Deep Learning Methods for Affect and Schizophrenia Symptom Estimation

Niki Maria Foteinopoulou

Submitted in partial fulfillment of the requirements of the Degree of
Doctor of Philosophy

Supervisor: Prof. Ioannis Patras

School of Electronic Engineering and Computer Science

Queen Mary University of London

United Kingdom

December 2023

Statement of originality

I, Niki Maria Foteinopoulou, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author. Signature: Niki Maria Foteinopoulou

Associated publications

Details of collaboration and publications:

- **N. M. Foteinopoulou** and Ioannis Patras. “Machine Learning Approaches for Fine-Grained Symptom Estimation in Schizophrenia: A Comprehensive Review”, Journal manuscript under review (preprint arXiv:2310.16677)
- **N. M. Foteinopoulou** and Ioannis Patras. “EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition”, in Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkiye, 2024, pp. 1-10.
- **N. M. Foteinopoulou** and I. Patras, ‘Learning from Label Relationships in Human Affect’, in Proceedings of the 30th ACM International Conference on Multimedia, Lisboa Portugal: ACM, Oct. 2022, pp. 80–89.
- **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), Nara, Japan: IEEE, Sep. 2021, pp. 1–8.

Abstract

Human communication encompasses a rich mixture of verbal and non-verbal cues, significantly contributing to conveying emotion and mental state. While humans instinctively recognise such cues, explaining and defining them in natural language is challenging, rendering tasks in affective computing more complex than other discriminative tasks, e.g. object detection, as the ground truths are ambiguous in classic supervised settings. The primary focus of this thesis is to develop automated methods to predict human emotion and negative symptoms of schizophrenia, primarily from non-verbal facial cues. By acknowledging the substantial methodological and contextual overlap between predicting human emotion and assessing negative symptoms of schizophrenia—closely linked to affect and emotion—we adopt a concurrent approach, focusing on three key challenges predominantly tied to the nature of ground truth labels. More specifically, through this thesis, we aim to address a) the label uncertainty in human affect, which stems from the inherently noisy nature of human emotions and can be seen in practice by annotators' disagreement, b) labels that describe a broader behaviour resulting in low-label resolution and c) the vast variability of human affect, which results in subjective emotional descriptions when expressed in natural language. Firstly, we propose a method that addresses label uncertainty in continuous affect. We assume that each ground truth label is a univariate Gaussian distribution with a mean equal to the ground truth mean and an unknown variance that is predicted by the network. The Kullback-Leibler-based loss minimises the distance between the Gaussian ground truth and the Dirac delta prediction. We show that the proposed loss improves convergence and a relationship between the estimated variance and noisy samples. Secondly, we propose a deep learning approach for continuous affect and symptom estimation in long video samples that learns from the clip and the batch context. Contrary to previous works that addressed the problem either using statistical representations or trimming videos into shorter clips, we propose using features from the wider video when making a clip prediction. We also introduce a novel loss as an auxiliary task, named the relational regression loss that aligns the continuous label vector distances in the mini-batch to those of the latent features. The ablation studies show that both components offer significant performance improvements to both tasks. Finally, we develop a novel vision-language model that utilises sample-level text descriptions as natural language supervision to learn semantically rich representations for each sample to address the intra-class variability of emotional expression. Then, during inference, we use category-level descriptions for each emotion in a zero-shot approach rather than the typical class prototypes previously used in zero-shot Facial Expression Recognition. We also use the vision modality as a backbone for the downstream task of schizophrenia symptom estimation. The method shows significant improvement compared to baseline methods and outperforms previous works on both tasks, showing the benefit of more fine-grained approaches.

Acknowledgements

First of all, I must thank my supervisor, Prof. Ioannis Patras, for his help and support throughout my studies, but also for trusting me and giving me the opportunity to conduct this PhD research project. I would also like to thank the rest of the members of my supervisory team, Prof. Shaogang Gong and Prof. Matthew Purver, for their insightful guidance.

I would also like to thank all the past, present and visiting colleagues I met in MMV and QMUL over these years for the countless interesting discussions and memorable moments: Tingting, Chen, Giannis, James, Zheng, Jingjing, Davide, Dario, Zengqun, Debin, Yeming, Zhonglin, Ioanna, Alexandros, Zhaohan and Kit. I thank all of them very much and apologise for any involuntary omission from this list. Special thanks go to Dr Christos Tzelepis and Dr Georgios Zoumpourlis, the first people I met in the group, who instantly made me feel welcome and helped profoundly during my early research days. Also, special thanks to Ioanna, Kike, Kit, Stathis, Pavlos and Lydia, who proofread sections of this thesis.

Finally, I thank my friends, family, partner and Ajax (my dog) for their daily psychological support through this milestone.

This work was supported by the EPSRC DTP studentship (No. EP/R513106/1)

Contents

List of Figures

List of Tables

List of abbreviations

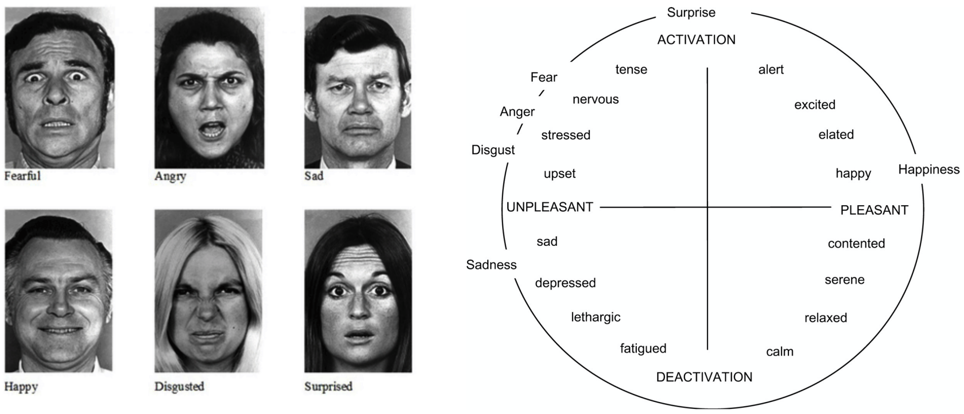
AI	Artificial Intelligence
AU	Action Units
CCC	Concordance Correlation Coefficient
CI	Confidence Interval
FC	Fully Connected
FER	Facial Expression Recognition
GMM	Gaussian Mixture Model
HCI	Human-Computer Interaction
KL	Kullback–Leibler
LLM	Large Language Model
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
PCA	Principle Component Analysis
PCC	Pearson's Correlation Coefficient
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
SVR	Support Vector Regression
VLM	Vision Language Model

Chapter 1

Introduction

Contents

Recent years have seen an explosion of technological advancements, often neologically described as the fourth industrial revolution. As such, the interactions of humans with automated systems have been integrated into the daily lives of people in most developed nations. The rise in Human-Computer Interaction (HCI) has even taken over areas traditionally requiring communication or coordination between humans, for example, in customer service [86]. However, these systems still largely depend on rigid choices rather than natural communication, including essential non-verbal and emotional cues, which comprise a large part of human communication and behaviour. Affective computing, a multi-disciplinary research area, attempts to bridge the gap by recognising, processing and simulating human affect [90] in an automated manner. There are several potential applications for affective computing, ranging from enhancing user experience in human-computer interaction, fostering more intuitive and empathetic interfaces to assisting doctors [77] and educators [128] where the technology can facilitate personalised and emotionally aware interactions, ultimately contributing to improved well-being



(a) Ekman & Friessen [35] six basic emotions

(b) Arousal-Valence circumplex [99]

Figure 1.1: Examples of the six basic emotions first proposed by Ekman & Friessen [35] a and the Arousal-Valence circumplex b (recreated from [39])

and learning outcomes respectively. While non-verbal cues can include a vast number of behaviours, such as body pose and hand gestures, the most common and universally recognised are related to facial expressions. Non-verbal emotional cues are not only important for effective communication but can also be indicators of mental illness as several disorders are associated with human affect, such as depression, schizophrenia or other associated illnesses [121].

While humans instinctively express and interpret emotional cues, particularly in the form of facial expressions, these are hard to explain in natural language, as is evident by the multiple models proposed to formalise human affect. More specifically, starting with a categorical interpretation, Ekman & Friessen [35] identified six basic emotions universally experienced in all cultures. However, while these six categories, namely *happiness*, *sadness*, *disgust*, *fear*, *surprise*, and *anger*, can easily be explained in natural language, there are several issues associated with this coarse categorisation. Namely, the coarse categories do not include any information on the intensity, thus creating substantial intra-category differences. In addition, emotional and mental states can be complex or not fully described as sub-categories of the basic six (e.g. *confused*). To address the shortcomings of the coarse categorical emotions, the continuous *Arousal-*

Valence circumplex was proposed [99]. This model of emotion introduces emotional and mental states as points on a two-dimensional space defined by the extent to which an emotion is associated with a sensation of energy –*Arousal*– and the extent to which an emotion reflects a positive state of mind –*Valence*. By treating each emotional state as a unique data point, the circumplex addresses the issues associated with coarse categorisation; however, continuous scales are not intuitive or easily explained in natural language, raising questions on the scale's universality and interoperability. Visual examples of the basic emotions and the circumplex can be seen in Fig. 1.1. To address the complexity of human emotional expression, several models, such as compound emotions or identifying fine-grain Action Units [36], have been proposed in psychology. However, the former may still suffer from ambiguity, and the latter can be very effective in describing an expression but not in translating to emotional terms.

Similarly, in clinical practice, the understanding of illnesses has evolved from broad definitions to fine-grained and detailed descriptions of specific symptoms, many of which are related to non-verbal behaviour and apparent affect. However, as we established previously, human behaviour, in general, can be abstract, leading to some disagreement in diagnosis and assessment even amongst experts [56, 104]. For example, the intensity of Blunted Affect, one of the negative symptoms of schizophrenia, is set relative to "normal" emotional responsiveness, which is very difficult to quantify and, therefore, can be open to individual interpretation. As these definitions have an element of subjectivity and unconscious bias, any automated system is susceptible to the same pitfalls trained human experts would be. In other words, whether looking into the very large intra-category differences of coarse categorical models or uncertainty of continuous labels, Machine Learning (ML) methods in affective computing are presented with a more noisy problem definition than in other ML tasks, e.g. object detection.

Further to the inherent noise associated with human behaviour, tasks in affective computing share additional challenges. More specifically, as affective tasks typically contain identifiable information, obtaining large datasets is difficult due to confidentiality concerns, particularly for mental health tasks where medical records are further constrained, leading to limited data availability. In addition, as human behaviour can refer to either very short bursts or wider behaviour, very often, samples in dynamic tasks are very long sequences associated with a single label. This low-label resolution, i.e. having a single label vector for a very long sequence, is problematic as in dynamic human behaviour, it can translate to aleatoric uncertainty in the data [68] or methodological constraints such as forgetting [46].

As we have established, affective computing tasks share a number of similarities in terms of the problem definitions, challenges and, therefore, the approaches used to address them. In this thesis, we address challenges that are common in several human behavioural tasks and, more specifically, in human affect and negative symptoms of schizophrenia (which are related to affect) by developing methodologies that can extend to both tasks concurrently. We note that the term "negative" in this context refers to behaviour that is not present in patients relative to the general population and not to the effect these symptoms have.

1.1 Problem Definition

The core aim of this thesis is to contribute to the development of a fully automated system for understanding emotional and mental states in the wild (i.e., not in a lab or under posed conditions) that can aid human experts. However, in contrast with other supervised computer vision tasks where the ground truth can be easily defined with very little variation between expert annotators (e.g. object detection), human behaviour tends to exhibit a more fluid and dynamic nature in comparison. Consequently, annotations utilised as ground truth labels in a supervised setting will inherently contain noise, as they are shaped by the annotator's interpretation of the subjects' behaviour. This is evident in the multiple models proposed by psychologists to explain the human emotional experience; whether addressing human affect with coarse categories or as a set of continuous labels, the ground truth remains ambiguous.

Categorical models define emotions as discrete categories; Ekman's [35] model of basic emotions defined seven emotions –*anger, disgust, happiness, sadness, fear, contempt*, and *surprise*– as the primary universal ones, both in terms of expression and understanding across cultures and individuals. Ekman's categorical model does not claim that the range of human emotions is limited to those seven categories but that all other emotional states originate from them. This translates to very large intra-class variability as each category does not account for intensity or complex emotions; for example, anger will include samples from *annoyance* to *rage*, which can have very different facial expressions associated with them and are influenced by factors such as culture, anatomy and context.

To address the shortcomings of categorical models, several continuous models are proposed, the arousal-valence circumplex [99] being the most common to map human affect. Dimensional models can measure intensity and theoretically capture the broad spectrum of human emotions. However, the arousal-valence axes are not self-explanatory, resulting in ambiguity when annotating and interpreting samples.

Similarly, in several mental health tasks, the symptom definitions and their associated intensities are described in natural language; however, as we have established the variability in human behaviour, symptom definitions can also be abstract and subjective. For example, a common non-verbal symptom of schizophrenia is Blunted Affect, where patients "exhibit less than normal facial expressions"; this, however, assumes a common universal definition for the "normal" range of expressiveness.

Such variations can be mitigated by training the expert annotators [11] but not completely eliminated [56]. While this label noise is inherent in all tasks related to human behaviour to various degrees, in this thesis, we focus on addressing the problem in the tasks of apparent emotion and schizophrenia symptom estimation. In addition to the label noise discussed in both tasks, negative symptoms of schizophrenia and apparent affect manifest themselves in similar non-verbal manners, i.e. facial expressions, body pose, vocal expression, etc. Thus, based on both similar manifestations in terms of the subject's behaviour and similar challenges with regard to the ground truth, it is intuitive to examine the two tasks concurrently.

There are several ways to address issues around ground truth, whether labels are categorical or continuous. Note that the thesis does not propose a new label model but explores methods that are aware of the problems associated with ground truth labels in affective computing. The first aim is to estimate the apparent continuous affect along the arousal and valence axis. We thus treat the problem as a regression. The second aim of the thesis is to estimate symptom severity of several negative symptoms of schizophrenia that are related to apparent affect and thus have similar difficulties both in the data and the label definitions. As humans have individual ways of expressing and describing apparent emotion, the final aim of this thesis is to infer emotion based on the natural language descriptions of each category rather than typical classification tasks. In Chapters 3 & 4, we focus on addressing issues in continuous labels, while in Chapter 5, we focus on issues associated with the definitions of categorical emotions.

More specifically, in Chapter 3, we directly address label uncertainty measured by annotators' variance. In addition to the main regression task, we also estimate the label noise of each sample and compare it to the annotators' disagreement at test time. In Chapter 4, we address issues associated with low-label resolution and representation learning in continuous tasks. We continue measuring the apparent affect on the arousal-valence circumplex but extend the methodology to the mental health task. Finally, in Chapter 5, we explore the use of descriptions of facial expressions in natural language for zero-shot classification of emotions.

1.1.1 Challenges

In this section, we will describe the main challenges of the problems we are addressing, namely, a) noisy labels, b) low-label resolution, and c) new and unseen emotional and mental states. Following each challenge definition, we briefly describe how this is addressed.

a) Noisy Labels: In this thesis, we use facial expressions and emotions to estimate symptom severity in patients with schizophrenia. However, as previously discussed, there is an inherent subjectivity in understanding and expressing emotion. Therefore, the annotators' unconscious bias will affect any data used for training and evaluating ML systems. To mitigate the bias, authors typically use multiple expert annotators for each sample and use the inter-annotator agreement to measure label quality when collecting and annotating datasets. In most cases, and for simplicity, researchers take the average label (for continuous Arousal-Valence) or the mode in categorical labels to address disagreement; however, such an approach disregards any information regarding the sample we would get from the variance. In other words, the label noise can give us information regarding how "easy" a sample is to learn. Furthermore, by modelling the label noise, we can learn from noisy samples and improve ML systems' understanding of human affect.

Methods using the categorical model often use the distribution of labels as soft supervision to better learn under label noise [65]; however, in continuous affect, label uncertainty is rarely addressed either explicitly or implicitly. As part of this thesis, we address and leverage information from noisy continuous labels by modelling the ground truth as a univariate Gaussian distribution with an unknown variance predicted by the network. The uncertainty-aware loss used during training improves several architectures' performance in static and dynamic continuous effect estimation. In addition, during inference, we show a weak relationship between annotators' disagreement and predicted variance, indicating the model's understanding of noise.

b) Low Label Resolution: In affect and mental health tasks, the labels typically refer to a wider behaviour dependent on the context rather than per-frame categorisation. While the latter, i.e. static Facial Expression Recognition (FER), is a valid task of affective computing, it completely disregards temporal information and does not reflect human emotional experience. As such, including a wider temporal context can significantly improve an automated system's understanding of apparent emotion and mental state, particularly in mental health, where understanding the underlying pattern of behaviour is crucial in diagnosis. However, this low label resolution creates several issues when estimating long sequences, with, first and foremost, limitations faced by modern hardware. In addition, recurrent architectures used for long-sequence analysis tend to face issues with vanishing gradients [46]. In contrast, Transformer [116] based architectures are typically data and resource-hungry, especially when long-sequence inputs are used. Furthermore, due to confidentiality constraints, affect and mental health datasets typically contain few samples.

To address the issue of low label resolution in a low-data regime, we propose using the context of each sample. This is done in two ways: directly by building a two-stage attention architecture that uses features from the video clips' temporal neighbourhood to directly introduce context information in the feature extraction. In addition, we introduce a novel loss that uses the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner, thus learning from the batch context. More specifically, the two-stage architecture is building upon previous work in action recognition [122] to improve clip-level prediction using the temporal context, using an improved architecture to reduce the number of trainable parameters and computational resources needed. Inspired by the success of Contrastive Learning [27] in single-label categorical tasks, we propose aligning the distances of latent feature representations with the distances of the label vectors in multi-label regression problems in affect and mental health. The introduction of the proposed loss as an auxiliary to the main regression acts as a strong regulariser, particularly when a very limited number of samples is available, directly addressing the data-hungry nature of Transformer architectures.

c) New and Unseen Emotional States: One of the main arguments against the categorical model of Ekman & Friessen [35] is the rigidity of the category definitions compared to human emotional expression and understanding. Even though the categorical model originally proposed the basic emotions as coarse categories from which fine-grained emotional categories stem, in most FER datasets, only the coarse categories are represented as they are easily defined and, in most cases, are self-explanatory in contrast to more fine-grained or continuous annotations. However, coarse categorisation suffers from several issues; namely, it does not address emotion intensity, there are vast intra-class differences, and it may be inadequate to describe the apparent emotion, particularly for downstream tasks.

This thesis proposes using sample-level descriptions for natural language supervision in dynamic zero-shot FER. Natural language supervision is not a novel paradigm on its own; however, this is a novel approach for zero-shot FER that typically learns representations using class prototypes. Furthermore, as we accept that the basic emotions are inadequate to describe all emotional states and, in many cases, show significant intra-class variation, we propose manipulating the latent space of description representations of basic emotions to obtain representations for compound emotions rather than creating additional prompts. The improved embedding space from natural language supervision can also be seen in the downstream task of schizophrenia symptom estimation, particularly for symptoms that are, by definition, related to FER.

1.2 Contributions

In this section, we list the main contributions of the thesis. In the first main chapter (Chapter 3), we train multiple backbone architectures with the proposed uncertainty-aware approach in video and static datasets. More specifically, the main contributions of this chapter can be listed as follows:

- We build upon the work of He *et al.* [50] to address label noise in continuous human affect. This novel approach addresses uncertainty stemming from annotators' disagreement in affective computing. The motivation behind it lies in the inherent ambiguity of continuous affect; rather than taking the mean of the annotators as the ground truth, which is the typical approach, we address each target label as a univariate Gaussian distribution with a mean and an unknown variance that the model predicts. To do so, we optimise two Multi-Layer Perceptron (MLP) heads, one predicting the mean and the other predicting the variance of the respective Gaussian label. This is achieved by optimising a Kullback–Leibler (KL) based loss that estimates the distance between the predicted Gaussian and a Dirac delta centred around the predicted mean.
- We show that the proposed approach improves the performance of several convolutional backbone architectures, as the KL-based loss allows for smoother convergence by penalising the network less for ambiguous samples.
- We finally show a weak relationship between estimated variance and label noise in two datasets, with both video and static image samples.

In the following chapter (Chapter 4), our contributions are twofold: introducing a novel loss that helps regularise the training in a low-data regime and building an architecture that uses context to improve clip-level predictions in low-label resolution tasks. The main contributions of this chapter can be listed as follows:

- We introduce a novel loss as an additional task to the primary regression task, named relational loss. This loss aims to align the intra-batch distances of the label vectors with the distances of the latent feature representations. We show that the improved latent space helps improve the main regression task by a large margin, particularly on small datasets.

- We also propose building upon previous work [122] to use features from each clip's context to improve prediction. To achieve this, we use a two-branch network with shared weights, where one branch extracts clip-level features and the other context-level features. We achieve shorter training times and more parameter-efficient networks by sharing weights between branches.
- We show how the methodology performs on the tasks of continuous affect and schizophrenia symptom severity estimation.

Finally, in the last main chapter of the thesis (Chapter 5), we propose fine-tuning Vision Language Models (VLM) in dynamic affect, by training on video-text pairs. Contrary to previous zero-shot works that use class prototypes for each emotional class, we propose to use sample-level descriptions to capture the intra-class variations in the latent feature space. The main contributions of this chapter can be listed as follows:

- As there are surprisingly only a handful of previous works of zero-shot FER, we are among the first to evaluate the capabilities of VLMs on the task.
- We introduce a method that manipulates the latent representations of basic emotions to represent compound emotions rather than creating additional prompts.
- Through both qualitative and quantitative assessments, we show that fine-tuning using sample-level descriptions improves representation learning on the domain.
- Extensive experiments on four popular dynamic FER datasets show the method's zero-shot capabilities. We also use the fine-tuned model as a feature extractor in the downstream task of symptom estimation.

1.3 Thesis structure

The rest of the thesis is structured as follows. In Chapter 2, we start by reviewing the literature related to schizophrenia symptom estimation and the various challenges identified. More specifically, we review works addressing uncertainty, representation learning, and Zero-Shot learning. In Chapter 3, our first methodological chapter, we follow by presenting our work on addressing label uncertainty in continuous affect. Chapter 4 presents the two-stage attention architecture and the proposed relational loss. In Chapter 5, we present our approach for zero-shot FER for basic and compound emotions. Finally, we draw our conclusions and opportunity for future work in Chapter 6.

Chapter 2

Literature Review

Contents

In Chapter 1, we discussed the applications of affective computing and introduced some preliminary knowledge on schizophrenia. We begin Chapter 2 by describing the various works on schizophrenia symptom estimation using audio-visual input in Section 2.1. We continue the chapter by discussing how previous works in computer vision in general and then human affect specifically address Data and Label Uncertainty in Section 2.2. We then present works in representation learning in Section 2.3 and addressing long sequences in Section 2.4. Finally, we present works using VLM, as natural language supervision has shown impressive capabilities in zero-shot discriminative tasks in Section 2.5.

2.1 Schizophrenia Symptom Estimation

As several mental health illnesses and disorders have non-verbal behaviour symptoms, understanding patients' affect is important in diagnosis and severity estimation. Depression is a mood disorder that has an impact on patients' affective state; similarly, several schizophrenia negative symptoms refer to patients' affect and expressions; therefore, there are important semantic parallels between continuous affect estimation and mental health assessment, so we examine them in parallel. More work has been performed on estimating depression severity than symptoms of schizophrenia, as there are no publicly available datasets for the latter.

Currently, mental health practitioners primarily rely on clinical interviews following a structured framework outlined in DSM-V [121] to assess individuals with schizophrenia. As such, leveraging audio-visual recordings of patients for diagnosis and symptom estimation presents a more intuitive approach that closely resembles real-world conditions than medical imaging or bio-signals.

In clinical practice, schizophrenia manifests itself in various aspects of a patient's behaviour, encompassing facial expressions, vocal patterns, and overall demeanour. Mental health practitioners directly gauge these behavioural symptoms as an estimate of the individual's illness state and progression. Since discrete values measure symptom severity, researchers have approached this problem primarily as either multi-label multi-class classification or multi-label regression tasks.

Tahir *et al.* [109, 110] addressed the problem of symptom severity estimation on the Positive Negative Symptom Scale (PANSS) [55] symptom scale, as both a classification and a regression task using Support Vector Machines (SVM) and Support Vector Regression (SVR) and with the use of hand-crafted features of non-verbal cues associated with conversations (e.g., interruption, natural turn, etc). However, such features require significant manual effort and thus are not fit to train large models with adequate generalisation.

As para-linguistic features have been proven to be crucial in estimating affect [20, 96], several works [24, 37, 18] use low-level descriptors (lld) from audio recordings of patients. Chakraborty *et al.* [24] used lld from clinical interviews and Principle Component Analysis (PCA) to reduce their dimensions. The authors trained several binary classifiers for high-low classification of each symptom on NSA16 [4]. Similarly, Boer *et al.* [18] extracted acoustic features using the OpenSMILE [38] toolkit and trained a set of Random Forest classifiers for a three-class classification task. However, binary and few class classifiers are not adequate to describe the full spectrum of the illness.

Similar to para-linguistic audio features, certain behavioural symptoms are manifested in the subjects' facial expressions and mannerisms. As a proof-of-concept, Barzilay *et al.* [12] extracted features related to the subject's facial expression per frame and in the whole video. These were then used to train a set of SVM models for affect subtypes in patients with schizophrenia, one for each of the five annotators. By highlighting the high disagreement between the human annotators, this study underlines the need for automated and consistent symptom assessment and diagnosis. Furthermore, as the classification process, in this case, is (in practice) conducted using personalised models for each annotator, the overall accuracy seems to depend on the annotator without creating a unified model or addressing the annotators' disagreement. Similarly, Abbas *et al.* [1] measured the head movement of subjects from smartphone front cameras. A linear regression was trained for symptom severity estimation, showing a negative relationship between head movement and high-symptom severity, particularly for negative symptoms, as would be expected based on the symptom definition. However, the results obtained are not robust, as would be expected from using a single feature.

Tron *et al.* [113, 115] recorded 34 schizophrenia patients and healthy controls during a clinical interview. From the video recordings, 23 Action Units (AU) [36] and their respective intensities were extracted for each frame. In [113], hand-crafted features, such as activation ratio and intensity, were used as descriptors of the patients' facial behaviour to train an SVM classifier on binary schizophrenia detection; in addition, the authors trained a ridge regression on symptom intensity within the patient pool of the collected dataset. Similarly, in a subsequent study [115], a k-means clustering algorithm was used first to assign each frame to a centroid, with the cluster centres representing facial expression prototypes. The use of AU and SVR is also adopted by Vijay *et al.* [117]; similarly to Tron *et al.* [113], the authors extracted AU from the whole recording session of a patient and constructed handcrafted features related to the AU prevalence and intensity to train a series of SVRs on symptom intensity. Bishay *et al.* [15] continued using AU as inputs to estimate symptom severity, as in [113, 115, 117]. More specifically, the authors took a staged approach, first training multiple VGG16 [105] networks on the detection of individual AUs. Contrary to previous works [113, 114, 115] that used hand-crafted features from frame level AUs, [15] used a Gaussian Mixture Model (GMM) followed by a Fisher Vector transformation to standardise the input to a fixed-length vector. The use of GMM and Fisher Vector transformation is streamlining and automating the process further; however, there is less control in feature selection and engineering. Finally, two Fully Connected (FC) layers were then used for the regression task, the first one estimating individual symptom intensities for three symptoms on the PANSS [55] negative scale or all expressive symptoms of the Clinical Assessment Interview for Negative Symptoms (CAINS) [40], and the second estimating the total negative score using the individual symptoms as input. While a clear progression from hand-crafted features to statistical representations can be seen in the literature, these methods disregard temporal relationships of features between frames and focus on AUs, which can potentially lead to omitted variable bias.

As ML for mental health is an emerging field, there are several open questions and future directions. The majority of studies included in this section use simpler linear methodologies to address the problem, with only two using deep learning approaches [15, 16]. This is in stark contrast to the majority of works in other domains, where deep learning methodologies are the dominant paradigm. As such, there is significant potential for methodological improvements in the domain of fine-grained symptom estimation. Furthermore, the majority of the works use models pre-trained on different tasks to extract features rather than using the raw image in their method, thus leading to potential omitted variable bias. Finally, the temporal relationships between features in works using audio-visual features are not utilised in previous works; therefore, there is significant work that can be done in the field, learning from the temporal dimension.

2.2 Addressing Data and Label Uncertainty

A significant amount of work has been done on data uncertainty in the form of noisy labels for classification tasks. Methodologies such as MixMatch [14], DivideMix [69], and FixMatch [108] are adopting a semi-supervised approach to address noisy labels and make a decision during training that splits samples into clean and noisy subsets. However, this approach, i.e., making a hard decision on uncertain samples, does not offer interpretability of the per-sample data uncertainty. In contrast, the proposed method adopts a continuous measure, which is derived per sample by a branch of the network.

Bayesian deep learning approaches have gained popularity in dealing with data uncertainty; for instance, for the task of image segmentation, Kendall and Gal [58] proposed a per-pixel regression uncertainty-aware approach. Similarly, modelling data uncertainty in latent space [103, 25, 100] has proven to improve face recognition. However, these works focus on data rather than label uncertainty. Moreover, in domains such as object detection [50] and temporal action localisation [124], data uncertainty is addressed by learning the variance of a continuous prediction value, i.e., the bounding box spatial boundaries of an object in an image or the temporal boundaries of an action in a video, by optimising a modified KL divergence loss function.

In these works, uncertainty is modelled per sample as a set of univariate Gaussian distributions of the predicted regression values, with both mean values and variances being predicted by the network. In contrast, instead of the predictions, the proposed method models the ground truth values as uni-variate Gaussians, for which the annotators' mean values are given and the variances are optimised using a KL-divergence-based loss term. Moreover, while data uncertainty modelling has been implemented in other regression problems, it appears that none of these works addresses the problem in continuous affect estimation.

Yannakakis et al. [131] propose comparing samples and ranking them rather than using absolute labels to address data uncertainty. This is an interesting approach to address label uncertainty; however, most datasets are annotated in a categorical or continuous manner and not in rankings. A recent work by Toisoul et al. [112] also evaluates against a “clean” subset of AffectNet [84], where samples are excluded when deemed noisy by a set of predefined rules. Their method performs better on the clean evaluation set, even though noisy sample labels are not corrected or excluded during training. Resigno et al. [97] propose the use of personal models for affect recognition to overcome generalisation issues due to physiological or cultural differences. However, the aforementioned works do not estimate the level of label uncertainty in affect estimation but rather attempt to clean the dataset of noisy samples. Han et al. [48] propose an uncertainty-aware methodology for continuous affect estimation by explicitly training on the inter-annotator disagreement as an additional task. Similarly, Chou and Lee [28] propose an ensemble methodology for speech emotion classification and use annotators' disagreement as a target during training. However, while their methodology improves on the baseline, showing the importance of uncertainty-aware models, it is dependent on individual annotations being available.

Several works have addressed the issue of uncertainty when multiple annotations of a given sample are available. Using a Gaussian process classification approach has been proven to outperform other approaches (e.g., majority voting) in multiple domains [98, 75]. These works explicitly handle uncertainty arising from annotators' disagreement. Similarly, ensemble architectures that model each annotator and implement decision-level fusion [47] for each sample show improvements against the baseline. However, such approaches require a large number of annotations per sample to model the annotation distribution and guarantee it is representative. By explicitly handling the uncertainty in Gaussian processes, the network learns the annotator's disagreement rather than the sample ambiguity. Furthermore, the latter approach of ensembles from individual annotator models does not provide sufficient information on the sample's uncertainty.

2.3 Learning Representations and Label Relations

In human affect problems and even more in mental state estimation, learning features representative of the behaviour rather than other entangled factors (e.g. identity) is paramount to the reliability of the final estimate. A number of works have addressed the issue of representation learning, with more recent developments in contrastive methodologies [27, 26], whether evaluating results on static image data or video datasets [94]. These self-supervised methodologies learn latent representations by teaching the architecture which data points are similar. By extending the idea of comparing samples, supervised contrastive frameworks propose that images [59] or videos [51] from the same class are treated as similar, which results in embeddings from the same class being more closely aligned. However, these works are trained on very large datasets which are not typically available for affective and mental health problems and have only been evaluated on classification problems. Kim *et al.* [60] implement an adversarial loss to learn better representations for continuous affect; however, arousal/valence values are binarised for the adversarial task. In our work, we explore the idea of learning representations by comparing sample similarities in a supervised approach; however, we implement a non-binary approach, which is more suitable for multi-label regression problems.

Several problems/datasets in the field of continuous affect and mental health have multiple labels in order to describe various affective attributes and psychological symptoms. Treating each label independently [31] ignores their potential correlations as well as increases training times significantly with each additional label. Several works investigate multi-label recognition problems using graph learning approaches to model label correlations and co-occurrences [119, 70]. However, such approaches do not learn from label similarities *between* samples and do not project these similarities to the latent representation space. In contrast, our work uses information from the inter-sample label similarities to learn better latent representations.

2.4 Addressing large sequences

The exploration of methods tailored to long-range video understanding is vital for human affect and mental state estimation, as long videos are typically more representative of real-life settings. Moreover, long temporal relationships intuitively should contribute to more accurate estimates of human affect and mental states. To address long video sequences, previous works have used a number of strategies. One such method is to pre-compute features [66, 15]; this, however, does not allow for end-to-end training and makes augmentation techniques more complicated (if feasible at all). Another strategy to address long video sequences is by using contextual features either in the form of intra-sample relations [138] or by exploring feature banks [122]. Both of these approaches utilise relations between the short-term actions, which is the temporal context of a clip. However, these methods have been evaluated on action recognition problems and have not been implemented in affect. Moreover, while action problems benefit from long-term context, they still have a much lower label resolution. Finally, when using feature banks context, features need to be pre-computed on pre-defined clips; therefore, feature quality does not improve with further training.

In our work, we build on the concept of exploiting contextual features; however, differently from [122], we use context features to improve clip-level prediction with end-to-end training. We also do not operate on pre-defined clips but rather dynamically compute features from them – with this approach, as training progresses, the network learns from better context features.

2.5 Vision and Language Models

The use of large contrastive pre-training for Vision-Language Models (VLM) has become popular, as these models have demonstrated impressive generalisation capabilities [95, 53, 137, 2]. As large VLM require large amounts of data and very high computational resources to achieve those results, the latest research on VLM has been mostly concentrated on three paths: (a) latent space manipulation, (b) leveraging pre-trained spatial features and fine-tuning a temporal module for video input, and (c) prompt learning.

Menon & Vondrick [82] use an ensemble of prompts generated by Large Language Models (LLMs) containing descriptive features of each class and show significant improvements in terms of accuracy as well as explainability of decisions. Ouali *et al.* [87] propose a method for latent space feature alignment in a target domain without the need for additional training. Similarly, Bain *et al.* [8] propose several methods for temporal pooling of frames, using pre-trained VLMs with very little or no additional training. Such approaches, although effective, use no information from the temporal dimension in the video, which is essential in human FER to understand macro and micro-expressions.

Large VLMs trained on static images have been used for video classification, particularly for action recognition. Lin *et al.* [72] propose using a lightweight Transformer Decoder over the CLIP [95] spatial features for downstream classification. Similarly, ActionCLIP [118] class labels are used for natural language supervision; therefore, in both approaches, the open vocabulary capabilities are lost. CLIP [95] has been used as a backbone in several video captioning works [78, 127, 79]. However, none of these works has been evaluated or trained on the domain of FER, where the behaviour is, in general, not as clearly defined.

To overcome the challenges of prompt engineering in VLMs, some works propose learning a set of tokens [140, 139, 88] to append to the class name, which can improve performance as the text-encoder acts more like a bag of words [132, 7]. Natural language supervision for facial expression recognition (FER) is a relatively unexplored idea, but there have been some preliminary works exploring this approach. For instance, CLIPER [67] has proposed prompt learning to improve closed dictionary FER. However, these tokens are class-specific and cannot be used in open dictionary settings for zero-shot classification.

Several works [9, 93, 125, 126] have proposed zero-shot frameworks for emotion recognition or emotional response recognition [134] by aligning class name embeddings to multi-media embeddings and then evaluating the method on unseen emotions. These methods, however, still rely on the use of hard labels and simpler class prototype embeddings (such as word2vec). As such, they do not take into consideration the intra-class differences or the underlying concepts in each class and do not capture semantically rich information in the latent representations.

Natural language supervision is not a new idea; however, data and hardware constraints only recently showed the impressive generalisation capabilities of VLM models. In the domain of human affect, we can make two main observations on the bibliography: a) there are very few works attempting zero-shot FER, and b) none of the methodologies using natural language supervision are evaluated on in-the-wild FER tasks. As such, we can conclude that there is a significant research gap on VLM models for human affect. In addition, as coarse categorisation of emotions can often be subjective and show large intra-class variation, emotional prototypes fail to capture more fine-grained

Table 2.1: List of Datasets Used in this Thesis

Name of the dataset	Size	Type	Available Annotation	Form of collection
AffectNet	1,000,000	Images	Arousal-Valence, Categorical Emotions	Wild
AFEW	1,809	Videos	Categorical Emotions	Wild
AMIGOS	40 Participants	Videos	Arousal-Valence	Laboratory controlled
DFEW	16,000	Videos	Categorical Emotions	Wild
FERV39K	39,000	Videos	Categorical Emotions	Wild
OMG Emotion Dataset	7371 utterances	Videos	Arousal-Valence, Categorical Emotions	Wild
MAFW	10,000	Videos	Categorical Emotions	Wild
NESS	69 Participants	Videos	Symptoms of Schizophrenia (Continuous)	Wild

emotional categories.

2.6 Datasets

A table of the affective datasets used in this thesis can be found in Tab. 2.1. The datasets used in this thesis, are chosen based on the experimental set-up of each chapter and to allow for a fair comparison with previous works in each task.

Chapter 3

Label Uncertainty in Human Affect

Chapter Abstract

As discussed in Chapter 1 of this thesis, we begin by addressing problems in continuous labels and, more specifically, directly addressing label noise in continuous affect. Continuous affect estimation is a problem where there is inherent uncertainty and subjectivity in the labels that accompany data samples – typically, datasets use the average of multiple annotations or self-reporting to obtain ground truth labels. In this chapter, we propose a method for uncertainty-aware continuous affect estimation that explicitly models the uncertainty of the ground truth label as a univariate Gaussian with a mean equal to the ground truth label and unknown variance. For each sample, the proposed neural network estimates the value of the target label (valence and arousal in our case) and the variance. The network is trained with a loss defined as the KL divergence between the estimation (valence/arousal) and the Gaussian around the ground truth. We show that, in two affect recognition problems with real data, the estimated variances are correlated with measures of uncertainty/error in the labels extracted by considering

multiple annotations of the data or by manually cleaning the dataset.[1]

Contents

3.1 Introduction

Affect recognition in the wild is a problem that traditionally uses the labels assigned by expert annotators or self-reporting as the ground truth. Even though the labels obtained in that manner are not as noisy as, for example, through social media scraping, there is an inherent element of subjectivity in the annotation that can be regarded as noise or bias. This subjectivity in available affect datasets can have an effect on the generalisation and interpretability of results.

In recent years, several works have attempted to address label uncertainty. DivideMix [69] introduces a methodology for training on noisy labels by leveraging a semi-supervised technique. The method simultaneously trains two networks and uses the per-sample training loss to co-divide the data into a clean- and a noisy-label subset. However, the methodology proposes a hard label correction by assigning pseudo-labels on noisy samples during training and requires co-training of two networks. In the regression framework, He et al. [50] model the difficulty in predicting object boundaries in object detection by estimating the uncertainty in predicting the bounding box in the form of variance and introducing a Kullback-Leibler (KL) based loss term that allows the estimation of the variance for each predicted boundary. However, none of the above

[1] Portions of this chapter are published: 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)

have been introduced in the domain of affective computing for continuous arousal and valence estimation.

In this work, we adopt a similar approach and build on the work of He et al. [50] in order to address the problem of label uncertainty in the domain of affective computing. We address the problem of affect estimation as a regression problem predicting a continuous value for arousal and valence. We propose to estimate the uncertainty of the label for each sample in the form of variance so that the model estimates both the target and the label variance. By contrast to approaches such as DivideMix [69] that model the distribution of the loss over multiple samples and make a hard decision between which samples are noisy and clean, our measure is continuous and is derived per sample by a branch of the network. Our network is trained on a KL-divergence-based loss using standard back-propagation. We evaluate the methodology on two continuous affect datasets, namely AMIGOS [83] for video affect estimation and AffectNet [84] for affect estimation in static images. We show that the derived measure is positively correlated to the variance of annotators in AMIGOS where multiple annotations are available. In AffectNet, where multiple annotations are not available, we use the rules proposed by [112] to obtain a clean and a noisy validation set and show that the estimated variances in the clean subset are lower than in the noisy one by performing a statistical significance test. Finally, we show that the proposed methodology consistently improves the performance in both datasets against their baselines.

The main contributions of this Chapter can be summarised as follows:

1. We propose addressing the problem of continuous affect estimation with label uncertainty by modelling the ground truth label as a univariate Gaussian distribution with unknown variance and training a network that learns to predict it. To the best of our knowledge, this is the first work doing so in this domain.
2. We show that the proposed methodology improves the performance upon the

adopted baselines on both image and video data affect recognition problems.

3. We quantitatively evaluate the predicted variance metric as a measure of uncertainty and show that it is positively correlated with the variance of multiple human annotators in AMIGOS and higher in part of AffectNet that were deemed to contain noisy samples.

The Chapter is organised as follows. Section 3.2 introduces our methodology, Section 4.3 introduces the experimental setup, Section 4.4 reports the results, and Section 5.5 concludes the Chapter.

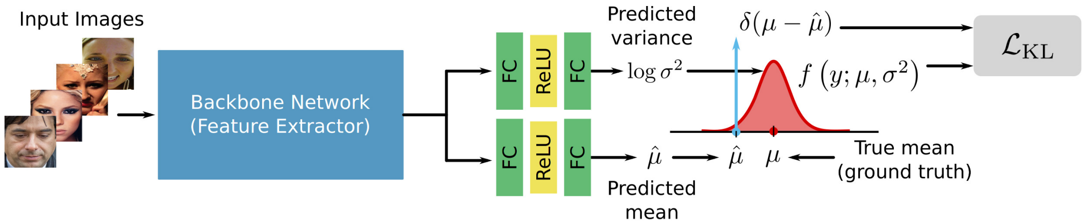


Figure 3.1: Proposed method overview: A backbone convolutional neural network is applied to input images in order to extract features which are subsequently used by two MLP heads in order to predict a) the variance σ^2 (top branch) and b) the mean $\hat{\mu}$ (bottom branch) of the annotation $y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ for a given training sample. A KL-divergence loss function is then used to measure the difference between the Gaussian distribution $f(y;\mu,\sigma^2)$ and the Dirac delta distribution $\delta(\mu-\hat{\mu})$.

3.2 Methodology

In tasks where multiple annotations per sample are available (specifically in emotion and affect recognition), majority voting or averaging over the given multiple labels approaches are typically followed in order to obtain a single ground truth label per sample. Such methods, however, neglect the uncertainty that is inherent in such annotations and that are introduced by multiple, usually disagreeing, annotators. Furthermore, multiple annotations per sample are not always available, making methodologies that explicitly handle label uncertainty in the data not applicable. In this section, we present our method for a) modelling the aforementioned uncertainty in the given annotations and b) using it in order to predict both the (ground truth) mean value of the label and its (unknown) variance. By doing so, we expect to estimate an interpretable metric for label uncertainty and improve the performance of affect estimation. An overview of the proposed method is shown in Fig. 3.1.

3.2.1 Ground truth uncertainty estimation

We begin by modelling the ground truth annotations as a set of independent uni-variate Gaussian distributions, for which we are given the true mean values (ground truth), and we try to predict both the mean values and the corresponding variances. More specifically, let $y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ denote an annotation label (e.g., the value of arousal for a given sample) with true mean value μ and unknown variance σ^2. For doing so, we jointly optimise a convolutional feature extractor backbone network and two MLP "heads", one predicting the mean and the other predicting the variance of the respective Gaussian, as shown in Fig. 3.1.

We achieve this by optimising a KL-divergence based loss function, $\mathcal{L}_{\mathrm{KL}}$, which measures the difference between the predicted Gaussian, which is uniquely expressed by its true mean μ and the predicted variance σ^2 and its density is given by $f(y; \mu, \sigma^2)$, and a Dirac delta distribution centred at the predicted mean value $\hat{\mu}$, with density given by $\delta(\mu - \hat{\mu})$ (see Fig. 3.1).

It is worth noting that, in order to impose positivity on the predicted variance and avoid exploding gradients, we implicitly predict its Napierian logarithm, $s = \log \sigma^2$, and use it as $\exp(s) = \sigma^2$, as we will show below. That is, as shown in Fig. 3.1, the top MLP predicts the logarithm of σ^2.

We note that KL-divergence is a distribution-wise asymmetric measure, which does not satisfy the triangle inequality and thus cannot serve as a true metric function. However, it is widely used for measuring the dissimilarity between statistical distributions [50, 124]. For instance, He et al. [50] incorporate a similar KL-divergence-based loss function for measuring the distance between a uni-variate Gaussian and a Dirac delta distribution.

By following similar arguments as in [50], we introduce a KL-divergence-based loss function given by

$$\mathcal{L}_{\mathrm{KL}} = \frac{(\mu - \hat{\mu})^2}{2\sigma^2} + \frac{\log \sigma^2}{2}, \tag{3.1}$$

when $|\mu - \hat{\mu}| \leq 1$, and by

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{\sigma^2}\left(|\mu - \hat{\mu}| - \frac{1}{2}\right) + \log \sigma^2, \tag{3.2}$$

when $|\mu - \hat{\mu}| > 1$. That is, in the cases where the predicted mean values are far from their true values (typically during the early training process), we use the latter modified smooth $\mathcal{L}_1$ loss term shown in (3.2), while after achieving certain convergence we use the former fine-grained and uncertainty-aware loss term (3.1).

We note that in contrast to [50] that model their regression predictions as uni-variate Gaussians and optimise their variances, we, instead, predict the variance of the ground truth values for our regression task. This reflects the intuition that affect labelling is prone to noise. The proposed loss takes into account the estimated variances of labels, unlike other losses traditionally used for regression problems (e.g. Mean Absolute Error or Mean Squared Error); for more ambiguous or noisy samples, we expect the model to estimate a higher variance.

3.2.2 Architectures

As discussed in the previous sections, in this work, we address the problem of data uncertainty on continuous affect estimation from both static images and videos. For affect estimation from static images, we set the general architecture presented in Fig. 3.1 so as the backbone feature extractor is implemented by a CNN architecture. More specifically, we have experimented with both VGG16 [106] and ResNet [49] architectures (see Fig. 3.2); however, the proposed methodology can be implemented on any appropriate network, as described in the previous section.

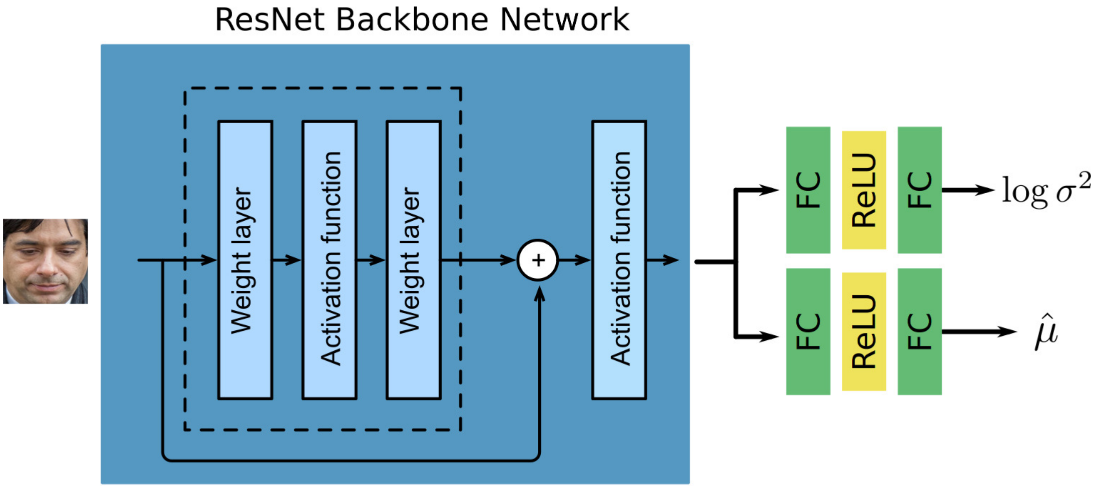


Figure 3.2: Residual CNN backbone architecture for extracting features from static images.

In the case of continuous affect estimation on untrimmed videos, our basic architecture (Fig. 3.1) is set so that video features are obtained using a CNN with a trainable NetVLAD [6] layer, as shown in Fig. 3.3. The NetVLAD architecture [6] is inspired by the Vector of Locally Aggregated Descriptors (VLAD), which is a pooling method that captures information about the statistics of local descriptors over the image, by storing the sum of residuals from cluster centres.

More specifically, the NetVLAD introduced in [6] can update the cluster centres during training; therefore, the layer can be introduced as a pooling layer in a standard convolutional architecture. The original NetVLAD layer is used to generate a $K \times D$ vector from a $W \times H \times D$ convolutional output, where K is the number of centroids to be used in the VLAD vectors, D is the number of channels of the last convolutional layer, and (W, H) are the spatial dimensions of the convolutional output, as shown in Fig. 3.3.

In this work, we modify the NetVLAD layer architecture to perform pooling along the temporal dimension instead of the spatial. The input to the network is a set of pre-computed features obtained during pre-training from each video frame. The network then performs convolutional and average pooling operations followed by ReLU activation across the temporal dimension and then uses the NetVLAD layer as a pooling layer to standardise the feature vector size. The proposed architecture using NetVLAD offers certain advantages; more specifically, it allows for the use of untrimmed video input and can handle longer sequences. It also offers a good performance versus simplicity trade-off.

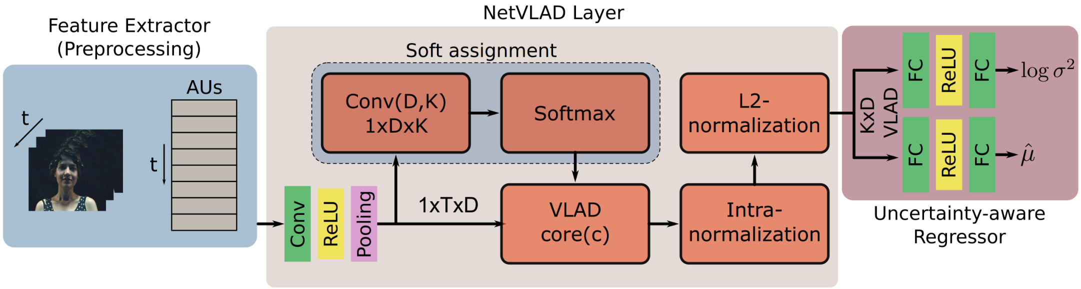


Figure 3.3: Video input architecture: Given an untrimmed video with t number of frames, we extract a vector of Action Units (AUs) per frame in the preprocessing phase. The AU time series is then used to train the NetVLAD architecture along with our uncertainty-aware regressor.

3.3 Experimental Setup

3.3.1 Datasets

AMIGOS The AMIGOS dataset [83] consists of audio-visual and physiological responses of participants (either alone or in a group) to a video stimulus. In this work, we use the responses of individuals; 40 participants watched sixteen short videos and four long ones. The former are defined as videos with length in the 50-150 second range. The responses are broken down to 20-second intervals and annotated by three annotators for *arousal* and *valence* on a scale from -1 to 1. We extracted the frames from the video with a framerate of 25 frames/sec and calculated the average score of the three annotators as the ground truth during training for the video segment. During testing, we use the variance of the annotators as an indication of uncertain or ambiguous samples and calculate the Pearson's Correlation Coefficient (PCC) between the estimated and the annotator's variance.

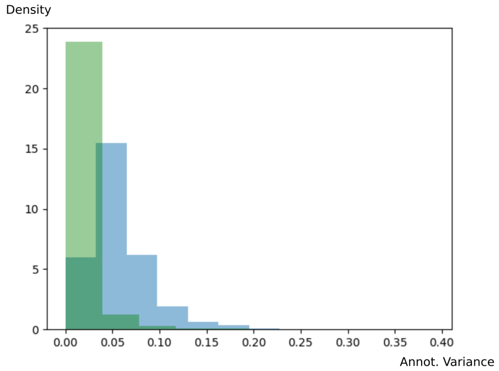


Figure 3.4: Histogram of annotators' variance in the AMIGOS dataset for *arousal* and *valence*.

As the individual annotator scores are available, we calculate the correlation matrices for arousal and valence as an indication of Inter-Annotator Agreement (IAA) in continuous affect estimation, as shown in Table 3.1. The correlations in the table indicate that there is disagreement between the annotators, particularly for arousal. A higher disagreement among annotators will introduce higher label uncertainty as it is an indication of the sample's ambiguity. By examining the histogram of variances of the available annotations in Fig. 3.4, we can see that while most samples will have low disagreement and thus low uncertainty, there is a significant number of samples with higher variance, particularly for arousal. There are multiple reasons for the annotators' disagreement in this and other datasets, however, we can summarise them along three pillars: (a) scale ambiguity, (b) sample ambiguity and (c) experimental set-up. The scale ambiguity refers to how *arousal* and *valence* are defined, which is introduced in Chapter 1 and further elaborated throughout the thesis. In a nutshell, they refer to continuous metrics with no absolute reference and are, therefore, open to interpretation. The principle has been extensively explored in previous works, where the label ambiguity is addressed directly by converting the regression problem to a ranking problem [131]. The sample ambiguity refers to noise in the input that may affect the perception of apparent emotion (by either humans or machines). An example of this could be an expression typically associated with happiness (e.g. smiling) compounded

Table 3.1: Correlation Coefficient of Annotators Scores for Arousal and Valence in the AMIGOS dataset

	Arousal			**Valence**		
	#1	**#2**	**#3**	**#1**	**#2**	**#3**
#1	1	0.54	0.62	1	0.7	0.73
#2	0.54	1	0.51	0.7	1	0.63
#3	0.62	0.51	1	0.73	0.63	1

with an expression typically associated with contempt (e.g. unilaterally raised brow). The perceived apparent emotion could then be different between annotators, where we see disagreement even among trained experts [11]. Finally, the experimental setup (that is, how the dataset was collected) could affect label noise. Given that the dataset measures a subject's response to a stimulus, the affective range is expected to be smaller than when subjects actively take action. This results in more subtle differences between data points that are harder for annotators to distinguish, particularly for arousal, which is, in essence, measuring the level of excitement.

AffectNet AffectNet [84] consists of more than one million facial images collected from the Internet. Approximately 440,000 are annotated manually for categorical emotions and continuous arousal and valence. In this work, we use the manually annotated samples of the eight emotion categories, namely, *Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger*, and *Contempt*, which include over 290,000 samples. Annotations from multiple annotators are not provided in the dataset.

3.3.2 Performance Measures

The performance of the proposed methodology and the baselines is assessed using three evaluation metrics, depending on the database. For experiments conducted on the AMIGOS database [83], we report the Mean Square Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i - \hat{\mu}_i\right)^2 , \tag{3.3}$$

where n is the number of videos in the database, μ_i is the ground truth and $\hat{\mu}_i$ is the predicted value, as discussed in Sect. 3.2. To better assess the performance of the regression task and to guarantee that results are comparable with other methods that apply transformations on the labels, we use Pearson's Correlation Coefficient (PCC), which for a pair of variables x, y with means $\bar{x}, \bar{y}$ is given by

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2 (y_i - \bar{y}_i)^2}} \in [-1, 1] \tag{3.4}$$

The above equation is used to evaluate both the performance of the regression when predicting the level of arousal/valence and the quality of the learnt variance, where a PCC close to one implies a one-to-one relationship between the two variables (so in simple terms higher is better). In addition to PCC, we also evaluate the performance of our method in the regression task in AffectNet using the Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)^2}, \tag{3.5}$$

where n, μ_i, and $\hat{\mu}_i$ denote the number of images, the ground truth, and the predicted value for arousal/valence, respectively.

3.3.3 Backbone and Implementation Details

Affect estimation in videos We evaluate the proposed method in the task of affect estimation in untrimmed videos using the AMIGOS [83] dataset. For this, we use ResNet50 as a backbone architecture (Fig. 3.2), which we have pre-trained on the CelebA [74] and the EmotioNet [13] datasets for the task of Action Units (AUs) recognition [35]. We use this pre-trained backbone in our preprocessing phase (Fig. 3.1) in order to extract features. Therefore, the input to the model is a time series of ten AUs per video segment. The AUs-based features extracted by the backbone are then used to train a simple CNN architecture using a NetVLAD [6] layer to produce a fixed-dimensional feature vector that is then fed forward to the regression and variance estimation fully connected (FC) layers, as shown in Fig. 3.3. We chose a trainable NetVLAD layer as a baseline since it offers a low simplicity-vs-performance trade-off. The 1D convolutional and average pooling layers are set with a kernel size of 7 and stride 5 and the same number of channels according to the input. As we do not downsample frames in the video sequence, we assume neighbouring frames will have similar values and, therefore, implement a larger kernel and stride. The NetVLAD layer is initialised with eight centroids. The training is performed in an end-to-end manner, and we follow a leave-one-subject-out cross-validation protocol for each subject in the individual database until the network converges. The network is trained using an ADAM optimiser with an initial learning rate of 0.01 multiplied by a factor of 0.1 every 100 epochs and a batch size of 512 on two NVIDIA RTX 2080 GPUs.

Affect estimation in static images In the case of affect estimation in static images, we evaluate the proposed method using both the VGG16 [106] and the ResNet50 [49] architectures as a backbone (Fig. 3.2), in order to assess the effect of the variance prediction and KL divergence loss. We also train a ResNet18 network and initialise convolutional layers with weights pre-trained on ImageNet. All networks are trained using Stochastic Gradient Descent (SGD) optimisation, with an initial learning rate of 0.0001 multiplied by 0.8 after 100 epochs and a batch size of 128 until convergence.

3.4 Results and Discussion

In order to assess the impact of the learned variances, we compare them with the corresponding variances induced by annotators' disagreement – when multiple annotators' scores are available, we can estimate uncertainty in the form of variance between annotators' scores. We propose to evaluate the learned variances against the annotator's variances at test time. It is worth noting that, unlike [98, 75, 28, 48], we do not use the annotator's variance in the training phase as a target, but instead we learn each annotation's variance from input and evaluate in the test phase.

Table 3.2: PCC of learned variance and annotators variance on AMIGOS dataset

	Arousal	**Valence**
Proposed method	0.34	0.31

In the AMIGOS dataset, we use the PCC, given by (4.3), to calculate the correlation between the learned and the annotators' variances, and we show the results in Table 3.2. We observe a higher PCC for arousal, which also had a lower IAA, as seen in Table 3.1. This is an indication of the model's understanding of ambiguity. Examples of clips with low and high predicted variance from the AMIGOS dataset are shown in Fig. 3.5.

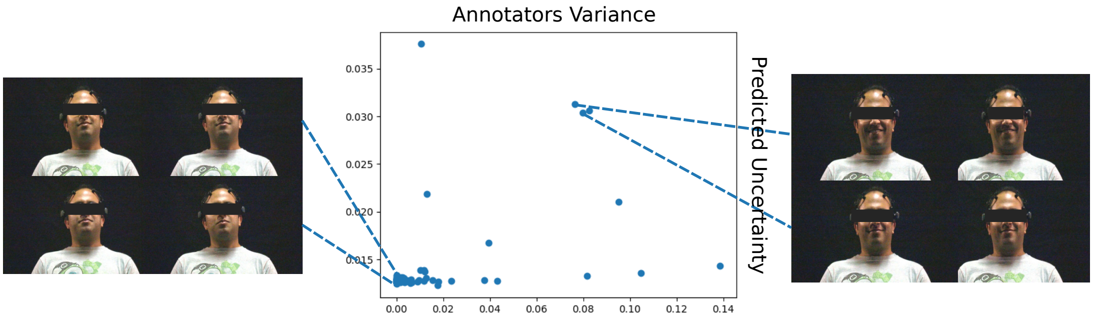


Figure 3.5: Examples of clips with low predicted variance (left – annotators assessments: $0.36, 0.12, 0.14$) and high predicted variance (right – annotators assessments: $0.77, 0.21, 0.49$) from a given subject.

In order to split the evaluation set of AffectNet into a clean and a noisy subset, we follow the rules proposed in [112]. That is, we split the evaluation set based on the categorical and continuous affect labels since multiple annotations per sample are not available. More specifically, for each sample in the evaluation set, we compare the categorical emotions to their theoretical equivalent in the arousal-valence circumplex and ensure that the assigned label for arousal and valence is in agreement with the arousal and valence of the categorical emotions. For example, a sample with the assigned emotion "Happy" in the categorical model but negative arousal would be excluded from the clean set. Examples from the two subsets can be seen in Fig. 3.7. In the top row, we show examples where the categorical emotion is consistent with the continuous arousal and valence, while in the bottom row, examples of noisy samples are presented. In total, 141 samples are flagged as noisy.

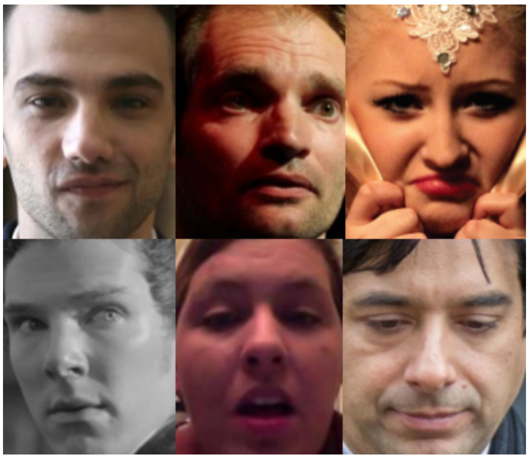


Figure 3.6: Examples of samples with clean (top) and noisy (bottom) labels. Top – from left to right, the assigned labels are: "Contempt, Arousal: 0.65, Valence:-0.65", "Fear, Arousal: 0.53, Valence: -0.06", "Sad, Arousal: -0.24, Valence: -0.66". Bottom – from left to right the assigned labels are: "Fear, Arousal: -0.32, Valence: -0.08", "Neutral, Arousal: -0.23, Valence: -0.37", "Neutral, Arousal: -0.29, Valence: 0.36".

We then estimate the variance for each sample in the subsets and compare the hypothesised population variances using a student t-test. The resulting average predicted variance for each subset is shown in Table 3.3. The estimated variances are obtained using the ResNet18 architecture initialised with ImageNet weights. Assuming the null hypothesis $H_0 : \sigma_{clean} = \sigma_{noisy}$ and the alternative hypothesis $H_1 : \sigma_{clean} < \sigma_{noisy}$, we perform a one-tailed Student's t-test. We compute t as follows

Table 3.3: Mean estimated variance for Arousal and Valence on AffectNet subsets. As the scale of the variance is small, results are shown in three significant figures.

	Samples	**Arousal(std)**	**Valence(std)**
AffectNet clean	3858	0.078(0.003)	0.079(0.004)
AffectNet noisy	141	0.082(0.003)	0.087(0.002)

$$t = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{3.6}$$

where x_i and s_i represent the means and variances of the two samples, respectively, and n_i is the respective sample size. With the values from Table 3.3, t is estimated at -0.91 and -1.96 for arousal and valence, respectively. The calculated p values with 139 degrees of freedom for arousal and valence are 0.18 and 0.025, respectively.

Therefore, we can reject the null hypothesis for valence at a 95% confidence interval but not for arousal. As the use of C.I. is dependent on both the problem and how much uncertainty is acceptable for it, we want to note that we can reject the null hypothesis for arousal with a lower C.I. of 80%. While the use of a lower C.I. is atypical for most tests of statistical significance, we want to emphasize that, in this case, a test with lower confidence successfully shows a relationship between estimated variance and label noise. The weak relationship, shown by accepting the null hypothesis with lower C.I., is also a testament to the difficulty of the problem, as well as evidence of other entangled factors affecting label noise. The distributions of the estimated uncertainty for the two subsets are shown in Fig. 3.7. In the plotted distributions, we can visually confirm the differences between the estimated uncertainty for arousal and valence between the sets. While there are some overlapping areas between the distribution of estimated variances of the clean and noisy sets, the mean of the distribution is higher for the noisy set on both targets. It is worth noting that, as we can see in Table 3.3, the noisy subset is much smaller than the clean one. The subset sizes are important in statistical tests, as they can affect confidence intervals or not be representative of the population. However, as a rule of thumb, sample sizes of 100 are expected to give statistically significant results [61, 45]. With this in mind, and assuming there are no statistical errors in sampling, which we would expect from a large enough dataset, we can conclude that the results with 139 degrees of freedom can be accepted as robust.

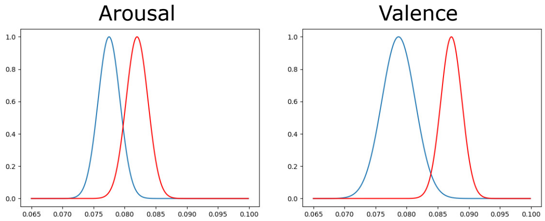


Figure 3.7: Distribution of the estimated uncertainty in Arousal (left) and Valence (right) for clean (blue) and noisy (red) labels in AffectNet.

In order to evaluate the proposed methodology and the impact of predicting variance on the overall model performance, we compare the architectures against their baseline

Table 3.4: Results on AffectNet using VGG16 and ResNet backbones

	Arousal		**Valence**	
	RMSE	**PCC**	**RMSE**	**PCC**
wideResNet	0.35	0.54	0.40	0.60
wideResNet proposed	0.35	0.55	0.41	0.61
VGG16	0.35	0.54	0.39	0.62
VGG16 proposed	**0.34**	0.55	0.40	0.62
ResNet18 (pre-trained)	0.34	0.55	0.40	0.62
ResNet18 (pre-trained) proposed	0.35	**0.57**	**0.39**	**0.63**

trained without variance prediction and an MSE loss. The results for AffectNet and AMIGOS are shown in Tables 3.4 and 3.5, respectively. We can see that the improvement in terms of PCC is consistent with estimation from both static image input and time-series input. In the AffectNet (static images), we have experimented with three different backbone architectures, namely VGG16 [106] and two variants of ResNet [49], obtaining consistent improvements in terms of the PCC. The architectures tested are simple uni-modal feed-forward networks as we aim to demonstrate the impact of uncertainty prediction. A higher predicted variance for an uncertain sample allows the network to learn from less ambiguous samples as the optimiser will prioritise lowering the $|\mu - \hat{\mu}|$ term in (3.1) and (3.2). Furthermore, by penalising the regression prediction less for uncertain samples, the predicted variance regularises the error.

Finally, for reference, we note that the results on AMIGOS align with previous work from [85], although not directly comparable, as different features and architectures are used. Specifically, in [85] Quantised Local Zernike Moments (QLZM) computed from the per frame facial landmarks were used to train an SVR and an LSTM architecture. In contrast, in our case, we used a simple frame-based estimation of a set of Facial Action Units. Moreover, while there are some methodological parallels between the NetVLAD architecture used and Fisher Vectors of QLZM used to train the SVR, recurrent methodologies better capture the temporal dimension of features, which is significant in continuous affect. The SVR architecture in [85] achieves a PCC of

Table 3.5: Results on AMIGOS using precomputed per frame Facial Action Units as input and a NetVLAD architecture.

	Arousal		**Valence**	
	MSE (std)	**PCC**	**MSE (std)**	**PCC**
NetVLAD	0.03 ($3e^{-}3$)	0.50	0.02 ($2e^{-}3$)	0.47
NetVLAD proposed	0.04($6e^{-}3$)	**0.53**	0.02($2e^{-}3$)	**0.52**

0.34 for both arousal and valence, while the LSTM architecture achieves 0.6 and 0.62, respectively.

3.5 Conclusion

Continuous affect estimation is an inherently uncertain problem due to the subjective and ambiguous nature of continuous labels. We have proposed estimating the level of continuous affect along with a certainty metric that represents the true variance in the label distribution of continuous arousal and valence. The methodology is inspired by work on other domains with label uncertainty, such as bounding box regression, but to our knowledge, this is the first work addressing the problem in affective computing by treating the ground truth as a Gaussian distribution and the predicted level of affect as a Dirac delta function. We evaluate our methodology on two datasets, AMIGOS [83] and AffectNet[84] for affect estimation from video and static images, respectively and find that it improves upon the baselines for all architectures tested. We also evaluate the learned uncertainty metric by comparing the learned variance against the annotators' variance when multiple annotations per sample are available. We find a positive correlation between the estimated uncertainty and the disagreement between annotators. When multiple annotations are not available, we compare the distribution of the predicted variance on clean and noisy evaluation subsets and find the estimated uncertainty in the clean set lower using a statistical test. The proposed methodology offers a measure for label uncertainty in continuous affect recognition.

Chapter 4

Label Relationships in Human Affect and Mental State Estimation

Chapter Abstract

In this second methodological chapter, we continue addressing human affect as a continuous task along the arousal-valence axis and extend the methodology to the task of schizophrenia symptom estimation. Human affect and mental state estimation in an automated manner face several difficulties, including learning from labels with poor or no temporal resolution, learning from few datasets with little data (often due to confidentiality constraints) and (very) long, in-the-wild videos. For these reasons, deep learning methodologies tend to overfit and arrive at latent representations with poor generalisation performance on the final regression task. To overcome this, in this chapter, we introduce two complementary contributions. First, we present a novel relational loss for multilabel regression and ordinal problems that regularises learning and leads to better generalisation. The proposed loss uses label vector inter-relational information to learn better latent representa-

tions by aligning batch label distances to the distances in the latent feature space. Second, we utilise a two-stage attention architecture that estimates a target for each clip by using features from the neighbouring clips as temporal context. We evaluate the proposed methodology on both continuous affect and schizophrenia severity estimation problems, as there are methodological and contextual parallels between the two. Experimental results demonstrate that the proposed methodology outperforms the baselines trained using the supervised regression loss and pre-training the network architecture with an unsupervised contrastive loss. In schizophrenia symptom estimation, the proposed methodology outperforms previous state-of-the-art by a large margin, achieving a PCC of up to 78%, performance close to that of human experts (85%) and much higher than previous works (uplift of up to 40%). In the case of affect recognition, we outperform previous vision-based methods in terms of CCC on both the OMG and the AMIGOS datasets. Specifically for AMIGOS, we outperform previous SoTA CCC for both arousal and valence by 9% and 13%, respectively. In the OMG dataset, we outperform previous vision works by up to 5% for both arousal and valence. [1]

Contents

4.1 Introduction

Understanding human affect and mental state is an active research area with multiple potential applications spanning fields such as education [128], healthcare [107], and entertainment [80, 102]. For example, by understanding human emotion, the user

[1] Portions of this chapter are published: N. M. Foteinopoulou and I. Patras, ‘Learning from Label Relationships in Human Affect’, in Proceedings of the 30th ACM International Conference on Multimedia, Lisboa Portugal: ACM, Oct. 2022, pp. 80–89

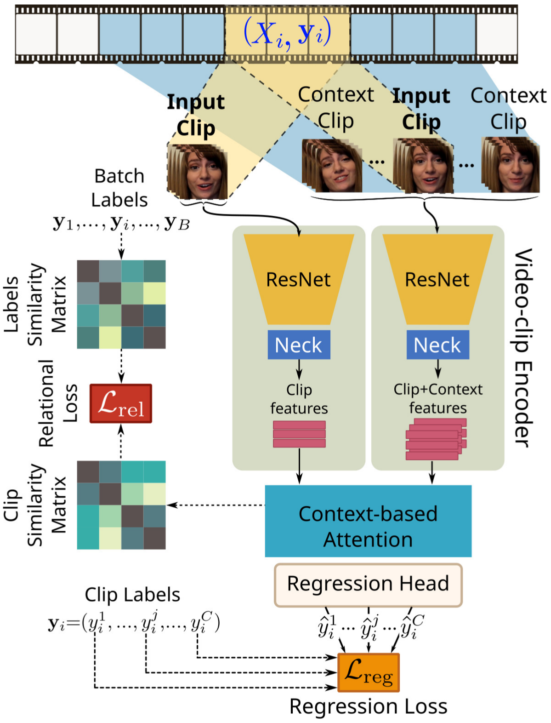


Figure 4.1: Preview of the proposed framework; the main contributions of this work are: (a)A two-stage architecture that uses features from the clip's neighbourhood to introduce context information in the feature extraction and (b) a novel relational regression loss that aims at learning from the label relationships of the samples during training

experience can be enhanced, and healthcare professionals can more effectively monitor the patient's emotional state. These problems can be treated either as a classification, using the basic human emotions [35] or by utilising continuous labels along the Arousal-Valence axes [99]. Similarly, in the domain of mental illness, several scales have been used by healthcare professionals to assess the severity of the symptoms, thus treating symptoms as a spectrum [5].

Regardless of which of the above labelling approaches is adopted, certain issues render the problem of human affect and mental state estimation challenging. Specifically, in-the-wild datasets tend to include long videos with low or no temporal label resolution – i.e., a set of labels describes the entire video. This typically occurs as affect and mental health symptom labels refer to abstract behaviour that is not easily captured and is not always objectively defined. The length of the video poses a major difficulty for Machine Learning methods due to GPU memory constraints. To address this issue, two main approaches are employed in the literature, namely, a) estimating sub-segments of the long videos [22, 76] and b) pre-computing features[136, 130, 17, 85]. For example, in MIMAMO [31] and the work of Peng *et al.* [89] a small number of frames is sampled from each clip. However, this disregards information from the remaining video and the clip context. Moreover, as affect and mental state descriptions often refer to a larger context, short clips might not be representative samples. Similarly, estimating per-frame predictions [81] disregards clip information and is also suffering from the lack of temporal information. Previous state-of-the-art works in symptom severity estimation [15] used statistical representations, such as Gaussian Mixture Models, on a set of per-frame extracted features. However, this approach does not learn from the temporal relationships of frame features. It also does not allow for end-to-end training and, therefore, does not allow for feature optimisation on the specific task. In order to exploit contextual information and improve clip-level features, Wu et al. [122] proposed the use of Long-Term Feature Banks for the problem of action recognition in videos. However, Long-Term Feature Banks [122] rely on a pre-computed set of features for the context that does not improve in quality during training. By contrast, in this work, we build upon [122] and use a context feature extractor that updates context features at each iteration, allowing for dynamically computing context features of random clips sampled from a longer video in an end-to-end manner, leading to much shorter training times.

Publicly available datasets for affect and mental health analysis are typically small, which often results in overfitting problems during training. As such, methods that lead to better representations with a small number of samples are paramount to the success of the final regression task. However, several recent works [59, 27, 51, 19] require pre-training (whether supervised or unsupervised) with very large datasets to achieve better representations before fine-tuning on the final task. In continuous affect estimation, Kim *et al.* [60] binarised labels and used an adversarial loss on the latent feature space; however, this approach ignores the continuous nature of Arousal/Valence dimensions.

In order to both alleviate the challenges due to long video input and to improve the feature representations so as to address the multi-label regression problems that arise in the domain of affect and mental health analysis, in this work, we propose a) a novel attention-based video-clip encoder that builds upon [122] and utilises the temporal dimension of the input clips and arrives at clip-level predictions that benefit from context clip information, and b) a novel relational regression loss function that aligns the distances in the latent clip-level representations/features to the distances of the labels of the clips in question. An overview of the proposed framework is shown in Fig. 4.1. Specifically, we propose to jointly train two network branches: a) one that uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of temporally neighbouring clips, which subsequently feed a regression head in order to infer the desired values and calculate the regression loss, and b) one that uses the proposed video-clip encoder to extract clip-level features from the input video clips, which subsequently feeds the regression head and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss. To the best of our knowledge, this is the first work that uses label relationships to improve feature representation learning. The proposed regression head employs an attention-based mechanism for fusing clip-level and context features and regressing to the desired continuous values. The main contributions of this Chapter can be summarised as follows:

- We build on [122] and propose a two-stage attention architecture that uses features from the clips' neighbourhood to introduce context information in the feature extraction. The architecture is novel in the domain of affect and mental state analysis and, unlike [122], it does not train a separate model to compute

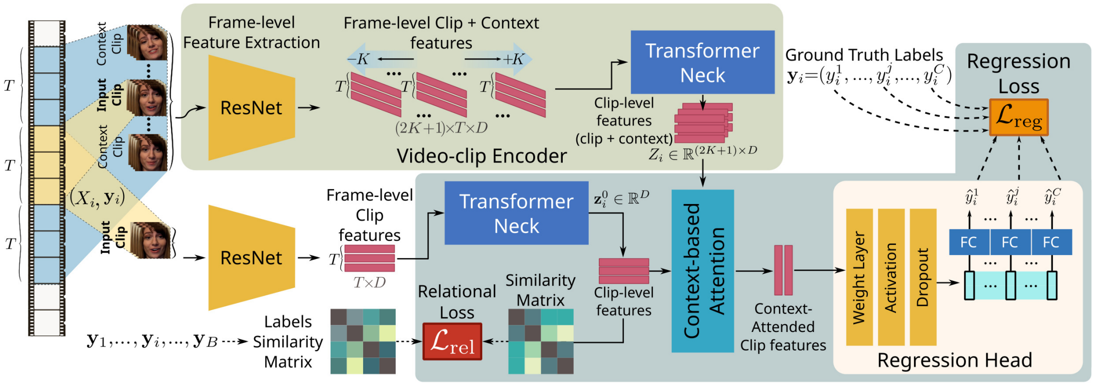


Figure 4.2: Overview of the proposed framework: (a) The *bottom branch* uses the proposed video-clip encoder (comprising of a ResNet frame-level and a Transformer clip-level feature extractors) to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss. (b) The *upper branch* uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of context clips from each of the input clips, which subsequently feed the context-based attention block in order to infer the desired values and calculate the regression loss. The context-based attention block fuses clip-level and context features and passes the context-attended clip features to the regression head that estimates the desired continuous values. Error is back-propagated only through the shaded region of the bottom branch.

context features but rather updates its weights during training – this leads to shorter training times.

- We introduce a novel loss, named relational regression loss, that aims at learning from the label relationships of the samples during training. This loss uses the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner. We show in the ablation studies that the improved latent representations obtained with the addition of the relational loss lead to improved regression output without the use of large datasets.

- We show that the methodology achieves results comparable to the state-of-the-art. Specifically, for symptom severity estimation of schizophrenia, our methodology outperforms the previous state of the art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

4.2 Proposed method

An overview of the proposed framework for the problem of multi-label regression from a sequence of clips is given in Fig. 5.1. In a nutshell, the proposed architecture consists of two branches with shared weights that incorporate two main components: a) a video-clip encoder employing a convolutional backbone network for frame-level feature extraction and b) a Transformer-based network leveraging the temporal relationships of the spatial features for clip-level feature extraction (Sect. 4.2.1). The clip and context features produced by the aforementioned branches are passed to a context-based attention block (Sect. 4.2.2) and a regression head (Sect. 4.2.3). The proposed method uses the context-based attention block to incorporate features from the two branches before passing them to the regression head, as shown in Fig. 5.1. The bottom branch uses the proposed video-clip encoder to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss (Sect. 4.2.4). The goal of the proposed relational loss, as an additional auxiliary task to the main regression, is to obtain a more discriminative set of latent clip-level features by aligning the label distances in the mini-batch to the latent feature distances. Finally, the upper branch uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of context clips from each of the input clips, which subsequently feed the regression head in order to infer the desired values and calculate the regression loss.

4.2.1 Video-clip encoder

Let $\mathcal{X}$ be a batch of labelled clips designed so as it contains consecutive clips taken from different video sequences; i.e., $\mathcal{X} = \{(X_i, \mathbf{y}_i)\}_{i=1}^{B}$, where $X_i \in \mathbb{R}^{T\times H\times W\times 3}$ denotes the i-th clip in the mini-batch, T denotes its duration in frames, H, W denote the frame height and width, $\mathbf{y}_i = (y_1, \ldots, y_C) \in \mathbb{R}^C$ denotes the corresponding ground truth label vector with continuous annotation for C classes, and B denotes the mini-batch size.

Given an input clip X_i, the proposed video-clip encoder extracts frame-level features by feeding them to a backbone convolutional network (e.g., a ResNet [49]), which subsequently feeds a Transformer-based network for extracting clip-level features, leveraging this way the temporal relationships of the calculated spatial features. In the proposed framework, we use the above video-clip encoder in both branches as shown in Fig. 5.1 – i.e., for calculating the clip-level features $\mathbf{z}_i^0 \in \mathbb{R}^D$ for the input clips X_i, $i = 1, \dots, B$ (bottom branch) and for calculating clip-level features $Z_i = \left(\mathbf{z}_i^{-K}, \dots, \mathbf{z}_i^0, \dots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ from each X_i along with a number K of context clips before and after it (upper branch).

4.2.2 Context-based Attention

As discussed above, for any given clip X_i and $2K$ context clips around it, the proposed video-clip encoders extract the clip-level features $\mathbf{z}_i^0 \in \mathbb{R}^D$ (corresponding to the input clip X_i alone) and $Z_i = \left(\mathbf{z}_i^{-K}, \dots, \mathbf{z}_i^0, \dots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ (corresponding to the input clip X_i and K clips before and K clips after it). These features are then fed to the regression head (Fig. 5.1), where they are first passed through an attention module before being concatenated. The resulting context-attended clip features are passed to the regression head for the final regression task.

4.2.3 Multi-label regression head

The context-attended clip features obtained through staged attention, as explained in the previous sections, is passed through an MLP regression head that predicts the regression values $\hat{\mathbf{y}}_i = (\hat{y}_i^1, \dots, \hat{y}_i^C)$, $i = 1, \dots, C$. Finally, we calculate the regression loss $\mathcal{L}_{\text{reg}}$ by either using the Root Mean Square Error (RMSE) or the Concordance Correlation Coefficient (CCC), depending on the task at hand, as we will discuss in Sect. 4.3.

4.2.4 Relational loss

At each training iteration, after having calculated (as discussed in Sect. 4.2.1) the clip-level features for the clips in a mini-batch, i.e., $\mathbf{z}_i^0 \in \mathbb{R}^D$, $i = 1, \dots, B$, we calculate the proposed relational loss as follows:

$$\mathcal{L}_{\text{rel}} = \sqrt{\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \left(\hat{M}_{i,j} - M_{i,j} \right)^2} \tag{4.1}$$

where $\hat{M} \in \mathbb{R}^{B \times B}$ denotes the cosine similarity matrix calculated on the clip-level features, whose (i, j)-th element is given as

$$\hat{M}_{i,j} = \frac{\mathbf{z}_i^0 \cdot \mathbf{z}_j^0}{\|\mathbf{z}_i^0\| \|\mathbf{z}_j^0\|},$$

and $M \in \mathbb{R}^{B \times B}$ denotes the cosine similarity matrix calculated on the ground truth labels, whose (i, j)-th element is given as

$$M_{i,j} = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|}.$$

It is worth noting that, for the calculation of the proposed relational loss, we use the clip-level features from the given clips without using any context clips, in contrast to the regression loss where additional context clips are being used, as discussed in Sect. 4.2.3. The total loss is then calculated as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{rel}}$, where λ is a weighting hyper-parameter which we discuss in Sect. 4.3.

4.2.5 Implementation details

Backbone frame-level feature extractor

We use a standard ResNet50 [49] pre-trained on VGGFace2 [21] and fine-tuned on FER2013 [44] as described in [3]. The classification layer of the pre-trained network was replaced with a fully connected (FC) layer that was fine-tuned for our task during the training of the network, followed by a ReLU [43] activation. The adopted backbone network receives an input of shape $H \times W \times 3$, where H, W are the height and width of the input frame, respectively, and are set to 224 pixels, and outputs a feature vector with 2048 dimensions for each frame. The per-frame feature vectors are stacked to a matrix of size $T \times 2048$ for each clip, where T is the number of frames of each input clip.

Transformer neck clip-level feature extractor

A transformer encoder architecture is employed to learn from the temporal relationships of the spatial feature vectors calculated by the convolutional frame-level feature extractor. The $T \times 2048$ features are positionally encoded and fed forward to a Transformer Encoder [116]. An element-wise addition is performed between the transformer encoder output and the frame-level features, followed by an average pooling operation along the temporal dimension, resulting in a D-dimensional clip-level representation, where $D = 2048$.

Context-base Attention

For each input clip X_i, the regression head takes as input both the clip-level features $\mathbf{z}_i^0 \in \mathbb{R}^D$ and the stacked context features $Z_i = \left(\mathbf{z}_i^{-K}, \ldots, \mathbf{z}_i^0, \ldots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ (Sect. 4.2.1). A modified non-local block [122] is then used as an attention operation, where clip-level features $\mathbf{z}_i^0$ are used as the query values to attend to features in Z_i, which are used as keys and values. The output context attention vector is concatenated with the clip-level features, resulting in a $2 \times D$ dimensional vector.

Regression head

The penultimate feature vector is obtained by passing the context-attended feature vector through an FC layer followed by a ReLU activation and a dropout layer.

Finally, in order to obtain the final regression predictions, we split the aforementioned penultimate feature vector into C subsets and attach an FC layer to each subset to obtain the final regression predictions. In the case of continuous affect estimation, we set $C = 2$ (i.e., for Arousal/Valence estimation), while for the schizophrenia symptom severity estimation, we set C accordingly to the number of symptoms provided by the scale at hand. Specifically, the CAINS-EXP scale has 4 symptoms in total; therefore, we set $C = 4$. The PANSS Negative scale has 7 symptoms in total, however, we select 3 for comparison with previous works [15, 115]. As the PANSS-NEG scale includes a number of symptoms we do not consider, we add an additional subset in the penultimate feature vector so that $C = 4$, which is only considered in the total score estimation. We note that, in the case of symptom severity estimation, additionally to each individual symptom prediction, we predict a total score (by using an additional FC layer) using the entire aforementioned penultimate feature vector. This is in contrast to [15], where the total score is estimated using the individual symptom scores.

4.3 Experimental setup

4.3.1 Datasets

NESS: The dataset was originally collected to study the effect of group body psychotherapy on negative symptoms of schizophrenia [91]. The participants in this study were recruited from mental health services from different parts of the UK. In total, 275 participants were interviewed at three different stages of the study: a) a baseline, b) at the end of the treatment, and c) after six months. Each clinical interview recording is between 40 and 120 minutes long and is performed in the wild, reflecting this way the conditions of real-life clinical interviews. Each interview is assessed in terms of two symptom scales, namely, PANSS [55] and CAINS [40]. Out of the total 275 patients, 110 were accepted to be recorded at baseline, 93 at the end of treatment, and 69 in the six-month follow-up. The videos in the dataset were recorded at various resolutions and frames per second. However, we standardised the resolution to 1920×1080 and fps to 25 frames/s for all videos, and we discarded videos where a face was not detected on more than 10% of the frames. Training and evaluation were performed on videos recorded at baseline for a fair comparison with works in the literature, i.e., 113 videos for 69 patients. All results reported on this dataset are based on a leave-one-patient-out cross-validation scheme, where videos were down-sampled to 3 fps. The values for "Total Negative" and "EXP - Total" in the PANSS and the CAINS scales, respectively, were scaled during training to match the range of individual symptoms (i.e., 1-7 for PANSS and 0-4 for CAINS).

OMG: The "OMG-Emotion Dataset" [10] consists of in-the-wild videos of recorded monologues and acting auditions collected from YouTube. Multiple annotators separated each clip into utterances and assigned labels for Arousal in the $[0, 1]$ scale and Valence in the $[-1, 1]$ scale. The dataset originally consisted of training, validation, and test sets with a total of 7371 utterances. As a number of videos have been removed since the publication of the dataset, we trained on 2071 and evaluated 1663 utterances. We also scaled Arousal to $[-1, 1]$ to match the range of Valence during training and inference.

AMIGOS: The AMIGOS dataset [83] consists of audio-visual and physiological responses of participants (either alone or in a group) to a video stimulus. In this work, we used the responses of individuals, i.e., where 40 participants watched 16 short videos and 4 long ones. The former were defined as videos of 50-150 seconds. The responses were broken down to 20-second intervals and annotated by three annotators for *Arousal* and *Valence* on a $[-1, 1]$ scale. We extracted the frames from the video (6 frames/s) and calculated the average score of the three annotators as the ground truth during training for the video segment. We trained the network following a leave-one-subject-out cross-validation scheme. At each fold, we randomly selected a subset of the training data, corresponding to 20% of samples. This is to show how the relational loss can achieve state-of-the-art results using a much smaller number of samples than conventional supervised methodologies.

Table 4.1: Performance (CCC) of the proposed method against baseline and other unimodal architectures (OMG).

	Arousal	Valence
Proposed	0.26	**0.48**
Proposed w/o K	0.29	0.46
Proposed w/o K w/o $\mathcal{L}_{rel}$	0.24	0.44
Proposed w/ $\mathcal{L}_{cont.}$	0.15	0.32
Peng *et al.* [89]	0.24	0.43
Kollias and Zafeiriou [63]	0.13	0.40
CNN-3RNN [64] (trained from scratch)	0.19	0.39
CNN-3RNN [64] (pre-trained on Affwild2 [62])	**0.33**	0.47

4.3.2 Augmentation

During training, we applied data augmentation to the spatial dimensions of all datasets. Specifically, we randomly changed the contrast, saturation, and hue of frames with a factor of 0.2, and we applied random horizontal flipping and random rotations (with a range of 30°). The same set of transformations was applied to all frames within a clip. Moreover, as clips with temporal length T were selected from a larger video, we considered the clipping along the temporal dimension as an augmentation approach. More specifically, from the video sequence, we selected a random initial frame and selected T consecutive frames to form a clip. Similarly, the context clips were defined as clips with T number of frames positioned before and after the current clip in the video sequence. We looped the video if the initial frame selected did not allow us to define a complete clip. The number of frames T was set to 32 for the experiments conducted on the NESS and 16 for the experiments conducted on the OMG and the AMIGOS datasets.

Table 4.2: Effect of number of frames T in terms of CCC (OMG).

	Arousal	Valence	Mean
$T = 8$	0.25	0.41	0.33
$T = 16$	**0.26**	**0.49**	**0.38**
$T = 32$	0.19	0.40	0.30

4.3.3 Training

During training, the hyperparameter λ that scales the relational loss was empirically set to 2 for experiments conducted on the NESS and the AMIGOS datasets and to 1 for experiments conducted on the OMG dataset. During testing, the clips were generated by a sliding window over the video sequence, resulting in non-overlapping clips; the average prediction of all clips in the video was calculated to estimate the final predicted label vector. The network was trained in an end-to-end manner with a batch size of 4, 8, and 16 for the NESS, the OMG, and the AMIGOS datasets, respectively, keeping the pre-trained weights of the ResNet-50 backbone frozen. We used an Adam optimizer with an initial learning rate of 10^{-4}, multiplied by 0.1 every 5 epochs, and weight decay $5 \cdot 10^{-3}$. The hyperparameter K that controls the context window size was set to 2 for the experiments on the NESS and to 1 for the experiments on the OMG and the AMIGOS datasets. The network incorporated an RMSE loss during training for the experiments conducted on the NESS and (1-CCC) for the experiments on the OMG and AMIGOS datasets, as proposed by previous works in continuous affect [112, 31].

4.3.4 Architecture Complexity

The proposed architecture has 90M trainable parameters, distributed as 4M in the backbone, 52M in the transformer neck and 33M in the context-aware attention and regression head. We note that even though the architecture uses two branches (one for clip-level features and one for context features), the two branches share weights, which significantly reduces the number of parameters. We also note that similarly to other state-of-the-art methods [31, 63], we use a ResNet50 as our backbone network, but in contrast to them that employ an RNN architecture to explore the temporal relationships, we instead use a Transformer Encoder module. As shown in [116], the self-attention layers of the Transformer are both faster and less complex than recurrent layers (RNN) when the sequence length is shorter than the feature dimensionality, which is the case in the current architecture. Hence, the proposed method is more efficient than RNN-based two-stream methods. We report that the inference time is, on average, 28.6ms (±2ms) for a clip prediction.

4.3.5 Performance Metrics

In order to assess the performance of the proposed method against the baselines and the state-of-the-art, we used the following four evaluation metrics, following the common practice in the related literature [15, 10, 85]. More specifically, for the experiments performed on the NESS dataset, we used the Mean Absolute Error (MAE), the RMSE and Pearson's Correlation Coefficient (PCC), which are given as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \tag{4.2}$$

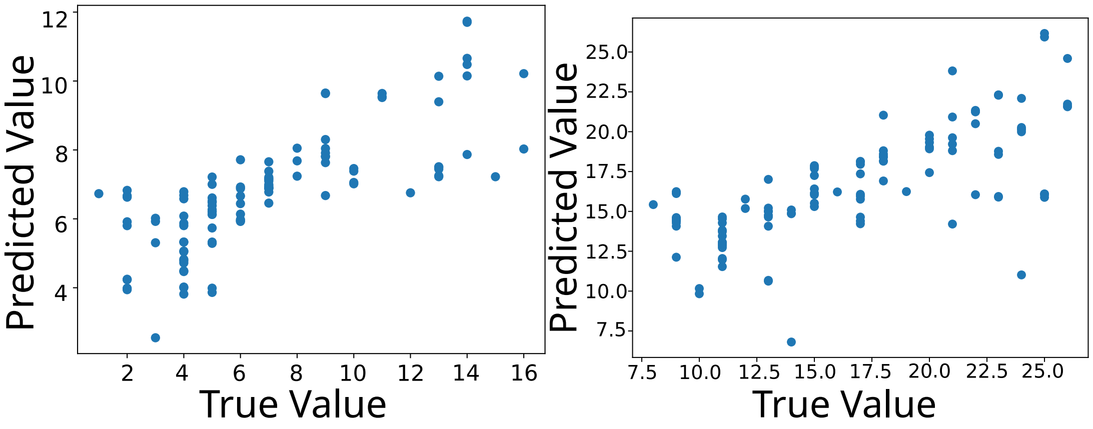
where n, y_i, and $\hat{y}_i$ denote the number of samples, the ground truth, and the predicted value for each label, respectively. To guarantee that reported results are comparable with other methods that apply transformations on the labels, we used Pearson's Correlation Coefficient (PCC), which for a pair of variables x, y with means $\bar{x}, \bar{y}$ is given by

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)\left(y_i - \bar{y}_i\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)^2\left(y_i - \bar{y}_i\right)^2}} \in [-1, 1] \tag{4.3}$$

For the experiments evaluated on the OMG and AMIGOS datasets, we used the Concordance Correlation Coefficient (CCC) as an evaluation metric. The CCC for sets $Y = \{y_1, \ldots, y_n\}$ and $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_n\}$, representing the ground truth and predicted values, is defined as:

$$CCC_{Y,\hat{Y}} = \frac{2\rho_{Y,\hat{Y}}\sigma_Y\sigma_{\hat{Y}}}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + \left(\mu_Y - \mu_{\hat{Y}}\right)^2} \in [-1, 1], \tag{4.4}$$

where μ_Y and $\mu_{\hat{Y}}$ denote the means, σ_Y^2 and $\sigma_{\hat{Y}}^2$ denote the variances of the variables Y and $\hat{Y}$, respectively, and $\rho_{Y,\hat{Y}}$ denotes the Pearson's correlation between the variables Y and $\hat{Y}$. Both eq. 4.3 and 4.4 show a high positive correlation when they approximate



(a) CAINS: EXP – Total | PCC: 0.77

PANSS: NEG – Total | PCC: 0.71

Figure 4.3: Scaled "Total Score" estimations of the proposed method on NESS using (a) CAINS and (b) PANSS scales.

(i.e. in the context of this work, higher values are better), while eq. 4.2 should be minimised.

4.4 Results and Discussion

In this section, we present the experimental evaluation of the proposed framework. We begin with our ablation study in Sect. 4.4.1 in order to demonstrate the effectiveness of our method with respect to various design options. Then, in Sect. 4.4.2, we present comparisons with state-of-the-art methods, where we show that the proposed method achieves results comparable to the state-of-the-art – specifically, for symptom severity estimation of schizophrenia; our method outperforms the previous state-of-the-art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

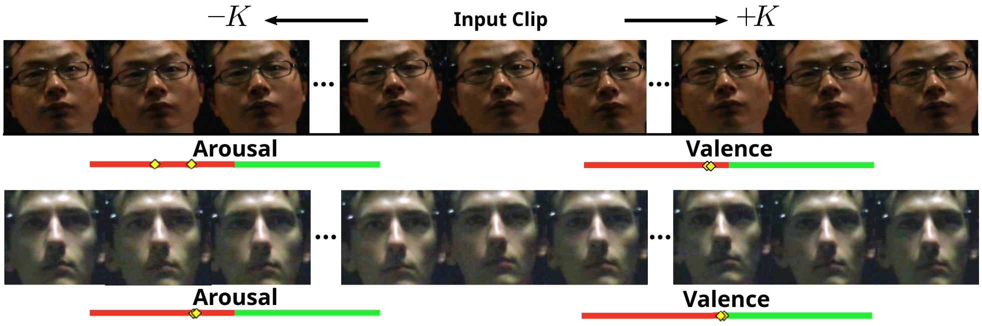


Figure 4.4: Examples of input clips, their context, and proposed method output on the AMIGOS [83] dataset: In the top row our method predicted A:-0.22, V:-0.12 (ground truth: A:-0.42, V:-0.12), and in the bottom row A:-0.29, V:-0.03 (ground truth: A:-0.29, V:-0.04).

Table 4.3: Ablation study on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Proposed	**0.87**	**1.16**	**0.78**	**0.74**	**0.95**	**0.47**	**0.64**	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**
Proposed w/o $\mathcal{L}_{rel}$	1.09	1.41	0.66	0.82	0.98	0.44	0.74	0.95	0.39	3.51	4.34	0.66
Proposed w/o $\mathcal{L}_{rel}$ w/o K	1.25	1.57	0.41	0.85	1.01	0.36	0.75	0.97	0.36	3.70	3.62	0.58
Proposed w/ $\mathcal{L}_{cont.}$	1.09	1.53	0.41	0.88	1.05	0.19	0.77	1.01	0.35	3.56	4.48	0.55

Table 4.4: Ablation study on the CAINS-EXP symptom scale.

	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Proposed	**0.56**	**0.72**	**0.75**	**0.65**	**0.89**	**0.71**	**0.71**	**0.89**	**0.76**	**0.60**	**0.82**	**0.54**	**1.88**	**2.6**	**0.77**
Proposed w/o $\mathcal{L}_{rel}$	0.59	0.78	0.63	0.75	0.98	0.60	0.72	0.96	0.59	0.62	0.85	0.51	2.12	2.94	0.71
Proposed w/o $\mathcal{L}_{rel}$ w/o K	1.06	1.33	0.64	1.06	1.36	0.59	1.14	1.37	0.62	0.77	1.02	0.44	3.76	4.72	0.48
Proposed w/ $\mathcal{L}_{cont.}$	1.12	1.37	0.45	1.06	1.35	0.54	1.19	1.49	0.41	0.84	1.11	0.26	3.87	4.80	0.41

Table 4.5: Effect of T on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
$T = 8$	0.97	1.31	0.67	**0.61**	**0.80**	**0.66**	**0.62**	0.88	0.55	3.27	4.14	0.64
$T = 16$	0.98	1.29	0.70	0.68	0.86	0.62	0.68	0.93	0.45	3.59	4.54	0.54
$T = 32$	**0.87**	**1.16**	**0.78**	0.74	0.95	0.47	0.64	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**

4.4.1 Ablation study

In order to examine the effect of the number of frames T in the overall method, we train the proposed methodology for $T = 8, 16, 32$ on the OMG and NESS datasets. The results of the ablation on T for the OMG dataset are shown in Table 4.2; we observe that the highest CCC for both arousal and valence is achieved when $T = 16$, closely followed by $T = 8$. The effect of T on the PANSS-NEG scale is shown in Table 4.5; we note that model performance is overall benefited by a larger T, with the exception of symptom N6, which is consistent with the symptom definition (i.e., Lack of Spontaneity and Flow in conversation, which is expected to be short-termed).

In order to investigate the effectiveness of the components of the proposed framework, we conducted an ablation study where we gradually excluded the incorporation of contextual clips and the proposed relational loss. For doing so, we trained a baseline network without context features and trained only on the standard regression loss (i.e., without the proposed relational loss), which we denote as "w/o K w/o $\mathcal{L}_{rel}$". We also trained a version of the network, including the context branch without the relational loss, which we denote as "w/o $\mathcal{L}_{rel}$". We finally conducted an experiment using an unsupervised contrastive pre-training, which we denote as "$\mathcal{L}_{cont.}$". In this scenario, we first pre-trained the clip-level feature extraction backbone in an unsupervised contrastive manner, and then we trained the regression head on top of the frozen backbone, using the regression loss. For the unsupervised contrastive loss, we sampled 2 clips from the same video as positive samples and considered samples from other videos as negatives.

The analysis results on the NESS [91] dataset are shown in Tables 4.3, 4.4 for the PANSS and CAINS scales, respectively. We see that the proposed network under the contrastive pre-training scenario has a similar performance to experiments where we trained with only the regression loss (shown as "w/o $\mathcal{L}_{rel}$") in terms of MAE/RMSE, however in terms of PCC the non-contrastive network still outperforms the contrastive methodology by a large margin. We attribute this to the size of the dataset that was required to learn discriminative features, as other unsupervised methodologies for representation learning [27, 26, 94] trained on very large datasets such as ImageNet [32] and Kinetics [57, 23]. Furthermore, the proposed relational clearly leads to a large improvement in the overall regression task against the baseline and the unsupervised contrastive loss using a small number of training samples. Contextual features also improved the overall regression performance, particularly for the MAE/RMSE metrics, with a more noticeable improvement in the Total Scores of the two scales.

Table 4.6: Performance of proposed method against state-of-the-art methods on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [113]	0.98	1.31	0.20	1.37	1.69	0.13	0.90	1.28	0.37	-	-	-
Tron *et al.* [115]	1.01	1.26	0.15	1.32	1.62	0.09	0.99	1.36	0.11	-	-	-
SchiNet [15]	**0.85**	1.20	0.27	1.25	1.51	0.25	0.84	1.18	0.42	3.30	4.17	0.29
Proposed	0.87	**1.16**	**0.78**	**0.74**	**0.95**	**0.47**	**0.64**	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**

Table 4.7: Performance of proposed methodology against other state-of-the-art on the CAINS-EXP symptom scale.

	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [113]	0.80	1.03	0.37	0.87	1.23	0.23	0.85	1.19	0.36	1.09	1.43	0.27	-	-	-
Tron *et al.* [115]	0.75	1.07	0.36	0.86	1.22	0.26	0.91	1.22	0.38	1.02	1.36	0.25	-	-	-
SchiNet [15]	0.66	0.93	0.46	0.77	1.10	0.27	0.90	1.15	0.36	0.98	1.30	0.30	2.67	3.34	0.45
Proposed	**0.56**	**0.72**	**0.75**	**0.65**	**0.89**	**0.71**	**0.71**	**0.89**	**0.76**	**0.60**	**0.82**	**0.54**	**1.88**	**2.60**	**0.77**

Table 4.8: Performance of the proposed method against baseline and other uni-modal architectures (AMIGOS).

	Arousal		Valence	
	PCC	CCC	PCC	CCC
Proposed	**0.69**	**0.68**	**0.75**	**0.74**
Proposed $\mathcal{L}_{CCC}$ w/o K w/o$\mathcal{L}_{rel}$	0.59	0.49	0.64	0.54
Proposed $\mathcal{L}_{RMSE}$ w/o K w/o $\mathcal{L}_{rel}$	0.60	0.39	0.55	0.40
Mou *et al.* [85]	0.60	0.59	0.62	0.61

The results of our ablation study on the OMG dataset [10] are presented in Table 4.1. Comparing the proposed methodology against its baseline (i.e., "w/o K w/o $\mathcal{L}_{rel}$"), we observe that the proposed relational loss which is the main contribution of this work improves the performance of the regression measured in terms of CCC, for both Arousal and Valence. Further incorporating the contextual features improved the CCC score for Valence but slightly lowered the CCC for Arousal. The drop in arousal affects the overall performance of the model when using K. We hypothesise that this may be due to the nature of the collected dataset (monologues, auditions, etc.), where changes in how calm or agitated a subject is are expected, thus making arousal short-lived, while the overall theme (in this case, how pleasant a scene is) would remain roughly the same. Compared to other works submitted to the challenge [63, 89], the proposed network and specifically the use of the novel relational loss, shows a clear improvement in terms of CCC for Valence, however [64] has the highest Arousal and overall performance on the OMG dataset. We also observe a clear advantage of the proposed method compared to the architecture pre-trained with contrastive loss, which appears to over-fit, and it may be encouraging the network to learn features of the subjects' identities rather than affective and mental states due to the nature of the problem and database size.

4.4.2 Comparison to state-of-the-art

In this section we present the results of the proposed method against state-of-the-art methods. The results for the NESS dataset [91] against previous works are shown in Tables 4.6 and 4.7, for PANSS and CAINS scales, respectively. We can see that the proposed methodology outperforms previous works across all the evaluated symptoms and scales by a large margin, particularly for PCC, achieving state-of-the-art results. Since the NESS dataset has been annotated by different healthcare professionals, we can compare the PCC achieved by the proposed method against the PCC of the annotators (mental health experts), which has a mean value of **0.85** [17, 91] on NESS. We observe that the proposed method achieves a PCC close to that of human experts for the "Total Negative" and "EXP-Total" scores in this dataset. In Fig. 4.3, we show the total score predictions for all videos for both scales in the NESS dataset. As the NESS dataset is imbalanced, with fewer patients having severe symptoms, we observe a higher error for patients with higher ground truth labels. Moreover, since we perform leave-one-patient-out cross-validation, there is a chance that no examples of high total scores are included in the training set of a given fold. This trend is consistent for both scales used to evaluate.

For experiments conducted on the OMG dataset [10], we compared the performance of the proposed method against other uni-modal multi-label works submitted to the "OMG-Emotion Behavior Challenge" – we show the results in Table 4.1, where we observe a clear improvement against previous works, for both Arousal and Valence in terms of CCC. We note that, to our knowledge, current state-of-the-art results for the OMG dataset are achieved by MIMAMO [31] with a CCC of 0.37 and 0.52 for Arousal and Valence respectively. However, as MIMAMO is a multi-modal approach (using RGB and inter-frame phase difference as input modalities) and is trained for a single target (i.e., Arousal or Valence) at a time, the results reported in [31] are not directly comparable to ours.

Finally, for the experiments conducted on the AMIGOS dataset [83], we compared the performance of the proposed methodology against previous state-of-the-art [85] for the face modality, and we show the results in Table 4.8. The proposed methodology leads to a clear improvement against both baselines, trained with an RMSE regression loss ($\mathcal{L}_{RMSE}$) and a CCC loss ($\mathcal{L}_{CCC}$). We also outperform previous state-of-the-art by a large margin for both Arousal and Valence, even though we trained on a subset of the training data at each fold. It is worth noting that on the AMIGOS dataset, the architecture that was pre-trained with a contrastive loss completely overfitted on the regression task. Thus, we chose to exclude it from the comparison. In Fig. 4.4, we see some visual examples of input clips, their context from the AMIGOS dataset [83] and the proposed methodology predictions against the ground truth.

4.5 Conclusion

In this work, we presented our method for dealing with challenges that arise in the domain of affect and mental health in multi-label regression problems. Specifically, we built on [122] and proposed a two-stage attention architecture that uses features from the clips' neighbourhood to introduce context information in the feature extraction. The architecture is novel in the domain of affect and mental state analysis and leads to shorter training times in comparison to state-of-the-art. Furthermore, we introduced a novel relational regression loss that aims to learn from the label relationships of the samples during training. The proposed loss uses the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner. We showed that the improved latent representations obtained with the addition of the relational regression loss led to improved regression output without the use of large datasets. Finally, we demonstrated the effectiveness of the proposed method on three datasets for schizophrenia symptom severity estimation and for continuous affect estimation, and we showed that our method achieves results comparable to the state-of-the-art – specifically for symptom severity estimation of schizophrenia, our methodology outperforms the previous state-of-the-art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

Chapter 5

EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition

In this final methodological chapter, we approach categorical emotions using natural language as the ground truth to address the large intra-class variation associated with coarse categorisation in the domain of human affect. Facial Expression Recognition (FER) is a crucial task in affective computing, but its conventional focus on the seven basic emotions limits its applicability to the complex and expanding emotional spectrum. To address the issue of new and unseen emotions present in dynamic in-the-wild FER, in this chapter, we propose a novel vision-language model that utilises sample-level text descriptions (i.e. captions of the context, expressions or emotional cues) as natural language supervision, aiming to enhance the learning of rich latent representations, for zero-shot classification. To test this, we evaluate using zero-shot classification of the model trained on sample-level descriptions on four popular dynamic FER datasets. Our findings show that this approach yields significant improvements compared to baseline methods. Specifically, for zero-shot video FER, we outperform CLIP by over 10% in Weighted Average Recall and 5% in Unweighted Average Recall on several datasets. Furthermore, we evaluate the representations from the network trained using sample-level descriptions on the downstream task of mental health symptom estimation, achieving performance comparable or superior to state-of-the-art methods and strong agreement with human experts. Namely, we achieve a Pearson's Correlation Coefficient of up to 0.85, comparable to human experts' agreement.[1]

Contents

5.1 Introduction

Facial Expression Recognition (FER) is a primary task of affective computing, with several practical applications in Human-Computer Interaction [29], education [128] and mental health [41], among others. To formalise the spectrum of human emotions, researchers have proposed several models. Ekman & Friesen [35] propose six emotions (later seven with the addition of "contempt") as the basis of human emotional expression. This is the most widely accepted model for FER; however, human emotional experience is significantly more complex and varied than the seven basic categories, with up to 27 distinct categories for emotion reported in recent studies [30]. A continuous arousal-valence scale has been proposed [99] as an alternative to creating emotion categories; however, the scale is neither objective [42] nor self-explanatory to human readers. Therefore, as more fine-grained definitions of emotion are proposed, and the

[1] Portions of this chapter are published: N. M. Foteinopoulou and Ioannis Patras. "EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition", in Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkiye, 2024, pp. 1-10

categorical models expand, there is also a need for automated systems to adjust to new definitions, unseen emotions and mental states or compound emotions. Zero-shot methodologies in Facial Emotion Recognition (FER) specifically tackle these challenges, providing effective solutions to unseen emotions and mental states.

Zero-shot Learning (ZSL) is a machine learning paradigm where a model can recognise and classify objects or concepts it has never been trained on by leveraging auxiliary information or attributes associated with those unseen classes [123]. In the context of emotion recognition, ZSL has traditionally been achieved by some label semantic embedding to produce emotion prototypes, typically word2vec [134, 9]. The use of hard labels, however, is failing to include semantically rich information in the prototypes as well as ignoring the subtle differences in expression between subjects. Recent developments in Vision-Language Models (VLM) [95, 53, 2], using image-caption pairs instead of hard-labels have demonstrated superior generalisation and zero-shot abilities. Furthermore, CLIP [95] has been used as the basis for several methods in action recognition [133, 118] or video retrieval and captioning [111, 79, 129, 78]. However, the use of VLMs in dynamic FER remains relatively unexplored. Recent preliminary works in FER have used video-language contrastive training [67, 71], primarily relying on class-level prompt learning. Nonetheless, this approach is akin to class supervision and is not designed to generalise to unseen behaviours.

In this work, we propose a novel approach to zero-shot FER from video inputs by jointly learning video and text embeddings, utilising the contrastive learning framework with natural language supervision. The network architecture is simple and trains a video and text encoder concurrently, as shown in Fig. 5.1. The text and image encoders are initialised by leveraging the knowledge of large-scale pre-trained CLIP [95], as typical dynamic FER datasets do not have enough samples to train from scratch, and we train an additional temporal module from scratch, similar to previous works in action recognition [72, 118]. Contrary to previous VLM or ZSL works in emotion recognition [9, 134, 67, 71] we use sample-level descriptions during training i.e. captions of the subject's facial expression and context, available in the MAFW [73] dataset. These act as soft labels and aim to achieve more semantically rich latent representations of the video samples. Then, during inference, we use class-level descriptions for each emotion. Specifically, we generate descriptions of each emotion in relation to the typical facial expressions associated with it. In the case of compound emotions, we propose manipulating the latent representation of the categories' descriptions in the embedding space rather than creating additional prompts. Specifically, as compound emotions are combinations of basic emotions, we propose averaging the embeddings of the components and adding them to the set of embeddings for each additional compound emotion. We show that the proposed methodology trained on sample-level descriptions shows generalisation capabilities invariant to domain shift as we perform zero-shot evaluation on multiple datasets. We also show that compared to the CLIP and FaRL baselines, the temporal information and domain-specific knowledge of the sample-level descriptions improve the zero-shot performance of FER. Finally, to show the generalisation ability of the representations we obtain from the video encoder, we adapt them to the domain of mental health. Using a simple MLP, trained in a fully supervised manner, we achieve results comparable to or outperforming previous state-of-the-art on estimating non-verbal symptoms of schizophrenia. We see a significant improvement compared to previous works, particularly on symptoms associated with affect and expression of emotions as well as total negative score, which is similar to human experts.

Our main contributions can be summarised as follows:

- This chapter introduces a novel zero-shot Facial Emotion Recognition (FER) paradigm from video input, employing sample-level descriptions and a dynamic model. This straightforward approach, which leverages CLIP [95], outperforms class-level descriptions and significantly improves zero-shot classification performance, particularly for under-represented emotions.

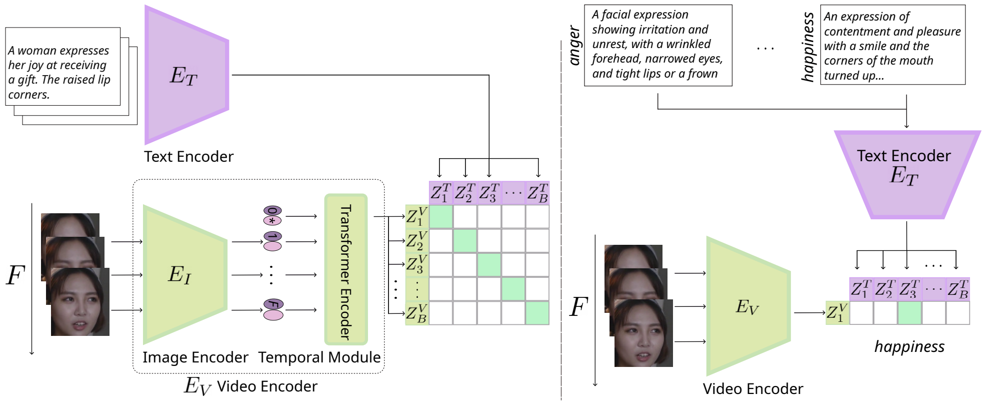
- We propose a novel method for representing compound emotions using average latent representations of basic emotions instead of concatenating or generating new prompts. This approach is more intuitive and efficient than prompt engineering and shows significant improvements across all metrics compared to prompt concatenation.

- Our proposed method, EmoCLIP, trained on the MAFW [73] dataset and evaluated on popular video FER datasets (AFEW [33], DFEW [54], FERV39K [120], and MAFW), achieves state-of-the-art performance on ZSL. Additionally, we evaluate the embeddings of the video encoder of EmoCLIP on the downstream task of schizophrenia symptom estimation using the NESS dataset [92]. We achieve results comparable to or better than previous state-of-the-art methods and comparable to human experts with a simple 2-layer MLP.

The remainder of this Chapter is organised as follows. Section 5.2 introduces our methodology, Section 5.3 reports the results on the tasks of zero-shot FER for simple and compound emotions, and the downstream task of schizophrenia symptom estimation, and Section 5.5 concludes the Chapter.

5.2 Methodology

An overview of the proposed method can be seen in Fig. 5.1. In a nutshell, we follow the CLIP [95] contrastive training paradigm to optimise a video and a text encoder jointly. The video and text encoders of the network are jointly trained using a contrastive loss over the cosine similarities of the video-text pairings in the mini-batch. More specifically, the video encoder (E_V) is composed of the CLIP image encoder (E_I) and a Transformer Encoder to learn the temporal relationships of the frame spatial representations. The text encoder (E_T) used in our approach is the CLIP text encoder. The weights of the image and text encoders in our model are initialised using the large pre-trained weights of CLIP [95] and finetuned on the target domain, as FER


A woman expresses her joy at receiving a gift. The raised lip corners.
E_T
Text Encoder
F
E_I
0
*
1
F
Transformer Encoder
Image Encoder
Temporal Module
E_V Video Encoder
Z_1^T Z_2^T Z_3^T $\cdots$ Z_B^T
Z_1^V
Z_2^V
Z_3^V
Z_B^V
anger
A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed eyes, and tight lips or a frown
...
happiness
An expression of contentment and pleasure with a smile and the corners of the mouth turned up...
Text Encoder
E_T
F
E_V
Video Encoder
Z_1^V
Z_1^T Z_2^T Z_3^T $\cdots$ Z_B^T
happiness

(a) Contrastive Training

(b) Inference

Figure 5.1: Overview of our method, EmoCLIP. During training (a), we use joint training to optimise the cosine similarity of video-text embedding pairs in the mini-batch. Sample-specific descriptions of the subject's facial expressions are used to train the model. During inference (b), we perform zero-shot classification using class-level descriptions for each of the emotion categories.

datasets are not large enough to train a VLM from scratch with adequate generalisation. Contrary to the previous video, VLM works in both action recognition [118, 72] and FER [67, 71], we propose using sample level descriptions for better representation learning, rather than embeddings of class prototypes. This leads to more semantically rich representations, which in turn allows for better generalisation.

We describe how we train the proposed method for dynamic FER in Section 5.2.1, and how we use it at inference time for simple and compound emotions in Section 5.2.3

5.2.1 Architecture and Training

The CLIP framework, proposed by Radford *et al.* [95], operates as a contrastive multi-modal system, encoding image and text features into a shared space and maximising the cosine similarity of matching image-text pairs (positives) while minimising the similarity of all other pairs (negatives) by optimising a cross-entropy loss over the similarity pairs. We adopt the framework in the dynamic paradigm by introducing a transformer encoder over the spatial features to learn from the temporal dimension.

More specifically, we utilise two separate encoders for processing video and text inputs. Given a video-text pair $x = \{x^V, x^T\}$, we obtain the video-text embeddings using the respective encoders so that $\mathbf{z}^V = E_V(x^V)$ and $\mathbf{z}^T = E_T(x^T)$, where $\mathbf{z}^V, \mathbf{z}^T \in \mathbb{R}^D$. The video and text embeddings, $\mathbf{z}^V$ and $\mathbf{z}^T$, respectively, obtained for each image-text pair in the mini-batch B, are utilised to generate a $B \times B$ matrix of cosine similarities. The diagonal elements of the matrix correspond to the B positive pairings, while the remaining elements represent $B^2 - B$ negative pairings. A cross-entropy loss is employed to maximise the similarity between the positives on the diagonal and minimise the similarity of the negatives.

The video encoder architecture is relatively simple and similar to architectures proposed for video captioning [79, 111]. Specifically, (E_V) is composed of the CLIP image encoder (E_I) that extracts frame-level features, which are then fed to a two-layer transformer encoder that acts as a temporal encoder. The state of the learnable classification token at the output of the transformer is used as the video embedding $\mathbf{z}^V$.

While the encoders E_V, E_T could be trained from scratch given a sufficiently large dataset of video-caption pairs, in the domain of FER, the only available dataset with such annotations is the MAFW [73] dataset, which is relatively small. To this end, we leverage the pre-trained CLIP [95] image and text encoders to initialise the weights of E_I and E_T in our architecture and fine-tune on the FER domain.

5.2.2 Inference

During inference, the cosine similarity of text and image embedding in the joined latent space is used as the basis for the classification. The prediction probability is then defined as:

$$P(y = i|x) = \frac{e^{\langle \mathbf{z}^V, \mathbf{z}_i^T \rangle / \tau}}{\sum_{j=1}^{N} e^{\langle \mathbf{z}^V, \mathbf{z}_j^T \rangle / \tau}}, \tag{5.1}$$

where τ is a learnable temperature parameter in CLIP and $\langle \cdot, \cdot \rangle$ is the cosine similarity.

5.2.3 Class Descriptions

In the case of basic emotions, we provide class descriptions in the form of natural language obtained from LLMs, rather than using a prompt in the form of *'an expression of {emotion}'*, to match the information-rich sample level descriptions. We note that these are different descriptions than the ones used during training, as in the latter, we use sample-level video-text pairs, as can be seen in Fig. 5.1. As such, our method is performing zero-shot classification during inference. Specifically, to obtain the class-level descriptions, we prompt ChatGPT with the input:

Q: What are the facial expressions associated with {emotion}?

Class Name	Description
anger	A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed eyes, and tight lips or a frown
disgust	An expression of repulsion and displeasure, with a raised upper lip, a scrunched nose, and a downturned mouth
fear	An expression of tension and withdrawal, with wide-open eyes, raised eyebrows, and a slightly open mouth. The face may appear physically tense or frozen in fear
happiness	An expression of contentment and pleasure, with a smile and the corners of the mouth turned up, often accompanied by crinkling around the eyes. The face may appear relaxed and at ease
neutral	An expression of calm and neutrality, with a neutral mouth and no particular indication of emotion. The eyebrows are usually not raised or furrowed
sadness	An expression of sadness and sorrow, with a downturned mouth or frown, and sometimes tears or a tightness around the eyes. The face may appear physically withdrawn or resigned
surprise	An expression of shock and astonishment, with wide-open eyes and raised eyebrows, sometimes accompanied by a gasp or an open mouth
contempt	An expression of disdain and superiority, with a slight smirk or sneer, often accompanied by a raised eyebrow or a lopsided smile
anxiety	An expression of worry and apprehension, with furrowed eyebrows and a tight mouth. The eyes may appear wide and darting, and the face may appear physically tense or worried
helplessness	An expression of defeat and resignation, with the eyes looking down and the mouth turned down. The eyebrows may be furrowed, and the face may appear resigned or resigned
disappointment	An expression of frustration and disillusionment, with a slight frown or drooping of the mouth. The eyebrows may be lowered or furrowed, and the face may appear physically drawn or tired

Table 5.1: Class descriptions for each emotion used during inference

We then curate the generated responses to exclude irrelevant information, such as body pose or emotional cues (such as "a sad expression" for helplessness). For example, *"A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed* *eyes, and tight lips or a frown"* is the generated description for *"anger"*. Examples of prompts can be seen in Fig. 5.1 & 5.2; the full set of class descriptions is given in Table 5.1.

The MAFW [73] dataset does not include sample-level descriptions for the neutral category; as such, we generate descriptions for neutral samples by randomly selecting and concatenating two prompts generated from ChatGPT for the neutral category. The full list of prompts generated for the category can be seen in Table. 5.2.

Neutral Category Description Components
A lack of emotional expression, as if the person's face is in a resting state.
The facial muscles are generally relaxed, creating a smooth and even appearance.
The mouth is typically closed or slightly open, with the lips not turned up or down.
The eyebrows are in a neutral position, not furrowed or raised, and the eyes are generally looking straight ahead or slightly down.
While the face may not show any specific emotions, the expression can still convey a sense of attentiveness or alertness.

Table 5.2: Descriptions used for neutral category samples during training

This strategy is proposed over the prompt templates used in CLIP [95], as the prompt in the form of *'an expression of {emotion}'* would imply a universal definition for each emotion, that the text encoder has learnt along with the underlying behaviours and facial expressions associated with each emotion. Additionally, while the CLIP prompts have shown impressive results on clearly defined objects in images, emotions are significantly more vague with large intra-class variation and open to interpretation, both in terms of expression and understanding. Finally, the use of LLM, in this case, ChatGPT, over hand-crafted prompts is proposed to avoid introducing the authors' bias in the prompts.

Compound emotions are a complex combination of basic emotions, such as "happily surprised" [30, 73], that are used to identify a wider range of human facial expressions. As, by definition, compound emotions are combinations of basic emotions; we propose a new approach to constructing the latent $\mathbf{z}^T$ representation. Instead of treating them as independent emotional states and creating additional prompts, we use the pre-normalised latent representations of the components to compose the new compound

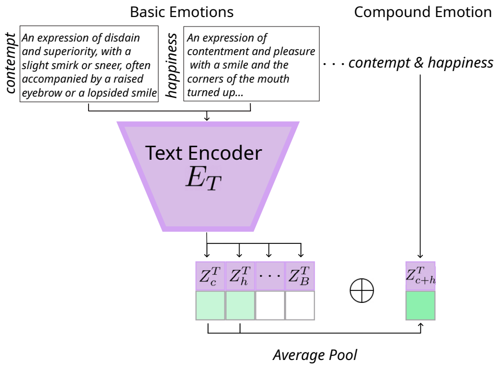

Basic Emotions
Compound Emotion
contempt
An expression of disdain and superiority, with a slight smirk or sneer, often accompanied by a raised eyebrow or a lopsided smile
happiness
An expression of contentment and pleasure with a smile and the corners of the mouth turned up...
. . . contempt & happiness
Text Encoder
E_T
Z_c^T
Z_h^T
· · ·
Z_B^T
⊕
Z_{c+h}^T
Average Pool

Figure 5.2: EmoCLIP manipulates the latent space of basic emotions to create representations for compound emotions. We take the average latent representation of the components and concatenate it to the set of representations for each new compound emotion.

emotion as shown in Fig. 5.2.

Formally, for any new compound emotion, we calculate its latent representation $\mathbf{z}_n^T$ as the average of the latent representations of its C component emotions so that:

$$\mathbf{z}_n^T = \frac{1}{C}\sum_{c=0}^{C}\mathbf{z}_c^T \tag{5.2}$$

The resulting vector representations are then concatenated to the set of class representations. For the final classification, the cosine similarity between the class (basic and compound) and video embeddings is calculated as described in Section 5.2.2.

5.3 Experimental Results

In this section, we present the experimental set-up (Section 5.3.1), the ablation study (Section 5.3.2), and the experimental evaluation of the proposed framework in the Zero-Shot paradigm (Section 5.3.3) as well as the Downstream task of schizophrenia symptom estimation (Section 5.3.4).

5.3.1 Experimental Setup

Datasets: We train on the **MAFW** [73] dataset, as to our knowledge, it is the only FER dataset that includes sample-level descriptions of the facial expression without explicitly mentioning emotion. The dataset contains $10k$ audio-video clips with categorical annotations for 11 emotions, accompanied by short descriptions of facial expressions in two languages, English and Mandarin Chinese. The dataset is divided into five folds for evaluation. In this work, we use the English descriptions.

We evaluate our method on three additional FER datasets on the seven basic emotions. The **AFEW** [33] dataset contains 1,809 clips from movies, divided into three subsets. We train the method from scratch and report results on the validation set. **DFEW** [54] is composed of 16,000 videos from movies, split into five folds for cross-validation. **FERV39K** [120] is a large dataset with around 39,000 clips annotated by 30 annotators, and we report results on the test set.

Finally, we evaluate the video encoder of EmoCLIP as trained on the MAFW dataset, on the downstream task of schizophrenia symptom estimation using a subset of the **NESS**[92] dataset, as described in [15, 41]. The subset includes 113 in-the-wild baseline clinical interviews from 69 patients and two symptom scales, PANSS[55] and CAINS[40]. To ensure a fair comparison with previous works, we used leave-one-patient-out cross-validation. The values of “Total Negative” and “EXP - Total” in PANSS and CAINS scales, respectively, were scaled during training to match the range of individual symptoms. **Metrics:** For the evaluation, we use Unweighted Average Recall (UAR), also known as balanced accuracy, and Weighted Average Recall (WAR), which is equivalent to accuracy and are given by eq. 5.3 and 5.4 respectively:

$$UAR = \frac{1}{|C|}\sum_{c=0}^{|C|}\frac{TP_c}{TP_c + FN_c} \tag{5.3}$$

$$WAR = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

where C is the set of classes and TP, TN, FP, FN refer to *True Positive, True Negative, False Positive* and *False Negative* respectively. For the compound expressions, we also use the F1 score defined in eq. 5.5(also known as the harmonic mean between precision and recall) and the Area Under the ROC Curve (AUC):

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{5.5}$$

For all metrics, a higher value indicates a better model performance. For ease of reading, we present all UAR/WAR values multiplied by 100.

Training Details: The CLIP image encoders used in all experiments have a ViT-B/32 architecture. During training, we apply augmentation to the spatial dimensions of the video inputs for all datasets. Specifically, we apply a random horizontal flip to all frames in a sequence with a probability of 50%. We also apply a random rotation with a range of 6°and random centre crop. In the temporal dimension, we empirically trim all videos to $T = 32$ number of frames and downsample by a factor of 4 (which results in the entire video being included for the majority of samples). As the NESS dataset contains significantly longer sequences, during inference, we average the prediction of all the clips as described in [41].

We use a Stochastic Gradient Descent Optimiser (SGD) with a learning rate of 10^{-3} for all experiments. We finetune the pre-trained parameters of the CLIP backbone using a different learning rate of 10^{-6} for the image and text encoders.

5.3.2 Ablation Study

In order to examine the effect of different components of EmoCLIP and evaluate the prompting strategy used in this work, we perform several experiments using baseline CLIP and our proposed method, EmoCLIP. More specifically, as discussed in the Methodology Section 5.2, during inference, we use class descriptions generated with the help of ChatGPT; this is different to the prompt ensemble of Radford *et al.* [95] who use several prompt templates in the form of *'an expression of {emotion}'* and average their

Mode	Architecture	Contrastive Pre-training	UAR	WAR
Supervised	C3D [73]	-	31.17	42.25
	Resnet18_LSTM [73]	-	28.08	39.38
	VIT_LSTM [73]	-	32.67	45.56
	C3D_LSTM [73]	-	29.75	43.76
	T-ESFL [73]	-	33.28	**48.18**
	EmoCLIP (LP)	-	30.26	44.23
	EmoCLIP (Frozen backbone)	MAFW [class descriptions]	**34.24**	41.46
Zero-shot	CLIP [95]	Laion-400m	20.40	21.16.
	FaRL - ViT-B/16 [137]	Laion Face-20M	14.07	7.70
	EmoCLIP	MAFW [sample descriptions]	**25.86**	**33.49**

Table 5.3: Performance of the proposed method on the MAFW [73] dataset on 11-class single expression classification against other SOTA architectures in a supervised and zero-shot setting.

latent representation vectors during inference. Furthermore, as CLIP [95] is a static model, we conduct two zero-shot evaluation strategies, first on the middle frame and then by averaging the latent representations of all frames, i.e. performing frame ensemble. To further show the necessity for a temporal layer, we finetune a CLIP [95] architecture on the MAFW [73] dataset using frame ensembling, which has a negative effect on CLIP's performance on FER. We theorise that frame ensembling in FER tasks negatively affects training as averaging out frame representations in most cases will consider noisy and keyframes equally, thus muddling the video latent features [68], which confuses the network.

The proposed architecture, which incorporates a temporal layer, has significant performance improvements compared to the baseline CLIP architectures. Finally, we compare the performance of EmoCLIP with a frozen backbone to that of our proposed method. The results of the ablation study can be seen in Tab. 5.5. As the prompt templates used in CLIP [95] are similar to the format “A photo of class name”, they are more suitable for static images rather than video. This is corroborated by the performance of the CLIP [95] baseline using prompt ensemble vs our class descriptions. Additionally, as emotional expression is not static, we also observe an increase in the baseline performance using frame ensembling compared to evaluating on the middle frame. Such an outcome is intuitive as dynamic FER has a wider temporal context that needs to be considered rather than a single frame. Furthermore, the temporal relationships between frames hold important information in FER tasks. The proposed architecture, including a temporal module, offers a large increase in performance for both Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). By finetuning the backbone to the FER domain, we see a further increase in performance by approximately 2% for each metric.

5.3.3 Zero-Shot Evaluation

To evaluate the effectiveness of the proposed method, we compare it with pre-trained CLIP [95] and FaRL [137] models in a Zero-shot setting. As both of these methodologies are trained on static images, we take the average of the latent representations of all frames in a video to compute the video embedding and calculate the cosine similarity with the text embeddings. We show the performance of our method against the CLIP and FaRL baselines, with a frozen CLIP backbone and the finetuned image-text encoders on the 11 class classification of MAFW [73] in Table 5.3. We note that even though FaRL [137] is trained on a subset of the Laion dataset [101] filtered to include samples of faces, the model trained on FaRL performs significantly worse than both the CLIP baseline and our proposed method. We also note that the FaRL pre-trained weights are only available for the ViT-B/16 architecture, which may explain the difference in performance. Furthermore, while FaRL is trained with image-text pairs of faces, these are not necessarily related to facial expression, so we hypothesise that the FaRL embedding space is significantly more niche compared to CLIP but not in a direction that is beneficial for zero-shot FER.

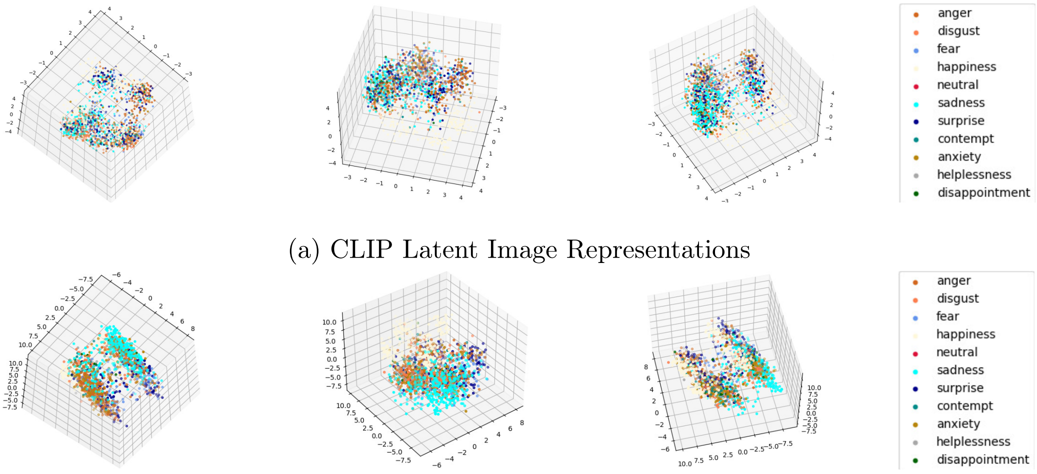
To further investigate the improvement of our method vs baseline CLIP, we use PCA to reduce the high-dimensional latent image vectors to three dimensions (which explain approximately 70% of the variance) and plot them in a 3D scatter plot, as shown in Fig. 5.3. We see that by fine-tuning to the FER domain, the categorical emotions form more distinct clusters than in CLIP, which is also reflected in the classification performance of our method.

For reference, we also train our architecture with class-level descriptions, and using an MLP head with two fully connected layers (Linear probe shown as LP on the table) over the EmoCLIP video encoder, we show the performance against several supervised architectures as reported in [73]. We see that the architecture trained using the class descriptions outperforms previous methods in terms of UAR and is comparable with others in terms of WAR. These results indicate that the contrastive vision-language approach leads to more semantically rich and discriminating latent representations, even in a supervised setting. The difference in the two metrics is somewhat expected, as FER datasets are typically imbalanced.

Mode	Architecture	Repr. Avg	UAR	WAR	F1	AUC
Supervised	C3D [73]	-	**9.51**	28.12	6.73	74.54
	Resnet18_LSTM [73]	-	6.93	26.6	5.56	68.86
	VIT_LSTM [73]	-	8.72	32.24	**7.59**	75.33
	C3D_LSTM [73]	-	7.34	28.19	5.67	65.65
	T-ESFL [73]	-	9.15	**34.35**	7.18	**75.63**
Zero-shot	Random	-	2.38	7.72	0.34	50.00
	CLIP [95]	✗	4.72	5.25	2.44	51.89
	CLIP [95]	✓	4.14	5.35	2.46	**53.07**
	FaRL [137]	✗	3.03	4.66	2.16	51.01
	FaRL [137]	✓	4.00	5.75	2.56	51.10
	EmoCLIP	✗	5.24	15.34	3.80	51.30
	EmoCLIP	✓	**6.58**	**18.53**	**4.78**	**52.59**

Table 5.4: Zero-shot classification on the 43 compound expressions of the MAFW [73] dataset. Supervised methods are included as a reference.

Furthermore, we present experimental results on the classification of 43 compound emotions in the MAFW dataset in Table 5.4. We evaluate the performance of our proposed method, EmoCLIP, against a baseline approach of using concatenated prompts, as well as CLIP and FaRL baselines. Specifically, we concatenate the class descriptions for each compound emotion and use this as class prompt input. We demonstrate that EmoCLIP outperforms the baseline approach for all metrics. Moreover, we note that in the 43 emotions classification, both CLIP and FaRL perform significantly worse than EmoCLIP and have performance comparable to random (where only the majority class



(a) CLIP Latent Image Representations

(b) EmoCLIP Latent Image Representations

Figure 5.3: Latent representation of the MAFW dataset using CLIP (a) and EmoCLIP (b) for each emotion category.

is predicted). We theorise this is due to the lack of temporal understanding of the static models. Furthermore, without fine-tuning the target domain, the class descriptions are not discriminate enough for the zero-shot classification of emotions, as previously discussed. However, the representation average method on both baselines improves the performance of the static models on the compound emotions task. We also report the results of several supervised methods for reference, which perform significantly better than zero-shot approaches as expected.

Finally, we evaluate the performance of our proposed method using sample-level descriptions from MAFW [73] on four widely used video FER datasets and compare it with the CLIP baseline as shown in Table 5.6. Additionally, in line with previous works in zero-shot emotion classification [9, 125, 134, 93] we train our architecture using class-level descriptions and evaluate using leave-one-class-out (loco) cross-validation. We note that we cannot directly compare with these architectures, as they involve either different modalities (eg. audio, pose) [9, 125, 93, 126] or a different task [134], we adopt however, their experimental set-up using our architecture to show how natural language supervision and semantically rich class descriptions can help improve zero-shot FER performance.

We observe that the EmoCLIP trained on MAFW [73] sample-level descriptions show impressive generalisation ability on all datasets that we evaluate. Specifically, for AFEW [33], MAFW [73] and DFEW [54], we see that the EmoCLIP model is outperforming both the loco experiment and the CLIP [95] baseline. Furthermore, the generalisation of the method is resistant to domain shift from unseen datasets, as we observe from the significant performance increase between the CLIP [95] baseline and EmoCLIP. We note that for FERV39K [120], the loco experiment has a higher performance than the sample-wise training; however, it is very important to stress that the FERV39K [120] is significantly larger than the base dataset (over 3x more samples) therefore methods trained on it would have an advantage, particularly as in the loco experiment there is no domain shift.

For reference and to provide context, we include the performance of the architecture in a supervised setting on all four datasets. We want to point out that the architecture trained using class descriptions is significantly outperforming the linear probe architecture, particularly in terms of UAR, showing again that natural language supervision can provide significant advantages even in a fully supervised setting.

Architecture	**Temporal Layer**	**Prompt Ensemble**	**Frame Ensemble**	**Pre-training**	**UAR**	**WAR**
CLIP [95]			✓	Laion-400m	20.40	21.16
CLIP [95]		✓		Laion-400m	19.77	18.64
CLIP [95]		✓	✓	Laion-400m	19.46	17.61
CLIP [95]				Laion-400m	16.97	21.69
EmoCLIP (Frozen backbone)	✓			MAFW	23.6	31.36
EmoCLIP	✓			MAFW	**25.86**	**33.49**

Table 5.5: Performance of the proposed method, EmoCLIP, on the MAFW [73] dataset on 11-class single expression classification against the baseline, with different frame aggregation and prompting strategies.

			DFEW		AFEW		FERV39K		MAFW	
			(7 classes)		(7 classes)		(7 classes)		(11 classes)	
	Architecture	Training labels	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
Supervised	EmoCLIP	[class description]	**58.04**	**62.12**	**44.32**	**46.19**	**31.41**	36.18	**34.24**	41.46
	EmoCLIP (LP)	[class]	50.29	62.09	33.74	38.85	30.58	**43.54**	30.26	**44.21**
Zero-shot	CLIP [95]	[image caption]	19.86	10.60	23.05	11.80	20.99	17.10	20.04	21
	EmoCLIP (leave-one-class-out)	[class description]	22.85	24.96	35.11	27.57	**39.35**	**41.60**	24.12	24.74
	EmoCLIP	[video caption]	**36.76**	**46.27**	**36.13**	**39.90**	26.73	35.30	**25.86**	**33.49**

Table 5.6: Evaluation of EmoCLIP using sample descriptions vs class-level description as natural language supervision, on four video FER datasets.

5.3.4 Downstream Task

We evaluate the representations obtained by the proposed method on the downstream task of schizophrenia symptom estimation on two scales, namely the PANSS [55] and CAINS [40] scales. As symptoms in both scales have ordinal labels, we address the problem of symptom estimation as a multi-label regression. We evaluate the proposed method in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Pearson's Correlation Coefficient (PCC) for a fair comparison with previous works in the literature. We train the linear probe architecture, with the video encoder obtained through contrastive pre-training on the MAFW [73] frozen, and updating only the MLP weights. The results against previous state-of-the-art (SoTA) can be seen on Tables 5.7 & 5.8 for the PANSS [55] and CAINS [40] scales respectively. The proposed method, pre-trained contrastively on sample level descriptions, outperforms or is comparable to previous methods on all symptoms, particularly for "N1: Blunted Affect" and "Facial Expression" on the PANSS and CAINS scales, which are by definition most related to FER.

As multiple healthcare professionals annotate the NESS [92] dataset, we can compare the PCC achieved by the method to the annotators' PCC, which has a mean value of 0.85 across all symptoms [15, 92]. The proposed method on the downstream task achieves performance comparable to that of human experts on the total scores of both scales, as we can see in Tables 5.7 & 5.8 particularly for the total scores.

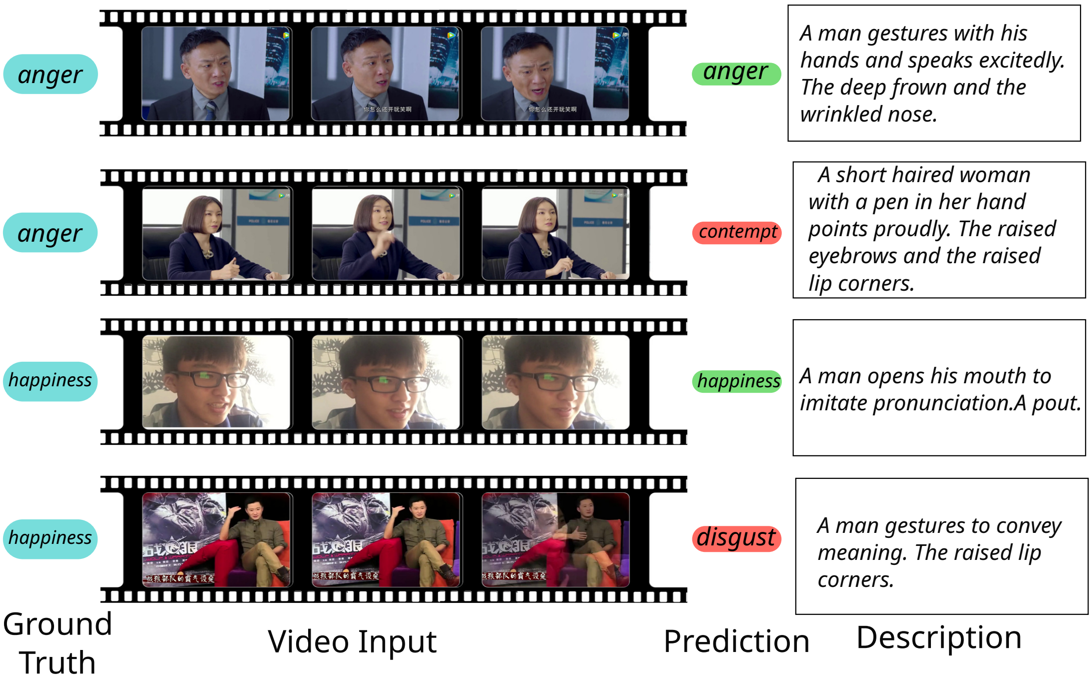
	N3: Poor Rapport			N6: Lack of Spont.			N1: Blunted Affect			Total Negative Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [113]	0.98	1.31	.20	1.37	1.69	.13	0.90	1.28	.37	-	-	-
Tron *et al.* [115]	1.01	1.26	.15	1.32	1.62	.09	0.99	1.36	.11	-	-	-
SchiNet [15]	**0.85**	1.20	.27	1.25	1.51	.25	0.84	1.18	.42	3.30	4.17	.29
Relational [41]	0.87	**1.16**	**.78**	**0.74**	**0.95**	**.47**	0.64	0.87	.56	2.80	3.78	.71
EmoCLIP (LP)	0.89	1.28	.72	0.91	1.15	.27	**0.56**	**0.79**	**.63**	**2.31**	**3.01**	**.85**

Table 5.7: Performance on the downstream task against other SoTA (PANSS-NEG).

	Facial Exp.			Vocal Exp.			Expr. Gestures			Quant. of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [113]	0.80	1.03	.37	0.87	1.23	.23	0.85	1.19	.36	1.09	1.43	.27	-	-	-
Tron *et al.* [115]	0.75	1.07	.36	0.86	1.22	.26	0.91	1.22	.38	1.02	1.36	.25	-	-	-
SchiNet [15]	0.66	0.93	.46	0.77	1.10	.27	0.90	1.15	.36	0.98	1.30	.30	2.67	3.34	.45
Relational [41]	0.56	0.72	.75	0.65	0.89	.71	0.71	0.89	**.76**	**0.60**	**0.82**	**.54**	1.88	2.60	.77
EmoCLIP (LP)	**0.49**	**0.65**	**.77**	**0.59**	**0.83**	**.74**	**0.66**	**0.85**	.74	0.64	0.89	.50	**1.32**	**2.52**	**.83**

Table 5.8: Performance on the downstream task against other SoTA (CAINS-EXP).

5.4 Qualitative Analysis



A man gestures with his hands and speaks excitedly. The deep frown and the wrinkled nose.

A short haired woman with a pen in her hand points proudly. The raised eyebrows and the raised lip corners.

A man opens his mouth to imitate pronunciation.A pout.

A man gestures to convey meaning. The raised lip corners.

Figure 5.4: Example frames of correctly and incorrectly classified samples from the MAFW [73] dataset, for anger (top) and happiness (bottom). The sample-level descriptions are included for reference but are not used during inference.

Figure 5.4 showcases a selection of correctly and incorrectly classified instances from the MAFW dataset. As the interpretation of emotion for humans is dependent on not only the facial expression but also the context, we see that some samples are harder than others to classify. It appears that examples with higher emotional intensity and more animated subjects appear to be easier to classify. Conversely, subtle expressions of emotion are prone to misclassification. For instance, in the anger examples depicted in Figure 5.4, the man displays multiple anger-associated expressions, such as a furrowed brow, whereas the woman exhibits a calmer demeanour. Similarly, in the happiness examples, the appropriately classified sample features a smiling man, while the incorrectly classified example shows a man with a subtle facial expression while speaking.

5.5 Conclusion

In this work, we presented a novel contrastive pre-training paradigm for video FER, trained on video-text pairs with sample-level descriptions without any class-level information. While contrastive learning and natural language supervision have been used in other domains, zero-shot emotion recognition remains surprisingly unexplored, with works focusing on creating class prototypes with simpler word encoding methods [9, 93, 125, 126]. Emotional prototypes, however, disregard the intra-class variation that is inherently present in FER tasks. To overcome the limitations of training on coarse emotional categories, EmoCLIP is trained on sample-level descriptions. We evaluate our method on four popular FER video datasets [33, 54, 120, 73] and test using zero-shot evaluation on the basic emotions as well as compound emotions. Our method outperforms the CLIP baseline by a large margin and shows impressive generalisation ability on unseen datasets and emotions. To our knowledge, this is the first work to train with sample-level descriptions for FER and to propose zero-shot evaluation using semantically rich class descriptions in the domain. We also evaluated the EmoCLIP video encoder features on schizophrenia symptom estimation, outperforming previous state-of-the-art methods and achieving performance comparable to human experts in terms of PCC.

Chapter 6

Conclusions

Contents

6.1 Discussion and Conclusions

In this thesis, we worked towards a fully automated system for the assessment of mental illness behavioural symptoms, specifically the negative symptoms of schizophrenia in conditions that resemble real-life clinical interviews. As this is a very large research area with few available datasets, we identify three sub-problems. First, we examine apparent continuous affect as analysing facial non-verbal behaviour forms the underlying basis of most affective tasks. Second, we extend such works to symptom severity estimation and show that the two problems can be examined in parallel. Third, we develop a method for zero-shot facial expression recognition to identify new and unseen emotions and mental states using natural language descriptions.

We started by developing architectures for continuous affect estimation, i.e. Arousal-Valence estimation, as there are several similarities between continuous affect and symptom estimation. This relationship is inherent as both tasks heavily rely on the non-verbal behaviour of the subject. Affect is directly measured for symptoms such as Blunted Affect on the PANSS [55] or Facial Expression on CAINS [40]. However, as there are no universal definitions for apparent emotion, we first developed an uncertainty-aware method for continuous affect estimation that addresses the inherent ambiguity in affective computing tasks. The proposed method addresses each label as a univariate Gaussian distribution where the mean is the ground truth (annotator's mean) and an unknown variance predicted by the model. The Kullback–Leibler (KL)-based loss minimises the distance between the Gaussian label and the network's prediction, a dirac delta. Such an approach is novel in continuous affect and improves the performance of several backbone architectures. The contributions of this work are two-fold: first, we show that the proposed methodology improves performance for static image and video input and second, we show a weak relationship between the predicted variance and noisy samples.

The KL-based loss acts as a regularisation method for noisy samples by penalising less uncertain samples. At the same time, for less noisy samples, the loss prioritises minimising the distance between the estimated mean and the annotators' mean. Using several feed-forward neural network architectures with two MLP heads (one for the main task and an additional one estimating label uncertainty), we show that the proposed loss improves the network prediction, particularly in terms of Pearson's Correlation Coefficient (PCC).

In addition, when individual annotations are available, we calculate the PCC between the estimated variance and annotators' variance and show a weak positive correlation for both Arousal and Valence. When individual annotations are not available, we compare the categorical annotations against their theoretical equivalent on the continuous Arousal Valence circumplex and identify a sub-group of noisy samples where the annotations are in disagreement. For the noisy sub-group, we show that the estimated variance is higher than the variance of the clean sub-group. The method is evaluated on two popular affect datasets, namely the AMIGOS [83] and AffectNet [84], which are composed of video and static images, respectively.

In the second main chapter of this thesis, we continue evaluating our methodologies on continuous affect and extend the architecture to symptom severity estimation. The main idea of the second methodology presented in this thesis is to use context to improve the network predictions. Context, in this case, has a dual meaning as it refers to the temporal neighbourhood of a clip sampled from a wider video and the other samples during training.

Affect and mental health symptoms often refer to a wider behaviour that cannot easily be localised along the temporal dimension; as a result, particularly for mental health, video samples tend to be very long sequences of frames, which is problematic in terms of architectural and hardware limitations. Typically, to address this issue, researchers used statistical representations [113, 15] which disregard temporal relationships of features or made predictions on shorter clips taken from wider videos that can lead to the "flickering" problem [135] i.e. a model producing inaccurate and unrelated predictions for individual frames or clips, disregarding the temporal context and leading to inconsistent or erratic output sequences. Our proposed architecture uses a two-staged transformer architecture to improve individual clip predictions while moving away from statistical representations of features. The architecture comprises two networks with shared weights; one extracts clip-level features, while the second extracts features from neighbouring clips that act as context features. The clip and context features are fed into an attention block before passing the attended clip features to the regression head. Updating the context branch weights and, consequently, features during training significantly reduces training times and parameters compared to previous work [122]. The experimental results, particularly the ablation studies, show that including features from neighbouring clips greatly contributes to the prediction accuracy for both tasks, i.e. schizophrenia symptom estimation and continuous affect.

In the same chapter, we propose using sample context by introducing the relational loss. More specifically, inspired by contrastive approaches in single-label categorical tasks, we propose aligning the distances of the continuous label vectors to the distances of the clip latent features. The introduction of the relational loss is the main contribution of Chapter 2, the strength of which lies in its simplicity. As we have seen in several categorical tasks, latent feature alignment typically results in better predictions; however, this idea has not been extended to continuous tasks. Experimental results show that by aligning the clip-level feature distances to the label vector distances, the architecture's performance is significantly improved, particularly for smaller datasets.

Finally, in the last methodological chapter of this thesis, we explore the idea of natural language as a method of describing emotional states instead of the classic supervised setting using Ekman's [35] categorical model. Such an approach is intuitive among humans; however, it remains relatively unexplored in affective computing. More specifically, we propose jointly training video and text encoders by optimising the cosine similarity of video-text embedding pairs in the mini-batch. While previous works in FER have used class prototypes during training, we proposed using sample-level descriptions of the subject's facial expressions. During inference, we evaluate using class-level descriptions of each emotion concerning the facial expression instead of descriptions as *"a photo of {emotion}"*. While the approach is still evaluated on categorical emotions predefined by each dataset, we show experimentally that the network maintains impressive zero-shot capabilities across datasets. For compound emotions that are combinations of basic emotions, we propose linearly combining the latent features of the components rather than creating additional category descriptions. Combining basic emotional descriptions is consistent with Ekman 's [35] universality of the basic emotions but addresses limitations associated with the large intra-class variations in the basic categories. The semantically rich latent space obtained when training with sample-level descriptions helps achieve results comparable to or outperforming previous work on the downstream task of schizophrenia symptom estimation.

One of the main themes of our analysis revolves around the use of labels in human affect, whether by acknowledging label uncertainty (Chapter 3), labels' context (Chapter 4) or natural language to describe emotional states (Chapter 5), we show that significant performance improvements can be achieved by understanding the inherent ambiguity and intra-class variation in affective computing. Another central theme is the transferability of methods within affective computing tasks. While mental health symptom estimation has unique challenges associated with the problem's nature, we can draw methodological parallels that help improve task-specific performance by looking at the more general problem of understanding human behaviour.

6.2 Strengths and Limitations

In the methodological chapters of this thesis, we address three challenges related to the ground truth labels of affective computing tasks: uncertainty, context and definitions in Chapters 3, 4 and 5 respectively. The proposed approaches differ from traditional supervised settings where a network makes a single categorical or continuous prediction; as we have shown experimentally, performance improvements can be made by understanding the nuances of the ground truth.

In classic supervised learning tasks, each input has a target vector that, in tasks such as object detection, is very objectively defined and visually grounded. However, as established in Chapter 1, human behaviour is more abstract and subjective, even for human experts. Previous research typically takes the mean or majority of annotations during training and inference to address this inherent ambiguity. In Chapter 3, we show that by acknowledging label uncertainty during training in both dynamic and static tasks, we can learn continuous affect more robustly. In addition, we show that there is a relationship between estimated uncertainty and the annotators' disagreement without explicitly training on the latter. Such an approach is novel in continuous affect; however, the estimated uncertainty can only be evaluated if the annotators' disagreement is available. Furthermore, as typically affective datasets have a limited number of samples, the estimated uncertainty could greatly be influenced by the label distribution and available samples.

As annotating large datasets is costly, several pre-training methods such as contrastive learning [27, 26, 59] have been proposed in the literature focusing primarily on categorical tasks. In Chapter 4, we propose a novel approach inspired by supervised contrastive learning to regularise learning in multi-label continuous tasks and learn better latent representations. In addition, we show that incorporating features from the wider temporal context can further improve the method's performance on both tasks, i.e., continuous affect estimation and schizophrenia symptom estimation. However, hardware constraints still are a significant limitation to consider when using context features in very long videos, e.g. in the schizophrenia symptom estimation task where it is still impossible to include all frames when making a prediction, thus leading to some information loss. Furthermore, when very long videos with low-label resolution are analysed in human behavioural tasks, we can expect that not all frames exhibit the behaviour [68]. As such, inherent aleatoric uncertainty is not addressed by the methods in this thesis.

As discussed in Chapter 1, several models are proposed for human affect [35, 99]. However, the human emotional experience is very diverse, and humans use a plethora of descriptions for their emotional states, which are not accurately described by the coarse basic emotions categorical model. Furthermore, the Arousal-Valence circumplex is not easily explained in natural language, which is a significant obstacle in the interpretability of continuous methods. Given the subjectivity of human emotions, there is a motivation for approaches that can identify new and unseen emotional states, i.e. zero-shot approaches in FER. In Chapter 2, we see that very few methods attempt zero-shot FER and that these focus on learning from coarse categories. Such approaches ignore the intra-category variance and sample-level differences in emotional expression; therefore, their latent representations would not include semantically rich information regarding the samples' apparent emotion and thus would be less robust to new emotional states. In Chapter 5, we propose adopting a natural language supervision approach using sample-level descriptions, which is novel in the domain of dynamic FER. We show that the proposed approach is superior to class prototypes and generalises better to unseen datasets and emotions. However, as affective datasets are typically small zero-shot approaches in FER still underperform compared to fully-supervised methods and lack the performance capabilities of VLMs in other tasks.

6.3 Wider Implications and Potential Applications

This thesis has contributed to automated emotional and mental state estimation. As briefly mentioned in Chapter 1, there are several potential applications for affective computing in general, mainly focusing on assisting professionals and improving human-computer interaction. The most straightforward application of the proposed methods is in a clinical setting, assisting healthcare professionals with patient assessment by analysing non-verbal behaviour. Such an application would help develop personalised medicine in mental health, considering individual emotional and mental states for more tailored treatment. From a methodological perspective, we show that by addressing affect and mental health intersectionally, we can research towards a more unified human behavioural model.

However, there are several implications to consider in affective computing tasks, specifically around ethics, privacy and potential misuse of such technology. Algorithmic bias is a major concern, particularly in domains such as affective computing, with limited training data available to researchers. A low data regime can be very problematic in affective computing and even more so in mental health tasks, as it is very easy for models to learn from unrelated characteristics (e.g. skin colour, environment) instead of the underlying behaviour. Bias in diagnostic tools bears significant risks for public health, and efforts need to be made to mitigate bias in both datasets and algorithms. One promising direction is the use of synthetic data [34]; however, research in the domain is still in its infancy, particularly for dynamic datasets. In addition, automated methods may misinterpret subtle signs of mental health conditions, leading to incorrect diagnoses or recommendations. This risk can be avoided with human validation, so it is important to stress that any automated assessment and diagnostic tools should aim to assist rather than replace experts.

An additional advantage of synthetic data is anonymisation and privacy. The low-data regime discussed in the previous paragraph is a direct result of the sensitivity of the tasks. Anonymisation allows for the safe collection of datasets without the risk of leaking private or sensitive information such as medical records, a major concern that inhibits research in mental health tasks.

6.4 Future work

Although the proposed methods for symptom estimation show promising results, ML in mental health is still an emerging field. As a result, further improvement in terms of performance and generalisation abilities is needed to achieve a reliability level acceptable for clinical application. The datasets and methods proposed are limited to schizophrenia patients, completely disregarding other illnesses with similar symptoms or pathology, for example, schizoaffective disorder that shares some but not all of the symptoms present in schizophrenia [121]. Still, there is no indication of how the methods proposed in this thesis would generalise to patients with those conditions, particularly as several symptoms in schizophrenia are correlated to each other [113, 115], which is a form of bias in the annotations provided. There is, therefore, an opportunity for a fine-grained approach that looks into a more general pool of subjects with shared symptoms but different conditions.

In addition, with mental health data being limited due to confidentiality constraints, there is a significant research gap for zero-shot approaches in schizophrenia specifically and mental health symptom estimation in general. In this thesis, we explore zero-shot approaches based on natural language descriptions of emotional states in categorical FER; however, extending the zero-shot method to mental health is currently constrained by language model limitations. More specifically, negative symptoms define behaviour that is absent in patients relative to the general population; therefore, the latent representations of the class descriptions in the symptom estimation task need to account for negated prompts, that is, which is a major limitation in the capabilities of Large Language Models currently [52]. As such, there is a significant opportunity to address the negated prompts in zero-shot affect and mental health symptom estimation.

Appendix **A**

Appendix A

Contents

A.1 Positive and Negative Symptom Scale (PANSS)

The Positive and Negative Syndrome Scale (PANSS) is a scale used for measuring the symptom severity of schizophrenia. It is widely used for the assessment of patients and to measure illness progression.

PANSS rates 30 symptoms in three symptom categories: (a) Positive Symptoms, (b) Negative Symptoms, and (c) General Psychopathology. Each of the 30 items is accompanied by a specific definition and detailed anchoring criteria for all seven rating points given by [55]. These seven points represent increasing levels of psychopathology, as follows:

1. Absent

2. Minimal

3. Mild

4. Moderate
5. Moderate Severe
6. Severe
7. Extreme

The Positive Symptoms refer to behaviour present in patients with schizophrenia but not the general population. The seven symptoms on the positive scale are listed as follows:

1. Delusions
2. Conceptual disorganization
3. Hallucinatory behaviour
4. Excitement
5. Grandiosity
6. Suspiciousness/Persecution
7. Hostility

The Negative Symptoms refer to a lack of function in patients with schizophrenia relative to the general population. The seven symptoms of the negative scale are listed as follows:

1. Blunted affect
2. Emotional withdrawal
3. Poor rapport

4. Passive/apathetic social withdrawal
5. Difficulty in abstract thinking
6. Lack of spontaneity & flow of conversation
7. Stereotyped thinking

Finally, sixteen symptoms are associated with the general psychopathology scale:

1. Somatic concern
2. Anxiety
3. Guilt feelings
4. Tension
5. Mannerisms & posturing
6. Depression
7. Motor retardation
8. Uncooperativeness
9. Unusual thought content
10. Disorientation
11. Poor attention
12. Lack of judgement & insight
13. Disturbance of volition
14. Poor impulse control
15. Preoccupation
16. Active social avoidance

A.2 Clinical Assessment Interview for Negative Symptoms (CAINS)

The Clinical Assessment Interview for Negative Symptoms (CAINS) [40] is a “second generation” scale for the assessment of negative symptoms in schizophrenia; it was developed using an iterative, empirical approach and includes items assessing motivation, pleasure, and emotion expression. CAINS consists of two negative scales rated separately: Motivation & Pleasure and Expression. The Motivation & Pleasure scale has nine symptoms, and the Expression scale has four. Each symptom has a value between zero and four, representing increasing levels:

0. No impairment
1. Mild deficit
2. Moderate deficit
3. Moderately severe deficit
4. Severe deficit

The motivation and pleasure scale measures impairment in motivation for social relationships, school/work activities and recreation and lists the following symptoms:

1. Motivation for Close Family/Spouse/Partner Relationships
2. Motivation for Close Friendships & Romantic Relationships
3. Frequency of Pleasurable Social Activities - Past Week
4. Frequency of Expected Pleasure from Social Activities - Next Week
5. Motivation for Work & School Activities

6. Frequency of Expected Pleasure from Work & School Activities - Next Week

7. Motivation for Recreational Activities

8. Frequency of Pleasurable Recreational Activities - Past Week

9. Frequency of Expected Pleasure from Recreational Activities - Next Week

The expression scale measures impairment in the expression of emotion and speech. The rating is based on observations during the clinical interviews. The expression symptoms in CAINS are as follows:

1. Facial expression

2. Vocal expression

3. Expressive gestures

4. Quantity of Speech

Bibliography

[1] A. Abbas, V. Yadav, E. Smith, E. Ramjas, S. B. Rutter, C. Benavidez, V. Koesmahargyo, L. Zhang, L. Guan, P. Rosenfield, M. Perez-Rodriguez, and I. R. Galatzer-Levy. Computer Vision-Based Assessment of Motor Functioning in Schizophrenia: Use of Smartphones for Remote Measurement of Schizophrenia Symptomatology. Digital Biomarkers, 5(1):29–36, Jan. 2021. 15

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems, 2022. 21, 67

[3] S. Albanie and A. Vedaldi. Learning Grimaces by Watching TV. In BMVC 2016, Oct. 2016. arXiv: 1610.02255. 49

[4] L. Alphs, A. Summerfelt, H. Lann, and R. Muller. The negative symptom assessment: a new instrument to assess negative symptoms of schizophrenia. Psychopharmacology Bulletin, 25(2):159–163, 1989. 15

[5] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, fifth edition edition, May 2013. 43

[6] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6):1437–1451, June 2018. 30, 34

[7] P. Bagad, M. Tapaswi, and C. G. M. Snoek. Test of Time: Instilling Video-Language Models with a Sense of Time. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Apr. 2023. 22

[8] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. A CLIP-Hitchhiker's Guide to Long Video Retrieval, May 2022. 21

[9] A. Banerjee, U. Bhattacharya, and A. Bera. Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1):3–10, June 2022. Number: 1. 22, 67, 68, 80, 86

[10] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter. The OMG-Emotion Behavior Dataset. In 2018 International Joint Conference on Neural Networks (IJCNN), pages 1–7, July 2018. 52, 55, 62, 63

[11] M. Barthet, C. Trivedi, K. Pinitas, E. Xylakis, K. Makantasis, A. Liapis, and G. N. Yannakakis. Knowing your annotator: Rapidly testing the reliability of affect annotation. arXiv preprint arXiv:2308.16029, 2023. 5, 33

[12] R. Barzilay, N. Israel, A. Krivoy, R. Sagy, S. Kamhi-Nesher, O. Loebstein, L. Wolf, and G. Shoval. Predicting affect classification in mental status examination using machine learning face action recognition system: A pilot study in schizophrenia patients. Frontiers in Psychiatry, 10, 2019. 15

[13] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. arXiv preprint arXiv:1703.01210, 2017. 34

[14] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32, 2019. 17

[15] M. Bishay, P. Palasek, S. Priebe, and I. Patras. SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis. IEEE Transactions

on Affective Computing, 12(4):949–961, Oct. 2019. 16, 17, 20, 44, 51, 55, 61, 75, 83, 84, 89

[16] M. Bishay, S. Priebe, and I. Patras. Can automatic facial expression analysis be used for treatment outcome estimation in schizophrenia? In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1632–1636. IEEE, 2019. 17

[17] M. A. T. Bishay. Automatic Facial Expression Analysis in Diagnosis and Treatment of Schizophrenia. Thesis, Queen Mary University of London, 2020. 44, 63

[18] J. N. d. Boer, A. E. Voppel, S. G. Brederoo, H. G. Schnack, K. P. Truong, F. N. K. Wijnen, and I. E. C. Sommer. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. Psychological Medicine, 53(4):1302–1312, Mar. 2023. Publisher: Cambridge University Press. 15

[19] A. Bulat, S. Cheng, J. Yang, A. Garbett, E. Sanchez, and G. Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. arXiv:2103.16554 [cs], Mar. 2021. arXiv: 2103.16554. 44

[20] F. Burkhardt, A. Derington, M. Kahlau, K. Scherer, F. Eyben, and B. Schuller. Masking speech contents by random splicing: is emotional expression preserved? In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023. 15

[21] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 67–74. IEEE, 2018. 49

[22] W. Carneiro de Melo, E. Granger, and A. Hadid. A Deep Multiscale Spatiotemporal Network for Assessing Depression from Facial Dynamics. IEEE Transactions on Affective Computing, pages 1–1, 2020. 44

[23] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A Short Note about Kinetics-600. Aug. 2018. 60

[24] D. Chakraborty, Z. Yang, Y. Tahir, T. Maszczyk, J. Dauwels, N. Thalmann, J. Zheng, Y. Maniam, N. Amirah, B. L. Tan, and J. Lee. Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6024–6028, Apr. 2018. ISSN: 2379-190X. 15

[25] J. Chang, Z. Lan, C. Cheng, and Y. Wei. Data uncertainty learning in face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 17

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 19, 60, 92

[27] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems, 33:22243–22255, 2020. 8, 19, 44, 60, 92

[28] H. Chou and 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, May 2019. ISSN: 2379-190X. 18, 35

[29] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, pages 1–18, 2021. 66

[30] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences, 114(38):E7900–E7909, 2017. 66, 73

[31] D. Deng, Z. Chen, Y. Zhou, and B. Shi. MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03):2621–2628, Apr. 2020. 20, 44, 54, 63

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009. 60

[33] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. IEEE multimedia, 19(3):34, 2012. 69, 75, 81, 86

[34] M. D'Incà, C. Tzelepis, I. Patras, and N. Sebe. Improving fairness using vision-language driven image augmentation. arXiv preprint arXiv:2311.01573, 2023. 94

[35] P. Ekman and W. V. Friesen. Facial action coding system: Investigator's guide. Consulting Psychologists Press, 1978. vi, 2, 5, 9, 34, 43, 66, 90, 92

[36] R. Ekman. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997. 3, 15

[37] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos. Vocal acoustic analysis and machine learning for the identification of schizophrenia. Research on Biomedical Engineering, 37(1):33–46, Mar. 2021. 15

[38] 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. 15

[39] L. Feldman Barrett and J. A. Russell. Independence and bipolarity in the structure of current affect. Journal of personality and social psychology, 74(4):967, 1998. vi, 2

[40] C. Forbes, J. J. Blanchard, M. Bennett, W. P. Horan, A. Kring, and R. Gur. Initial development and preliminary validation of a new negative symptom measure: The Clinical Assessment Interview for Negative Symptoms (CAINS). Schizophrenia Research, 124(1-3):36–42, Dec. 2010. 16, 51, 75, 83, 88, 99

[41] N. M. Foteinopoulou and I. Patras. Learning from Label Relationships in Human Affect. In Proceedings of the 30th ACM International Conference on Multimedia, pages 80–89, Lisboa Portugal, Oct. 2022. ACM. 66, 75, 76, 84

[42] 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, Nara, Japan, Sept. 2021. IEEE. 66

[43] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 49

[44] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414 [cs, stat], July 2013. arXiv: 1307.0414. 49

[45] R. L. Gorsuch. Factor analysis: Classic edition. Routledge, 2014. 38

[46] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014. 4, 8

[47] M. Guan, V. Gulshan, A. Dai, and G. Hinton. Who Said What: Modeling Individual Labelers Improves Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Apr. 2018. Number: 1. 19

[48] 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, MM '17, pages 890–897, New York, NY, USA, Oct. 2017. Association for Computing Machinery. 18, 35

[49] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016. 29, 35, 39, 48, 49

[50] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang. Bounding Box Regression With Uncertainty for Accurate Object Detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2883–2892, Long Beach, CA, USA, June 2019. IEEE. 10, 17, 25, 26, 28, 29

[51] T. Hirata, Y. Mukuta, and T. Harada. Making Video Recognition Models Robust to Common Corruptions With Supervised Contrastive Learning. In ACM Multimedia Asia, MMAsia '21, pages 1–6, New York, NY, USA, Dec. 2021. Association for Computing Machinery. 19, 44

[52] J. Jang, S. Ye, and M. Seo. Can large language models truly understand prompts? a case study with negated prompts. In A. Albalak, C. Zhou, C. Raffel, D. Ramachandran, S. Ruder, and X. Ma, editors, Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop, volume 203 of Proceedings of Machine Learning Research, pages 52–62. PMLR, 03 Dec 2023. 95

[53] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR, 2021. 21, 67

[54] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In Proceedings of the 28th ACM international conference on multimedia, pages 2881–2889, 2020. 69, 75, 81, 86

[55] S. R. Kay, A. Fiszbein, and L. A. Opler. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. Schizophrenia Bulletin, 13(2):261–276, Jan. 1987. 14, 16, 51, 75, 83, 88, 96

[56] S. R. Kay, L. A. Opler, and J.-P. Lindenmayer. Reliability and validity of the positive and negative syndrome scale for schizophrenics. Psychiatry research, 23(1):99–110, 1988. 3, 5

[57] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. May 2017. 60

[58] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5574–5584, 2017. 17

[59] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning. In Advances in Neural Information Processing Systems, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 19, 44, 92

[60] D. Kim and B. C. Song. Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 35(7):5948–5956, May 2021. Number: 7. 19, 45

[61] P. Kline. An easy guide to factor analysis. Routledge, 2014. 38

[62] D. Kollias and S. Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770, 2018. 53

[63] D. Kollias and S. Zafeiriou. A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild. arXiv:1805.01452 [cs, eess, stat], Dec. 2019. 53, 54, 62

[64] D. Kollias and S. Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. IEEE Transactions on Affective Computing, 12(3):595–606, 2021. 53, 62

[65] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, and A. Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 6088–6097, January 2023. 7

[66] F. Li, C. Gan, X. Liu, Y. Bian, X. Long, Y. Li, Z. Li, J. Zhou, and S. Wen. Temporal Modeling Approaches for Large-scale Youtube-8M Video Understanding. arXiv:1707.04555 [cs], July 2017. arXiv: 1707.04555. 20

[67] H. Li, H. Niu, Z. Zhu, and F. Zhao. CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition, Feb. 2023. arXiv:2303.00193 [cs]. 22, 67, 68, 70

[68] H. Li, M. Sui, Z. Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. arXiv preprint arXiv:2206.04975, 2022. 4, 77, 92

[69] J. Li, R. Socher, and S. C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. arXiv:2002.07394 [cs], Feb. 2020. arXiv: 2002.07394. 17, 25, 26

[70] X. Li, H. Wu, M. Li, and H. Liu. Multi-label video classification via coupling attentional multiple instance learning with label relation graph. Pattern Recognition Letters, 156:53–59, 2022. 20

[71] Y. Li, M. Wang, M. Gong, Y. Lu, and L. Liu. FER-former: Multi-modal Transformer for Facial Expression Recognition, Mar. 2023. arXiv:2303.12997 [cs]. 67, 68, 70

[72] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li. Frozen CLIP Models are Efficient Video Learners. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, Computer Vision – ECCV 2022, Lecture Notes in Computer Science, pages 388–404, Cham, 2022. Springer Nature Switzerland. 21, 67, 70

[73] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, pages 24–32, New York, NY, USA, Oct. 2022. Association for Computing Machinery. ix, xi, 68, 69, 71, 73, 75, 77, 78, 79, 80, 81, 82, 83, 85, 86

[74] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15(2018):11, 2018. 34

[75] C. Long and G. Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015. 19, 35

[76] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, page 646–652, New York, NY, USA, 2018. Association for Computing Machinery. 44

[77] A. Luneski, P. D. Bamidis, and M. Hitoglou-Antoniadou. Affective computing and medical informatics: state of the art in emotion-aware medical applications. Studies in health technology and informatics, 136:517, 2008. 1

[78] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing, 508:293–304, Oct. 2022. 22, 67

[79] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, pages 638–647. Association for Computing Machinery, Oct. 2022. 22, 67, 71

[80] D. Melhart, A. Liapis, and G. N. Yannakakis. Towards general models of player experience: A study within genres. In 2021 IEEE Conference on Games (CoG), pages 01–08, 2021. 42

[81] W. C. d. Melo, E. Granger, and A. Hadid. Depression Detection Based on Deep Distribution Learning. In 2019 IEEE International Conference on Image Processing (ICIP), pages 4544–4548, Sept. 2019. ISSN: 2381-8549. 44

[82] S. Menon and C. Vondrick. Visual classification via description from large language models. ICLR, 2023. 21

[83] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. IEEE

Transactions on Affective Computing, pages 1–1, 2018. viii, 26, 31, 33, 34, 40, 52, 57, 63, 64, 89

[84] 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, Jan. 2019. 18, 26, 33, 40, 89

[85] W. Mou, H. Gunes, and I. Patras. Alone versus In-a-group: A Multi-modal Framework for Automatic Affect Recognition. ACM Transactions on Multimedia Computing, Communications, and Applications, 15(2):1–23, June 2019. 39, 44, 55, 62, 63

[86] L. Nicolescu and M. T. Tudorache. Human-computer interaction in customer service: the experience with ai chatbots—a systematic literature review. Electronics, 11(10):1579, 2022. 1

[87] Y. Ouali, A. Bulat, B. Martinez, and G. Tzimiropoulos. Black Box Few-Shot Adaptation for Vision-Language models, Apr. 2023. 21

[88] S. Parisot, Y. Yang, and S. McDonagh. Learning to Name Classes for Vision and Language Models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Apr. 2023. 22

[89] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler. A deep network for arousal-valence emotion prediction with acoustic-visual cues. arXiv preprint arXiv:1805.00638, 2018. 44, 53, 62

[90] R. W. Picard. Affective computing: challenges. International Journal of Human-Computer Studies, 59(1-2):55–64, 2003. 1

[91] S. Priebe, M. Savill, T. Wykes, R. Bentall, U. Reininghaus, C. Lauber, S. Bremner, S. Eldridge, and F. Röhricht. Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial. The British Journal of Psychiatry, 209(1):54–61, 2016. 51, 59, 62, 63

[92] S. Priebe, M. Savill, T. Wykes, R. Bentall, U. Reininghaus, C. Lauber, S. Bremner, S. Eldridge, and F. Röhricht. Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial. The British Journal of Psychiatry, 209(1):54–61, 2016. 69, 75, 83

[93] F. Qi, X. Yang, and C. Xu. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1074–1083. Association for Computing Machinery, New York, NY, USA, Oct. 2021. 22, 80, 86

[94] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui. Spatiotemporal Contrastive Video Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6964–6974, June 2021. 19, 60

[95] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021. 21, 22, 67, 68, 69, 70, 71, 73, 76, 77, 78, 79, 81, 82

[96] Z. Ren, T. T. Nguyen, Y. Chang, and B. W. Schuller. Fast yet effective speech emotion recognition with self-distillation. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023. 15

[97] M. Rescigno, M. Spezialetti, and S. Rossi. Personalized models for facial emotion recognition through transfer learning. Multimedia Tools and Applications, 79(47):35811–35828, Dec. 2020. 18

[98] F. Rodrigues, F. Pereira, and B. Ribeiro. Gaussian process classification and active learning with multiple annotators. In E. P. Xing and T. Jebara, editors, Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pages 433–441, Bejing, China, 22–24 Jun 2014. PMLR. 19, 35

[99] J. A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161–1178, 1980. 2, 3, 5, 43, 66, 92

[100] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos. Affective processes: Stochastic modelling of temporal context for emotion and facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9074–9084, June 2021. 17

[101] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, Nov. 2021. arXiv:2111.02114 [cs]. 78

[102] D. Setiono, D. Saputra, K. Putra, J. V. Moniaga, and A. Chowanda. Enhancing player experience in game with affective computing. Procedia Computer Science, 179:781–788, 2021. 5th International Conference on Computer Science and Computational Intelligence 2020. 42

[103] Y. Shi and A. K. Jain. Probabilistic Face Embeddings. pages 6902–6911, 2019. 17

[104] P. E. Shrout. Measurement reliability and agreement in psychiatry. Statistical methods in medical research, 7(3):301–317, 1998. 3

[105] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 16

[106] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs], Apr. 2015. arXiv: 1409.1556. 29, 35, 39

[107] E. Smith, E. A. Storch, H. Lavretsky, J. L. Cummings, and H. A. Eyre. Affective Computing for Brain Health Disorders. Springer International Publishing, Cham, 2020. 42

[108] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. arXiv:2001.07685 [cs, stat], Nov. 2020. arXiv: 2001.07685. 17

[109] Y. Tahir, D. Chakraborty, J. Dauwels, N. Thalmann, D. Thalmann, and J. Lee. Non-verbal speech analysis of interviews with schizophrenic patients. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5810–5814, Mar. 2016. ISSN: 2379-190X. 14

[110] Y. Tahir, Z. Yang, D. Chakraborty, N. Thalmann, D. Thalmann, Y. Maniam, N. A. b. A. Rashid, B.-L. Tan, J. L. C. Keong, and J. Dauwels. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. PLOS ONE, 14(4):e0214314, Apr. 2019. Publisher: Public Library of Science. 14

[111] M. Tang, Z. Wang, Z. LIU, F. Rao, D. Li, and X. Li. CLIP4Caption: CLIP for Video Caption. In Proceedings of the 29th ACM International Conference on Multimedia, MM '21, pages 4858–4862, New York, NY, USA, Oct. 2021. Association for Computing Machinery. 67, 71

[112] 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, Jan. 2021. Number: 1 Publisher: Nature Publishing Group. 18, 26, 36, 54

[113] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In International Symposium on Pervasive Computing Paradigms for Mental Health, pages 72–81. Springer, 2015. 15, 16, 61, 84, 89, 94

[114] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Differentiating facial incongruity and flatness in schizophrenia, using structured light camera data. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2427–2430, 2016. 16

[115] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pages 220–223. IEEE, 2016. 15, 16, 51, 61, 84, 94

[116] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. 8, 50, 54

[117] S. Vijay, T. Baltrusaitis, L. Pennant, D. Ongur, J. T. Baker, and L.-P. Morency. Computational study of psychosis symptoms and facial expressions. In Computing and mental health workshop at CHI, volume 2, 2016. 16

[118] M. Wang, J. Xing, and Y. Liu. ActionCLIP: A New Paradigm for Video Action Recognition, Sept. 2021. arXiv:2109.08472 [cs]. 21, 67, 70

[119] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, and S. Wen. Multi-Label Classification with Label Graph Superimposing. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):12265–12272, Apr. 2020. 20

[120] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang. Ferv39k: a large-scale multi-scene dataset for facial expression recognition in

videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20922–20931, 2022. 69, 75, 81, 86

[121] J. B. Williams and M. First. Diagnostic and statistical manual of mental disorders. In Encyclopedia of social work. 2013. 2, 14, 94

[122] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. Girshick. Long-Term Feature Banks for Detailed Video Understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 8, 11, 20, 21, 44, 45, 50, 64, 89

[123] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9):2251–2265, 2018. 67

[124] T.-T. Xie, C. Tzelepis, and I. Patras. Boundary Uncertainty in a Single-Stage Temporal Action Localization Network. arXiv:2008.11170 [cs], Aug. 2020. arXiv: 2008.11170. 17, 28

[125] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller. Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes. 24, 2022. Conference Name: IEEE Transactions on Multimedia. 22, 80, 86

[126] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller. Zero-shot speech emotion recognition using generative learning with reconstructed prototypes. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023. 22, 80, 86

[127] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment, Sept. 2022. arXiv:2209.06430 [cs]. 22

[128] 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. 1, 42, 66

[129] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning, Feb. 2023. arXiv:2302.14115 [cs]. 67

[130] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo. Pose-based body language recognition for emotion and psychiatric symptom interpretation. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 294–301, 2021. 44

[131] G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions: An emerging approach. IEEE Transactions on Affective Computing, 2018. 18, 32

[132] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? In International Conference on Learning Representations, 2023. 22

[133] G. Zara, S. Roy, P. Rota, and E. Ricci. AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation, Apr. 2023. arXiv:2304.01110 [cs]. 67

[134] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang. Zero-shot emotion recognition via affective structural embedding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1151–1160, 2019. 22, 67, 68, 80

[135] F. Zhang, Y. Li, S. You, and Y. Fu. Learning temporal consistency for low light video enhancement from single images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4967–4976, June 2021. 89

[136] L.-W. Zhang, J. Li, S.-J. Wang, X.-H. Duan, W.-J. Yan, H.-Y. Xie, and S.-C. Huang. Spatio-temporal fusion for macro- and micro-expression spotting in long video sequences. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 734–741, 2020. 44

[137] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen. General Facial Representation Learning in a Visual-Linguistic Manner. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18676–18688, June 2022. 21, 77, 78, 79

[138] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng. Graph-Based High-Order Relation Modeling for Long-Term Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8984–8993, June 2021. 20

[139] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 22

[140] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV), 2022. 22