validation (both from German and Hungarian cultures), with the video streams recorded at 50 FPS, the audio data recorded at 48 kHZ and the ratings of valence and arousal annotated at 10 FPS. As the labels of test sets and Chinese culture are not publicly available, the evaluation results are reported only on the validation set.

B.0.2 Network Architectures

Backbone Models and Baselines

Different unimodal and multimodal backbones are trained for the dimensional affect recognition. Note that all the backbone models are trained using ground truth labels as the targets, and the predictions from these backbones are referred to as proxy labels.

Visual Backbones. For visual dimensional affect recognition on SEWA, two different static CNNs are used: ResNet-50 [He et al., 2016] and Emo-FAN [Bulat and Tzimiropoulos, 2017, Ntinou et al., 2021, Toisoul et al., 2021]. By training and evaluating different visual AP models using these two backbones, this work aims to verify the generality of APs effectiveness regardless of the underlying backbone model complexity.

[2] The partition details are kindly provided by the database owners.

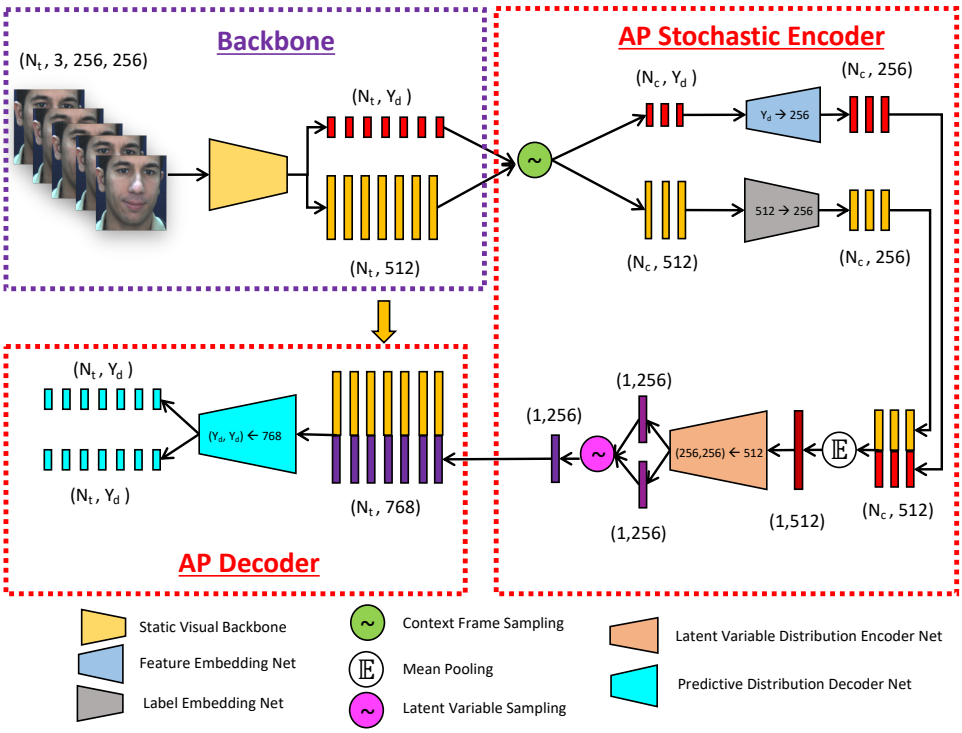


Figure B.1: Visual AP network (N_t and N_c denote the number of target and context frames respectively, and X_d and Y_d denote the dimensionality of features and labels respectively).

ResNet-50 [He et al., 2016] pre-trained on VGG-Face database [Parkhi et al., 2015] is extensively used for facial feature extraction in the existing temporal affect recognition models [Kollias et al., 2017, Kumar et al., 2018a]. In this implementation, its last convolution layer output feature maps are flattened into a 512 dimensional feature embedding. These features are fed into a 3-layer FC network (with 256 hidden units and 2 output units) to output the proxy labels.

EmoFAN backbone [Toisoul et al., 2021, Ntinou et al., 2021, Yang et al., 2020] comprises a feature extraction module designed for fine-grained facial analysis through pre-training on 2D face alignment task. As shown in Figure B.2, it includes a dimensional affect head composed of only convolutional layers for producing the dimensional affect predictions and image feature embeddings. To further improve the quality of facial features, this backbone is initialised with the weights of a model pre-trained[3] on AffectNet dataset [Mollahosseini et al., 2017].

Audio Backbones. In the AVEC19 CES, deep representation learning methods [Chen et al., 2019] demonstrated significantly better performance than the hand-crafted features [Ringeval et al., 2019]. Hence, this work adopted the audio feature learning method proposed in [Chen et al., 2019] for evaluating the $\mathcal{SPR}$-NP model. In this method, 128-dimensional features are extracted by applying VGGish [Hershey et al., 2017] pre-trained network to Mel-spectrogram images of the input audio signals (hop size and window length values set to 0.1s and 1s respectively). To differentiate the target speaker's features from the interlocutor's features, this work followed dimensionality-doubling strategy proposed in [Chen et al., 2017]. Note that while training this backbone model, only the last two fully connected layers

[3]Pretrained models of Toisoul et al [Toisoul et al., 2021] available at https://github.com/face-analysis/emonet

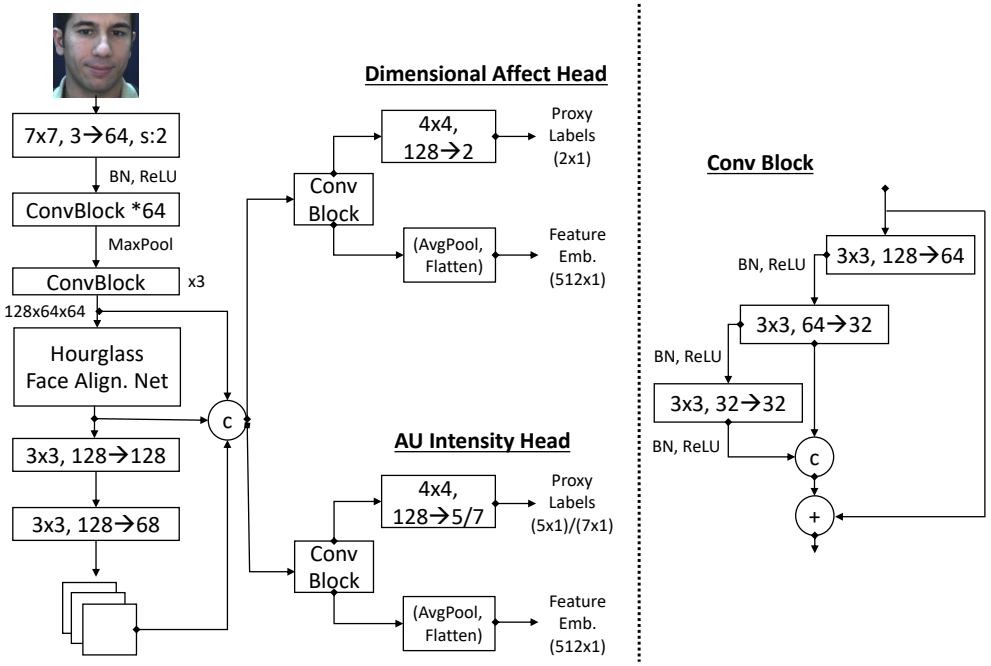


Figure B.2: EmoFAN backbone architecture [Toisoul et al., 2021, Ntinou et al., 2021, Yang et al., 2020] used for visual dimensional affect estimation.

of the VGGish pre-trained network are fine-tuned.

Audio-Visual Backbones. The network architecture of the audio-visual backbone is same as that of AVEC'19 CES challenge winners [Zhao et al., 2019], except that this work used EmoFAN backbone for the visual feature extraction. By following a simple uniformly weighted feature fusion strategy, the concatenated image and audio features are fed into a temporal regression model (a 3-layer GRU with 256 hidden units followed by one FC output layer) to produce the proxy labels.

Deterministic Temporal Regression Baselines. Two standard temporal models are trained on top of the EmoFAN backbone: a 2-layer bidirectional GRU network (with 256 hidden units) and a 3-layer self-attention [Vaswani et al., 2017] network with 16 multi-attention heads, followed by an output FC layer.

Audio-Visual Fusion Baselines. Three standard audio-visual feature

fusion baselines are evaluated in this work. First one uses uniformly weighted fusion (same as in the audio-visual backbone), and the remaining two are based on (a). globally weighted and (b). locally weighted fusion techniques. In the former, per-sequence weight vectors are used for fusing the audio and visual feature sequences, while the latter model uses per-frame weight vectors. Here, the dimensionality of unimodal weight vectors is same as that of their corresponding unimodal feature embeddings. To predict the weight vectors in both cases, first the unimodal feature sequences are passed through two different 1-layer GRU blocks with 256 hidden units. Per-sequence weight vectors are inferred by concatenating the last time step hidden vectors of the audio and visual GRUs and feeding the resultant vector into an FC layer. To predict the per-frame weight vectors, all time steps' hidden vectors from the audio and visual GRUs are concatenated and then passed through the same FC layer.

Multimodal Transformer. In addition to the above discussed instance-level fusion models, based on pair-wise crossmodal self-attention proposed in Tsai et al. [Tsai et al., 2019], a multimodal transformer is implemented as an additional fusion baseline. An audio-visual version of this transformer[4] is constructed by tailoring its original network architecture designed for text, audio and visual modalities. In this implementation, similar to the self-attention based deterministic temporal regression models, the network is composed a 3-layer self-attention network with 16 heads followed by an FC output layer.

[4]https://github.com/yaohungt/Multimodal-Transformer

B.0.3 Training of Backbones and Baselines

Image and Audio Features. The visual backbone models provide 512 dimensional feature vectors as face image representations used in both visual-only and audio-visual affect recognition tasks. To train the audio-only and audio-visual affect recognition models on AVEC'19 CES corpus, this work followed the same audio feature extraction method used by the corresponding challenge winners [Zhao et al., 2019]. In the case of audio data, the backbone model outputs a 256 dimensional feature vector per audio frame. Note that the VGGish backbone applied to input audio frame provides only a 128 dimensional vector, but the dimensionality doubling strategy transforms it into 256 dimensional vector to indicate the presence of interlocutor in the input audio frame.

Optimisation Details. All backbones and baselines are trained using Adam optimizer [Kingma and Ba, 2014] to minimise the inverse Concordance Correlation Coefficient (CCC) loss (1.0 - CCC) is used in addition to MSE [Kossaifi et al., 2020] for dimensional affect recognition. The visual backbones are trained on individual frames, with the batch size set to 32, and initial learning rate and weight decay values set to 1e-4 and 1e-5 respectively. When training the deterministic temporal baselines (BiGRUs and self-attention) on top of the static visual backbones, frame sequences are used as inputs with the sequence length set to 70 frames. Here, the batch size is 8, the learning rate and weight decay values are 5e-5 and 1e-4 respectively.

To train the audio and audio-visual backbones, and different fusion baselines, input sequences of 20 seconds duration are prepared, with 200 frames. The batch size, learning rate and weight decay values used for training all these models are 4, 5e-4 and 1e-4 respectively. For tuning the learning rate

in all different cases, Cosine annealing coupled with warm restarts [Loshchilov and Hutter, 2016] is used with the number of epochs for the first restart set to 1 and the multiplication factor set to 2. It is noticed that this warm restart technique stabilises the model training and also minimises the hyper parameter tuning iterations.

Appendix C

A Holistic Uncertainty Model of Temporal Affect: Datasets, Evaluation Metrics, and Backbone CNN Implementation

C.0.1 Datasets

For Dimensional Emotion Recognition this work used SEWA [Kossaifi et al., 2019], a large-scale continuous affect recognition dataset composed of 538 videos (10s-30s) collected from 398 subjects of 6 different cultures. Training, validation and test sets of SEWA contain 431, 53 and 53 videos respectively[1]. Each video is annotated with per-frame valence and arousal values in the range [-1, 1] at 50 frames per second.

[1] video ids of these sets were kindly provided by the dataset owners.

For Personality Traits Estimation, this work used ChaLearn [Ponce-López et al., 2016, Escalera et al., 2017], an in-the-wild database containing 10,000 clips (with the total duration of 41.6 hours and 4.5M frames) sourced from YouTube videos. Each clip contains only one subject and it is annotated with per-sequence apparent Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN). Each personality trait is annotated as a continuous value in the range [0, 1]. The standard train, validation and test partitions containing 6000, 2000 and 2000 videos respectively, were used.

C.0.2 Evaluation Metrics

For Dimensional Emotion Recognition this work used Lin's Concordance Correlation Coefficient (CCC) [Lawrence and Lin, 1989], a standard evaluation metric used for measuring the agreement between ground truth labels y^* and model predictions y^o.

$$CCC = \frac{\rho_{y^*y^o}.\sigma_{y^*}.\sigma_{y^o}}{\left(\mu_{y^*} - \mu_{y^o}\right)^2 + {\sigma_{y^*}}^2 + {\sigma_{y^o}}^2} \tag{C.1}$$

where $\rho_{y^*y^o}$ denotes the correlation coefficient between y^* and y^o, and (μ_{y^*}, μ_{y^o}) and $(\sigma_{y^*}, \sigma_{y^o})$ denote the mean and standard deviation values of y^* and y^o.

For Personality Recognition, following the existing works [Güçlütürk et al., 2016, Wei et al., 2017, Song et al., 2021], three evaluation metrics were used: Pearson's Correlation Coefficient (PCC), Root Mean Square Error (RMSE) and, mean Accuracy (Acc) [Ponce-López et al., 2016] for each trait,

$$PCC = \frac{cov(p^*, p^o)}{\sigma_{p^*}\sigma_{p^o}} \tag{C.2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p^* - p^o)^2} \tag{C.3}$$

$$Acc = 1 - \frac{1}{N}\sum_{i=1}^{N}|p^* - p^o| \tag{C.4}$$

where p^* and p^o denote the ground truth and predicted values of a personality trait, and N is the number of videos.

C.0.3 Backbone CNN for Face Image Feature Extraction

To extract low-dimensional features from face image sequences in both emotion recognition and personality recognition tasks, this work used a 2D backbone CNN proposed in [Toisoul et al., 2021] which demonstrated impressive generalisation performance on the emotion recognition task. Unlike the commonly employed ResNet based backbones pretrained on VGG Face dataset (face recognition task) [Kollias et al., 2019, Tellamekala and Valstar, 2019], the backbone CNN proposed in [Toisoul et al., 2021] exploits the facial features learned for 2D face alignment task through transfer learning. For extracting facial features with better generalisation capacity, this backbone CNN pretrained[2] on AffectNet dataset [Mollahosseini et al., 2017] was used for emotion recognition from static face images. The backbone CNN was fine tuned along with the other modules of emotion recognition and personality recognition. This model takes as input a 2D face image of dimensions 256x256x3 and outputs a 1D embedding of dimensions 256x1. During training, the following augmentations were applied: random

[2]Pretrained models of Toisoul et al [Toisoul et al., 2021] available at https://github.com/face-analysis/emonet

translation by +/- 20 pixels, random rotation by 30^o, random scaling by a factor of 0.25, and horizontal flipping with a probability of 0.5.

Appendix D

CHEF Experiments: Datasets, Network Architectures, Backbones and Baselines

D.0.1 Network Architectures

Visual CNN Feature Encoder Network is implemented based on the model of EmoFAN [Toisoul et al., 2021, Sanchez et al., 2021], a 2D CNN designed for facial feature extraction using only convolution layers. In this work, a pre-training baseline of this model trained on 2D face alignment tasks included as a baseline. An additional pre-training baseline of EmoFAN is also considered, in which the model is pre-trained on image-based emotion recognition using the AffectNet dataset [Mollahosseini et al., 2017]. Following prior works [Toisoul et al., 2021, Sanchez et al., 2021], this work also modifies the EmoFAN output layer in order to extract 512-dimensional facial embedding vectors.

Audio CNN Feature Encoder Network is implemented by adopting a

deep acoustic feature learning method originally proposed in [Chen et al., 2019]. This network is based on a 2D CNN model, dubbed as VGGish [Hershey et al., 2017] network, which operates on the 2D Mel-spectrogram images of the input audio signals (hop size and window length values set to 0.1s and 1s respectively). Given the Mel-spectrogram image of an audio segment as input, this encoder outputs a 128-dimensional feature vector. To differentiate the target speaker's features from the interlocutor's features, this work followed dimensionality-doubling strategy proposed in [Chen et al., 2017]. Similar to the 2D face alignment and AffectNet pre-training baselines of the visual CNN model, two baselines of this network are included in the experimental analysis. These baselines are pre-trained using the audio event recognition on the AudioSet corpus [Gemmeke et al., 2017] and speech emotion recognition on the IEMOCAP datasets.

D.0.2 Downstream Task Evaluation of the SSL Pre-Trained CNN Encoders

Dimensional Emotion Recognition Datasets. For evaluating the SSL-pretrained visual and audio CNN encoders on unimodal emotion recognition tasks, this work used two benchmark in-the-wild datasets, SEWA [Kossaifi et al., 2019] and the AVEC'19 Cross-cultural Emotion Sub-Challenge (CES) [Ringeval et al., 2019], respectively.

SEWA data was collected during computer-based naturalistic dyadic interactions and contains 538 face videos of 398 subjects from 6 different cultures. Each video is annotated with per-frame continuous-valued valence and arousal annotations in the range of -1 to 1 at 50 frames per second (FPS). The numbers of videos used for training, validation, and

testing[1] are 431, 53, and 53, respectively, with the duration in the range of 10 s to 30 s.

AVEC'19 CES Corpus provides the audio recordings of interactions between pairs of individuals from German, Hungarian and Chinese cultures. As the labels of test sets and Chinese culture are not publicly available, this work used the German and Hungarian training and validation sets. It contains 64 videos for training, and 32 videos for validation, with a total duration of roughly 160 minutes and 65 minutes, respectively. Audio data is recorded at 48 kHZ and the ratings of valence and arousal are presented at 10 FPS. Liking dimension of this dataset is not used in this work as the liking recognition typically needs linguistic features that are explicitly derived [Ringeval et al., 2018], whereas the focus is only on the audio-modality here.

Network Architecture for Emotion Recognition. Here, the evaluation focuses on unimodal temporal emotion recognition, a 3-layer BiGRU-RNN with 256 hidden units and a fully connected output layer are included on top of the SSL pre-trained CNN feature encoder for sequential prediction. Given a sequence of face images or Mel-spectrogram images as input, first the CNN extracts per-frame embedding vectors, which are sequentially processed by the BiGRU-RNN module to predict per-frame valence and arousal values. Note that both these modules, CNN and BiGRU-RNN, are trained in an end-to-end manner, using the ground truth emotion labels for supervision.

Evaluation Metric. Lin's Concordance Correlation Coefficient (CCC) [Lawrence and Lin, 1989] is used to measure the the agreement between the predicted

[1]The details of the train, validation, and test partitions were kindly provided by the database owners.

emotions y^o and their ground truth labels y^*

$$CCC = \frac{\rho_{y^*y^o}.\sigma_{y^*}.\sigma_{y^o}}{(\mu_{y^*} - \mu_{y^o})^2 + {\sigma_{y^*}}^2 + {\sigma_{y^o}}^2}, \tag{D.1}$$

where $\rho_{y^*y^o}$ denotes the Pearson's coefficient of correlation between y^* and y^o, and (μ_{y^*}, μ_{y^o}) and $(\sigma_{y^*}, \sigma_{y^o})$ denote their mean and standard deviation values, respectively.

Optimisation Details. To train all the emotion recognition models, a hybrid loss function is used here, $L = L_{MSE} + L_{iCCC}$ where $L_{iCCC} = 1 - CCC(Y_{gt}, Y_{pred})$ and L_{MSE} is the mean square error between Y_{gt} and Y_{pred}. The dropout values in the BiGRU-RNN and the final FC layers are set to 0.5 and 0.25, respectively, and L_2 regularisation is applied by setting the weight decay value to 1e-4. Each mini-batch is composed of 4 sequences, with each sequence containing 100 frames. The initial learning rate value is 1e-4, and it is tuned using a cosine annealing based scheduler with warm restarts enabled [Loshchilov and Hutter, 2016].