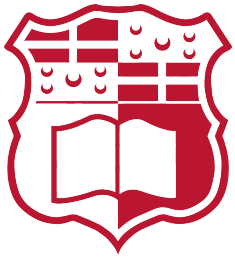


Advancing Affect Modelling via Representation Learning

PhD Dissertation

Kosmas Pinitas

Supervised by Professor Georgios N. Yannakakis

Co-supervised by Dr Konstantinos Makantasis and Professor Antonios Liapis

Institute of Digital Games

University of Malta

March, 2025

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

WWW.UM.EDU.MT

First edition, September 21, 2025

L-Università ta' Malta

FACULTY/INSTITUTE/CENTRE/SCHOOL INSTITUTE OF DIGITAL GAMES

DECLARATION OF AUTHENTICITY FOR DOCTORAL STUDENTS

(a) Authenticity of Thesis/Dissertation

I hereby declare that I am the legitimate author of this Thesis/Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedure

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code N/A.

☑ As a Ph.D. student, as per Regulation 66 of the Doctor of Philosophy Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Sacred Theology student, as per Regulation 17 (3) of the Doctor of Sacred Theology Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Music student, as per Regulation 26 (2) of the Doctor of Music Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

□ As a Professional Doctorate student, as per Regulation 55 of the Professional Doctorate Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

02.2023

Abstract

Affect modelling, the process of constructing computational models capable of recognising and interpreting human emotions, has seen significant advancements with the rise of machine learning. However, key challenges still need to be addressed, particularly in learning generalisable affective representations across different modalities and scenarios, especially in contexts where data is scarce or incomplete. This thesis explores these challenges through the lens of representation learning, with a specific focus on contrastive learning principles. The research is structured across three main parts. First, we investigate the use of supervised contrastive learning to model affective states. Through the development of novel methods, we demonstrate improvements in learning multimodal representations of affect, as evidenced by experiments on datasets such as RECOLA and AGAIN. The second part addresses the challenge of missing modalities in affective data. By leveraging privileged information during training, we introduce techniques that bridge the gap between controlled and in-the-wild affect modelling. Additional experiments demonstrate the robustness of these techniques across multiple modalities and datasets. Finally, the thesis tackles the problem of learning affective representations from a small number of samples, proposing a novel approach using contrastive learning to generate robust representations even in data-constrained environments. This work demonstrates the applicability of these methods across various contexts, including cross-game engagement prediction. The thesis concludes with a discussion of the limitations of the proposed methods and potential directions for future research, including the exploration of more diverse datasets and techniques to further enhance model generalisation and robustness in affective computing.

Dedicated to my family and friends

As you set out for Ithaca hope your road is a long one... (C. P. Cavafy)

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Georgios N. Yannakakis, Konstantinos Makantasis, and Antonios Liapis for their invaluable guidance, unwavering support, and insightful feedback throughout my PhD journey. Their expertise and encouragement have been instrumental in shaping this work, and I am truly grateful for their mentorship.

A special thanks to my fellow PhD students, a.k.a. my "Maltese" friends, who have shared this journey with me—the discussions, laughter, and much-needed coffee made this experience all the more meaningful. I am also grateful to Ahmed for his encouragement, insightful conversations, and, of course, the desserts, a good dessert is exactly what's needed to get through a PhD.

Most importantly, I want to thank my family for always being there for me. Their patience, encouragement, and support during difficult times meant everything, and I am truly grateful for it.

Contents

List of Figures

List of Tables

List of Abbreviations

1

Introduction

On the path to building human-centred artificial Artificial Intelligence (AI), one critical milestone is the development of systems capable of recognising, modelling, and responding to human emotions. While much of AI research has concentrated on areas such as problem-solving, knowledge representation, and natural language understanding (Bengio et al., 2013; Dutta, 1996; Weld et al., 2022), the integration of emotion into AI systems remains a relatively nascent but profoundly impactful endeavour. The intersection of intelligence and emotion holds the potential to transform not only our understanding of human cognition but also the way artificial systems interact with humans in dynamic, real-world scenarios. Human emotions are neither static nor isolated; they are continuously shaped by situational context, physiological processes, and individual experiences. Emotions alter cognitive functions, guide decision-making, and influence social behaviours, as articulated in theories such as the Somatic-Marker Hypothesis (Aday et al. (2017)). Thus, any AI system designed to navigate the social landscape must be capable of perceiving and adapting to these complex affective states. Such systems could create affective loops through human-computer interaction (Yannakakis and Togelius (2018)), fostering meaningful engagement and enabling applications that enhance user experiences, improve learning environments, and assist in therapeutic interventions.

To realise this vision, generalisable models of emotion are needed—models that not only perform well across diverse domains but also remain robust to contextual variability and data scarcity. Emotions often conceptualised as points in a multidimensional affective space (Russell, 1980), demand computational approaches that integrate multimodal signals, including visual, auditory, and physiological cues. Affective Computing (AC) is the interdisciplinary branch of AI that aims to build computational models that capture and interpret emotions. However, developing such models is not only a technical challenge but also an opportunity to bridge the gap between human emo-

tional intelligence and machine learning, paving the way for AI systems that interact seamlessly with humans across contexts. Building robust models of emotion from the ground up in uncontrolled real-world environments remains a significant challenge. This thesis takes a step toward advancing affective computing by focusing on representation learning, a key approach for developing scalable and generalisable models of affect. Representation learning enables the extraction of compact, meaningful, and multimodal embeddings that capture the subtle and complex nature of affective states. Emotions, being inherently multimodal and context-dependent, require computational methods capable of integrating diverse data sources—such as facial expressions, vocal intonations, and physiological signals—while maintaining robustness to noise, incomplete modalities, and variability across domains.

Although representation learning has become a central focus of machine learning and AI research, its application in affective computing remains relatively under-explored. Current approaches often rely on task-specific architectures that are fine-tuned for particular datasets or modalities, limiting their ability to generalise across diverse contexts or adapt effectively to new domains. This lack of generalisability poses significant barriers to deploying affective models in dynamic, real-world scenarios where emotional signals are subtle, multimodal, and subject to external noise. A significant challenge in affective computing is the limited availability of large, high-quality datasets, particularly for real-world, "in-the-wild" applications. Data collected in such environments is often noisy, incomplete, or lacks consistent annotations, making it difficult to develop reliable models. The scarcity and variability of data in affective computing necessitate innovative solutions to overcome these constraints. Addressing these challenges requires methods that are not only data-efficient but also robust to inconsistencies and capable of leveraging multimodal information. Without such advancements, the potential of affective computing to contribute to human-centred AI will remain unrealised.

1.1 | Problem Formulation

Affect modelling, the process of developing computational systems capable of recognising, interpreting, and responding to human emotions, has emerged as a cornerstone of human-centred AI (Picard, 2000). The ability to accurately capture and understand human emotional states is paramount in creating intelligent systems that interact meaningfully and effectively with people in real-world contexts. Emotions are an integral aspect of human experience, influencing not only immediate behaviours but also long-term decision-making, cognitive functions, and social interactions. They play a cru-

cial role in shaping how individuals perceive and respond to their environment, guiding adaptive actions in the face of changing circumstances (Adolphs and Anderson, 2018). Consequently, modelling emotions computationally has become critical in diverse applications, ranging from healthcare and education to entertainment and human-computer interaction. Despite the substantial progress enabled by advancements in machine learning, and deep learning affect modelling remains a highly complex and under-explored challenge, particularly in real-world settings where the diversity, ambiguity, and context-dependency of human emotions pose significant challenges.

One of the core issues in affect modelling is the need to capture and generalise across diverse modalities, such as facial expressions, speech, physiological signals, and behavioural patterns. Emotions are inherently multimodal, and their manifestation varies widely depending on individual differences, cultural contexts, and environmental factors. Learning representations that integrate multimodal data effectively while maintaining robustness to noise and missing information is a key technical challenge. The problem becomes even more acute in "in-the-wild" scenarios, where emotional signals are less controlled, often subtle, and frequently confounded by external factors. Unlike controlled laboratory settings, real-world environments introduce unpredictable conditions such as varying lighting, background noise, occlusions, and movement artefacts, which can degrade the quality of the collected data. Emotional expressions may also be less pronounced or deliberately masked in such settings, making their recognition even more difficult. These factors create a significant gap between the performance of affect models trained on controlled datasets and their ability to generalise effectively to real-world applications.

Another significant challenge in affect modelling is domain variability, which arises from the diverse contexts in which emotions are expressed and experienced. Emotional expressions and affective dynamics differ widely across application domains such as healthcare, education, entertainment, and gaming. Each domain is characterised by unique emotional triggers, behavioural patterns, and environmental factors, making it difficult for models trained in one context to generalise effectively to another. For example, in healthcare settings, emotions may be influenced by stress or vulnerability, requiring models to focus on subtle physiological signals. In education, emotions like frustration or curiosity may be primarily conveyed through facial expressions and speech. In contrast, engagement modelling in video games—a key focus of this thesis—requires capturing the distinct emotional dynamics shaped by game genres, mechanics, and aesthetics. Each domain presents a unique set of challenges, and addressing these requires methodologies capable of adapting to varying emotional landscapes while preserving critical domain-specific nuances.

Addressing the challenges of affect modelling necessitates the development of methods that are data-efficient, generalisable, and robust across diverse modalities and scenarios. Current approaches often struggle with the variability and complexity inherent in affective data, particularly in real-world settings where multimodal inputs may be noisy or incomplete, and domain-specific nuances further complicate the task. To overcome these obstacles, this thesis adopts a representation learning approach, which aims to extract compact, informative, and generalisable embeddings from raw data. These embeddings serve as a foundation for building affective models that can effectively integrate multimodal signals and adapt to varying contexts. A key focus of this work is the incorporation of contrastive representation learning principles into affect modelling. Contrastive learning has gained prominence in machine learning for its ability to improve the discriminative power of learned representations by emphasising similarities and differences between data samples. By aligning embeddings of semantically similar samples and contrasting them with dissimilar ones, contrastive learning enables models to capture subtle patterns in complex, multimodal data.

To this end, this thesis incorporates three complementary approaches—Contrastive Learning (CL), Few-Shot Learning (FSL), and Learning Using Privileged Information (LUPI)—to tackle, respectively, the three interrelated challenges at the core of this thesis: data scarcity, multimodal integration, and domain variability. Contrastive learning enhances representation quality and facilitates multimodal fusion by uncovering shared structures across modalities. Few-shot learning addresses the problem of data scarcity by allowing models to generalise effectively with minimal labelled samples, enabling scalable solutions for diverse contexts. This capability is essential in affective computing, where collecting and annotating large-scale datasets is often infeasible. Meanwhile, the LUPI framework bridges the gap between controlled and real-world settings by leveraging auxiliary information during training. By utilising privileged information—available only at training time—models can learn more robust representations that enhance their performance in challenging "in-the-wild" scenarios. These methods address the dual demands of generalisability and domain specificity, ultimately contributing to the creation of affective computing systems that perform reliably in dynamic, real-world environments.

1.2 | Research Questions and Objectives

The field of affect modelling is fraught with challenges, from the scarcity of labelled data to the inherent complexity of learning robust and generalisable affective representations

in different modalities and domains. Addressing these challenges requires a focused investigation into the limitations of current approaches and the development of novel techniques that balance adaptability, scalability, and data efficiency.

Aim

The aim of this thesis is to advance affect modelling by developing representation learning methods that enhance multimodal affective representations, mitigate the impact of missing modalities, and improve generalisation in data-scarce conditions.

Objectives

1. To investigate how supervised contrastive learning can improve the learning of multimodal affective representations by aligning semantically similar samples and contrasting dissimilar ones.

2. To examine strategies for handling missing modalities in affective datasets, using privileged information during training to bridge the gap between controlled (in-vitro) and real-world (in-vivo) affect modelling.

3. To explore how teacher models trained in controlled settings can guide student models under the LUPI framework, thereby improving the quality and generalisability of affective models.

4. To develop few-shot learning approaches that enable robust affective representations to be learned with minimal labelled data.

5. To analyse the impact of different input feature representations on affective model performance and to evaluate the capacity of the proposed methods to generalise to novel domains such as game-based engagement prediction.

These objectives consolidate the methodological strands of the thesis—contrastive learning, learning using privileged information, and few-shot learning—into a cohesive framework designed to address data scarcity, modality incompleteness, and domain variability in affective computing. Rather than formulating an extensive list of narrow research questions, this thesis is guided by a single overarching aim supported by five specific objectives. This structure is intended to provide a more cohesive framework that unifies the methodological contributions into a coherent research plan.

1.3 | Contributions

This thesis makes significant contributions to the field of affect modelling by addressing critical challenges related to data scarcity, multimodal integration, and domain generalisation. Through the lens of representation learning, the research introduces innovative techniques to develop robust, scalable, and generalisable affective computing models. The key contributions of this thesis are as follows:

Contrastive Representations of Affect

1. **Development of Contrastive Learning Frameworks for Affective Representations:**
   This thesis extends supervised contrastive learning principles to the domain of affect modelling, focusing on the integration of multimodal data such as facial action units, speech, and physiological signals. The proposed frameworks align semantically similar samples while contrasting them with dissimilar ones, enabling models to capture the richness and subtlety of affective signals, particularly in real-world conditions.

2. **Design of Positive and Negative Sample Definitions for Emotional Data:**
   Several methodologies have been explored for constructing meaningful positive and negative pairs tailored to the unique characteristics of emotional data. This contribution enhances the discriminative power of contrastive learning.

Learning from Missing Modalities

1. **Utilisation of Privileged Information for Multimodal Affect Modelling:**
   The thesis leverages the Learning Using Privileged Information framework to bridge the gap between controlled and in-the-wild scenarios. By incorporating auxiliary data available only during training, the proposed methods improve the robustness and generalisability of affective models, even when key modalities are missing during inference.

2. **Optimisation of Teacher-Student Models in Affect Modelling:**
   A teacher-student paradigm is used to guide the learning process using privileged information. The teacher model provides high-level abstractions and contextual insights during training, resulting in robust student models. This approach addresses the challenges posed by missing modalities and inconsistent data in real-world applications. Moreover, this work investigates two different teacher pre-

taining approaches via end-to-end modelling and supervised contrastive learning. The latter has been shown to produce more robust teachers which in turn yield student models of higher predictive power.

Affect Modelling with Limited Data

1. **Integration of Few-Shot Learning for Affective Representations:**
   This thesis develops a few-shot learning framework tailored to affective computing, enabling models to generalise effectively with minimal labelled data. By combining few-shot learning with contrastive learning principles, the proposed approach generates robust representations in data-constrained environments.

2. **Introduction of a Novel Loss Function for Few-Shot Learning**
   This thesis introduced a novel loss function for few-shot learning. This loss function is inspired by the widely used silhouette score and accounts for both inter-class cohesion and inter-class separation. This novel loss function is thoroughly tested across dissimilar affective computing datasets

3. **Evaluation of Input Feature Representations in Data-Limited Scenarios:**
   A comprehensive analysis is conducted to assess the impact of different input feature representations—such as multimodal data, audiovisual signals, and deep embeddings—on the performance and generalisation of affective models. The findings provide insights into the design of effective feature extraction pipelines for affect modelling.

4. **Application to Cross-Domain and Novel Contexts:**
   The proposed methods are validated across multiple domains, including cross-game engagement prediction, to evaluate their generalisability and scalability. This contribution demonstrates the practical applicability of the developed framework to novel domains and dynamic real-world scenarios.

5. **Development of the *GameVibe* Dataset:**
   A secondary contribution of this thesis is the design and collection of the *GameVibe* Dataset, a benchmark specifically tailored for affect modelling and experience prediction across first-person shooter games. This dataset was collected in collaboration with researchers and academics at the Institute of Digital Games of the University of Malta.

These contributions collectively advance the state of the art in affective computing by addressing critical gaps in representation learning for affect modelling. By integrating

contrastive learning, privileged information, and few-shot learning, this research provides a robust foundation for building affective systems that perform reliably across diverse and challenging environments.

1.4 | Publications

This section outlines the papers published during the course of the PhD studies. The first part focuses on works that directly contribute to this thesis, whereas the second part highlights papers that are not part of the thesis research. At the time of writing this thesis, a total of 10 papers have been published, and 2 more papers are currently under review or in the submission process. Of these 10 papers, 2 are journal articles, 4 have been published in conference proceedings and 4 have been published in workshops. Although all the research published during the PhD is centred around advancing affect modelling via representation learning, some of the work is more exploratory and does not directly align with the thesis but still contributes to the broader field of affective game computing and game research. (Yannakakis and Melhart, 2023)

1.4.1 | Part of the Thesis Work

The papers presented in this section are integral to the thesis work. Thus, they are listed along with the corresponding chapters they contribute to.

1. **Pinitas, K.**, Makantasis, K., Liapis, A., & Yannakakis, G. N. (2022, November). Supervised contrastive learning for affect modelling. In Proceedings of the 2022 International Conference on Multimodal Interaction (pp. 531-539).

   This paper contributes to Chapter 5 since it introduces Supervised Contrastive Learning for Affective Computing

2. Makantasis, K., **Pinitas, K.**, Liapis, A., & Yannakakis, G. N. (2023). From the lab to the wild: Affect modelling via privileged information. *IEEE Transactions on Affective Computing*.

   This paper contributes to Chapter 6. In particular, it was the first paper that extensively explored the Learning Using Privileged information across different affect modelling paradigms

3. Barthet, M., Kaselimi, M., **Pinitas, K.**, Makantasis, K., Liapis, A., & Yannakakis, G. N. (2024). *GameVibe*: a multimodal affective game corpus. *Nature Scientific Data*, 11(1), 1306.

This paper introduced *GameVibe*, a dataset tailored for addressing the challenge of domain generalisation for affective computing. Although not a primary contribution of this work it serves as a benchmark for evaluating the few-shot learning models of Chapter 7

4. **Pinitas, K.**, Makantasis, K. & G. N. Yannakakis. Across-game engagement modelling via few-shot learning. In Proceedings of the First Workshop on Computer Vision for Videogames, European Conference on Computer Vision (ECCV) (2024).

   This paper contributes to Chapter 7 as it introduces a novel framework for few-shot domain generalisation framework for affective computing

5. **Kosmas, P.**, Rasajski, N., Mankantasis, K., & Georgios, Y. (2024, October). Silhouette Distance Loss for Learning Few-Shot Contrastive Representations. In Classifier Learning from Difficult Data (pp. 32-39). PMLR.

   This paper contributes to Chapter 7 since it introduces the novel contrastive loss. The Silhouette Distance Loss is tailored for few-shot learning and an analysis of its behaviour is presented in this thesis.

1.4.2 | Out of Scope Work

The papers presented in this section describe supplementary exploratory studies conducted during the PhD, but not included in the thesis work. Since these studies do not contribute directly to the thesis, a summary of their content is provided below.

1. **Pinitas, K.**, Renaudie, D., Thomsen, M., Barthet, M., Makantasis, K., Liapis, A., & Yannakakis, G. N. (2023, October). Predicting Player Engagement in Tom Clancy's The Division 2: A Multimodal Approach via Pixels and Gamepad Actions. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 488-497).

   This study presents a large-scale multimodal dataset designed for the analysis and prediction of player engagement in commercial-standard video games. The dataset was collected from 25 participants playing the action role-playing game *Tom Clancy's The Division 2*, with engagement levels annotated continuously over time using a specialised annotation tool. After cleaning and processing, the dataset includes approximately 20 hours of annotated gameplay footage along with logged gamepad actions. Preliminary experiments are conducted to predict long-term player engagement using Convolutional Neural Network (CNN) architectures,

leveraging both in-game footage and game controller data. The results demonstrate the ability to predict player engagement with notable accuracy, particularly when fusing visual data with controller inputs. These findings support the hypothesis that long-term engagement, spanning up to an hour of gameplay, can be effectively predicted using only gameplay visuals and gamepad interactions.

2. **Pinitas, K.**, Makantasis, K., Liapis, A., & Yannakakis, G. N. (2022, July). RankNEAT: outperforming stochastic gradient search in preference learning tasks. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 1084-1092).

   This work introduces RankNEAT, an evolutionary algorithm that uses neuroevolution of augmenting topologies for preference learning. Unlike traditional gradient-based methods like RankNet, which struggle with noisy and subjective labels, RankNEAT optimises architectures and effectively selects features, reducing overfitting. Evaluated on predicting player arousal from gameplay footage across three games, RankNEAT consistently outperforms gradient-based approaches, demonstrating its efficiency and viability as an alternative for preference learning in affective computing.

3. Barthet, M., Trivedi, C., **Pinitas, K.**, Xylakis, E., Makantasis, K., Liapis, A., & Yannakakis, G. N. (2023, September). Knowing your annotator: Rapidly testing the reliability of affect annotation. In 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-8). IEEE.

   The time-intensive nature of affect annotation poses significant challenges to creating large-scale datasets with valid and reliable affect labels. To address the lack of tools for effectively assessing annotator reliability, this study proposes general quality assurance (QA) tests for real-time continuous annotation tasks. Assuming the use of audiovisual stimuli, such as videos, the research introduces and evaluates two QA tests: one visual and one auditory. These tests were validated with 20 annotators who first completed the QA tests and subsequently annotated engagement levels in gameplay videos. Results reveal that, as expected, trained annotators demonstrate greater reliability than even the most skilled untrained crowdworkers. Notably, the proposed QA tool predicts annotator reliability with high degrees of accuracy, reducing resource use, effort, and costs while maximising the reliability of labels in affective datasets.

4. Makantasis, K., **Pinitas, K.**, Liapis, A., & Yannakakis, G. N. (2022, October). The invariant ground truth of affect. In 2022 10th International Conference on Affec-

tive Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-8). IEEE.

This study integrates causation theory into affective computing to improve the reliability of affective annotations and models. It challenges traditional reliance on subjective affect labels by focusing on the causal relationships between affect elicitation, manifestation, and annotation, which remain invariant across tasks and participants. Using causation-inspired methods, the research detects outliers and builds robust affective models, validated within the context of digital games. Experimental results show improved outlier detection and model accuracy, marking a pioneering step toward generalised affect modelling through causation tools.

5. **Pinitas, K.**, Rasajski, N., Barthet, M., Kaselimi, M., Makantasis, K., Liapis, A., & Yannakakis, G. N. (2024). Varying the context to advance affect modelling: A study on game engagement prediction. In Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction (ACII).

   Affective computing faces a significant challenge: the limited ability of affective models to generalise across varying contextual factors within the same task. Despite being widely acknowledged, this issue persists due to the lack of large-scale datasets that encompass rich and diverse contextual information within a specific domain. To address this gap, this paper introduces *GameVibe*, a novel corpus specifically designed to tackle the problem of contextual diversity. The dataset is derived from 30 First-Person Shooter (FPS) games, representing a wide range of game modes and designs within the same domain. It includes 2 hours of annotated gameplay videos, with engagement levels continuously annotated by 20 participants. Preliminary analyses of this corpus highlight the challenges of generalising affective predictions across varying contexts in similar tasks within affective computing. These findings underscore the complexity of the issue and pave the way for future research, encouraging deeper exploration of this critical but under-explored aspect of affect modelling.

1.5 | Thesis Structure

This thesis is organised into several chapters, each addressing a critical aspect of the research undertaken and building towards the overarching goals of the study. The structure is designed to provide a logical progression from foundational concepts to the key contributions and findings of the work. Below is an outline of the thesis structure:

- **Chapter 2:** This chapter systematically explores the advancements in affective computing. It begins with an overview of affect modelling, tracing the evolution from unimodal to multimodal approaches powered by deep learning. The discussion then transitions to engagement and player modelling, highlighting their applications in interactive domains. Subsequent it delves into the role of contrastive learning, the LUPI paradigm, and few-shot learning methodologies, showcasing their contributions to overcoming challenges like data scarcity and generalisation in diverse scenarios.

- **Chapter 3:** This chapter outlines the key methodologies employed in this thesis, focusing on representation learning to capture complex patterns in data. It introduces contrastive representation learning for constructing discriminative latent spaces, the Learning Using Privileged Information paradigm to enhance learning with auxiliary data during training, and the role of teacher models in knowledge transfer. Additionally, the chapter explores few-shot learning as a solution for generalisation from limited labelled examples, emphasizing the importance of well-structured embeddings for task adaptability.

- **Chapter 4:** This chapter provides an overview of the three datasets used in this thesis: the RECOLA database, the AGAIN dataset, and the *GameVibe* corpus. These datasets span diverse modalities and domains, from dyadic interactions in RECOLA to game player arousal in AGAIN and viewer engagement in FPS gameplay in *GameVibe*. The inclusion of these datasets ensures robust evaluations of the proposed methodologies across varied contexts.

- **Chapter 5:** This chapter presents a novel approach to affect modelling using representation learning to derive affect-infused representations. By leveraging contrastive labels derived from affect annotations, the chapter demonstrates the pretraining of encoder architectures that effectively capture affective dynamics. Experiments show that Supervised Contrastive Learning (SCL) models outperform baselines, particularly in arousal prediction, and highlighted the importance of contrastive labelling strategies in improving representation quality. This work underscores the potential of SCL for developing robust, participant-agnostic affect modelling systems.

- **Chapter 6:** This chapter focuses on a methodology for affect modelling in real-world scenarios using privileged information and supervised contrastive learning. The chapter explores how privileged information, accessible only during training, facilitates the transfer of affect models from controlled environments to real-world

settings. Experiments demonstrate that student models, trained via knowledge transfer from teacher models with access to privileged data, achieve robust performance without relying on costly or intrusive modalities during deployment. This approach underscores its applicability in multimodal affect modelling tasks for real-world environments.

- **Chapter 7:** This chapter addressed the challenge of affect modelling with limited data by leveraging few-shot learning to enable generalisation across diverse domains with minimal labelled samples. The proposed framework decomposes multidomain affect classification into domain-specific few-shot tasks, allowing models to capture domain-specific patterns while addressing the inherent data scarcity in affective computing.
- **Chapter 8:** The final chapter summarises the key contributions of the thesis and reiterates its significance in advancing affect modelling and human-centred AI. It also discusses the limitations of the work and concludes with recommendations and directions for future research.

1.6 | Summary

This chapter presented the core research questions and the key challenges addressed throughout this thesis, establishing a structured roadmap to navigate these complexities. The primary objective is to advance affect modelling, particularly in dynamic, real-world environments. These settings pose unique challenges, including multimodal data integration, and learning from incomplete and limited labelled data. To address these issues, the thesis adopts state-of-the-art representation learning techniques, with an emphasis on contrastive learning to improve representation quality, few-shot learning to enhance generalisation with minimal labelled samples, and the Learning Using Privileged Information paradigm to leverage auxiliary data during training for more robust model development. In addition to addressing technical challenges, this chapter also provided an overview of the contributions made through various published works. It distinguished between papers directly contributing to the thesis and those serving as supplementary research, showcasing the breadth and depth of the work conducted during the PhD. These contributions include methodological innovations, dataset development, and domain-specific applications that collectively push the boundaries of affective computing. Finally, the chapter concluded with an outline of the thesis structure, highlighting the logical progression of topics and methodologies.

2

Literature Review

This chapter presents the essential background and notable literature. Section 2.1 provides an overview of affect modelling. Section 2.2 surveys literature on contrastive learning and emphasises contrastive representations for affect. Section 2.3 reviews work on learning using privileged information. Section 2.4 presents scientific papers in few-shot learning.

2.1 | Affect Modelling: Traditional and Contemporary Perspectives

This section surveys work on unimodal and multimodal affect modelling. Next it highlights notable contributions in modelling complex affect constructs such as engagement and moves on presenting papers on player modelling the field that models affect in games

2.1.1 | Unimodal Affect Modelling

Emotions can be elicited through various stimuli, such as speech (audio), and facial expressions (video) (Calvo et al., 2015; Picard, 2000). Thus it comes as no surprise that such modalities have been widely used to model affect. Traditionally, affect modelling based on visual information relied heavily on domain knowledge to manually create high-level visual features or utilise classic pattern recognition techniques such as SURF or Histogram of Oriented Gradients (Dahmane and Meunier, 2011; Zheng et al., 2010). However, the advent of deep learning has brought new possibilities since it has automated the representation extraction process, significantly advancing pixel-based affect

modelling. For instance, Baveye et al. (2015) were pioneers in applying deep learning for emotion recognition in videos, demonstrating that convolutional neural network representations show promise for analysing affective content in films. Subsequent studies, such as those by Ng et al. (2015) and Haddad et al. (2020), leveraged deep CNNs and transfer learning techniques to perform emotion recognition on small datasets and predict emotions from raw facial expression footage using 3D CNN, respectively.

The field of emotion recognition via audio has seen significant advancements over the years. Initially, the approach relied heavily on domain knowledge and handcrafted audio features. Experts would carefully design features that could capture various aspects of speech, such as pitch, energy, formant frequencies, and prosody, which are crucial for identifying emotional states. These features were then processed using classic pattern recognition methods like support vector machines, Gaussian mixture models, and hidden Markov models (El Ayadi et al., 2011; Schuller et al., 2003). This methodology required extensive feature engineering and expertise, making it time-consuming and limited in its ability to generalise across different datasets and conditions.

Similarly to pixel-based affect modelling, deep learning brought about a paradigm shift in emotion recognition via audio. Unlike traditional methods, deep learning techniques can learn hierarchical feature representations from raw data, reducing the reliance on manual feature extraction. Huang et al. (2014) was one of the first studies that demonstrated this shift by employing contractive CNNs to learn candidate features from audio signals. These CNNs were designed to contract the input space, making the learned features more robust to variations in the data. The extracted features were then refined through a semi-CNN, a specialised network structure that focuses on capturing affect-salient features, which are particularly relevant for emotion recognition. This deep learning approach significantly outperformed traditional methods, showcasing the ability of CNNs to capture complex patterns in audio data that are related to emotional states.

In another notable study, Kwon et al. (2021) introduced a lightweight dilated CNN model for speech emotion recognition. This model utilised dilated convolutions, which allow the network to have a larger receptive field without increasing the number of parameters. This is particularly useful for capturing temporal dependencies in speech signals over a longer context, which is crucial for understanding the prosodic and intonational features that convey emotion. This model also employed a multi-learning approach, which involves using multiple loss functions or tasks to enhance the learning

process. This approach helps the model to generalise better and improves its ability to recognise emotions across different speakers and datasets. The model was evaluated on several benchmark datasets and achieved high recognition accuracy.

Physiological responses are a crucial modality for capturing emotions, as they provide quantifiable indicators of affective states. These responses include changes in brain activity, heart rate, skin conductivity, and other autonomic nervous system functions, all of which can reflect different emotional states (Bradley and Lang, 2000). For instance, increased heart rate and galvanic skin response are often associated with heightened arousal or stress. Once again, early psychophysiological research in this area focused on designing handcrafted features to map physiological signals to specific emotional states. These features included metrics like heart rate variability, amplitude of skin conductance responses, and patterns in brainwave frequencies. Techniques from classic pattern recognition and statistical modelling were then applied to analyse these features and classify the corresponding emotional states (Holmgård et al., 2015; Mandryk and Atkins, 2007).

The introduction of deep learning techniques has transformed physiology-based affect modelling. Martinez et al. (2013) were pioneers in applying deep learning to this field, utilizing convolutional neural networks to automatically extract and analyse complex features from physiological data. CNNs are particularly well-suited for this task because they can learn spatial hierarchies of features directly from raw data, capturing patterns of emotional states. This approach eliminated the need for manual feature extraction, enabling the discovery of subtle and complex relationships within the data.

Further advancements were made by Harper and Southern (2020), who developed an end-to-end model for classifying emotional valence based on heartbeat data. Their model not only employed CNNs for feature extraction but also integrated a Bayesian framework to quantify the uncertainty of predictions. This addition is significant because it estimated confidence in the model's outputs, which is crucial in applications where understanding the reliability of predictions can inform subsequent decisions or actions. Giannakakis et al. (2019) contributed to the field by employing a multi-kernel 1D CNN to analyse heart rate variability (HRV), a complex and informative physiological marker. By using multiple kernels, the model could capture different aspects of the HRV signal, such as time-domain, frequency-domain, and non-linear characteristics. This comprehensive analysis allowed for the identification of unique signatures associated with stress, demonstrating the potential of deep learning models to differentiate between various emotional and physiological states based on signal characteristics.

This thesis considers both multimodal and unimodal scenarios, ensuring a comprehensive analysis of affect modelling across different data configurations. Rather

than relying solely on simple end-to-end modelling, it explores advanced representation learning techniques to enhance the robustness of affective models. By leveraging methods such as supervised contrastive learning, learning using privileged information, and few-shot representation learning, this work aims to uncover more meaningful and transferable representations, ultimately improving generalisation across diverse domains and conditions.

2.1.2 | Multimodal Affect Modelling

The integration of multiple user modalities to affect models, known as multimodal affect modelling, has become a significant area of research due to its potential to provide a more comprehensive understanding of emotional states by combining different types of data (Abdullah et al., 2021; Sebe et al., 2005). This approach leverages the strengths of various modalities—such as visual cues, audio signals, and physiological data—to capture the nuanced and multifaceted nature of human emotions more accurately than unimodal systems.

One of the pioneering studies in this field by Martínez and Yannakakis (2014) investigated the fusion of different modalities using deep learning techniques. This study highlighted the advantages of combining data from multiple sources, such as facial expressions, speech, and physiological signals, to improve the accuracy of emotion recognition systems. By employing deep learning, the researchers were able to learn feature representations from each modality, which were then fused to make predictions about the user's emotional state. This approach marked a significant advancement over traditional methods, which often relied on handcrafted features and could not fully exploit the rich information available from multiple data sources. Tzirakis et al. (2017) provided additional evidence of the superiority of deep learning over traditional methods in multimodal affect modelling. They utilised a combination of a CNN and a deep residual network to extract features from both speech and video modalities. The deep residual network, in particular, allowed for the creation of a deeper model, thus capturing more complex patterns in the data. The results from this study highlighted the enhanced capability of deep learning architectures to integrate and process diverse types of data for more accurate emotion recognition.

In another innovative study, Guo et al. (2019) compared different combinations of modalities, including eye images, eye movement data, and electroencephalograms. Their research demonstrated that these different modalities provide complementary information, which can be used to improve the accuracy of emotion recognition systems. This study also highlighted the importance of selecting appropriate modalities and fusion

techniques to maximise the effectiveness of multimodal systems. Zhang et al. (2021) took a unique approach by utilizing a Convolutional Long Short-Term Memory (LSTM) network and a 1D CNN to extract spatio-temporal and bio-sensing features. The Convolutional LSTM was used to capture the temporal dynamics in video data, while the 1D CNN was employed to analyse bio-sensing data, such as heart rate or skin conductivity. This dual approach allowed for a comprehensive analysis of both the temporal and physiological aspects of emotion, further demonstrating the capabilities of deep learning in multimodal affect modelling.

While previous studies have focused on deep learning-based multimodal fusion for affect modelling, this work extends beyond simple end-to-end approaches by incorporating advanced representation learning techniques. It considers both multimodal and unimodal scenarios, allowing for a more flexible and comprehensive analysis of affective states. By leveraging methods such as supervised contrastive learning, learning using privileged information, and few-shot learning, this study aims to improve the robustness and generalisation of affect models, ensuring they remain effective across varying data conditions.

2.1.3 | Modelling of Engagement

Engagement encompasses cognitive, affective, and behavioral components, thus playing a vital role in Human-Computer Interaction (HCI) (Appleton et al., 2006; Bindl and Parker, 2010). It is a multifaceted construct that reflects a user's involvement, interest, and emotional connection with an interface or activity. Consequently understanding and modelling engagement are essential steps for creating more effective and adaptive HCI systems. Several studies have focused on different aspects of user engagement, leveraging various techniques to measure and predict it. For example, Dermouche and Pelachaud (2019) developed an LSTM-based model designed to predict real-time engagement during dyadic interactions. Their model utilised data from facial expressions, head movements, and gaze, providing a comprehensive view of the user's engagement level. LSTM networks were particularly advantageous due to their ability to capture temporal dependencies in the data, such as changes in facial expressions over time, which are critical for accurately predicting engagement.

In education, Ting et al. (2013) employed Bayesian Networks to model student engagement in virtual learning environments. Bayesian Networks are probabilistic graphical models that represent a set of variables and their conditional dependencies. This approach allowed the researchers to model the complex interplay across factors that influence student engagement, such as motivation, attention, and emotional state. By

analysing these relationships, they could predict the level of engagement and identify students at risk of disengagement, thus enabling timely interventions to improve educational outcomes. Fan et al. (2016) took a different approach by developing a robotic coach system aimed at managing multi-user engagement. This system was designed to interact with multiple users simultaneously, using multiple cues such as speech, gestures, and facial expressions to assess and enhance engagement levels. The robotic coach could adapt its behaviour based on real-time feedback from users, providing personalised encouragement and adjusting the difficulty of tasks to maintain optimal engagement. This study demonstrated the potential of intelligent systems to actively manage and foster engagement in group settings.

Apart from engagement modelling, this study also focuses on modelling the core affect dimensions of arousal and valence, aiming to improve emotion recognition through advanced representation learning. While prior work predominantly employs multimodal fusion and sequence-based models, this research explores both multimodal and unimodal scenarios, leveraging techniques such as supervised contrastive learning and learning using privileged information. By enhancing feature representations and improving generalisation, this study moves beyond traditional predictive approaches to develop more robust affective models.

2.1.4 | Player Modelling

In the context of games, player modelling involves the creation of computational models that accurately predict a player's behaviours and emotions while interacting with a game (Yannakakis and Togelius, 2018). This field followed a similar path to most affective computing domains since early studies primarily relied on hand-crafted features to capture various aspects of the gaming experience. For example, Frommel et al. (2018) utilised data from a graphics tablet, along with gameplay performance metrics, to predict the emotional states of players. This approach involved analysing the players' interactions and inputs to derive features indicative of their emotional responses. Similarly, Melhart et al. (2021b) focused on hand-crafted features, specifically those that describe the state context and player actions, to develop general models of player arousal. This highlighted the importance of contextual information in understanding player experiences, such as game events, and environmental settings. By incorporating these features, the models could better predict fluctuations in player arousal, providing insights into how different game elements influence emotional responses.

The advancements in machine learning, particularly deep learning, allowed research to move towards using raw gameplay data to model player experiences. Makantasis

et al. (2019) leveraged Convolutional Neural Networks to predict player arousal directly from raw gameplay footage. This approach utilised the rich visual and contextual information available in video data, allowing the model to learn complex patterns associated with player arousal without the need for extensive manual feature engineering. The use of CNNs marked a significant advancement in the ability to capture and analyse the nuanced visual and auditory elements that contribute to a player's experience. Moreover, evolutionary approaches have also been explored in the context of player modelling, for example in early attempts to predict arousal from gameplay videos using preference learning methods (Pinitas et al., 2022a). While indicative of the breadth of techniques applied in this domain, such studies remain exploratory and illustrate the diversity of methods rather than establishing dominant paradigms.

Beyond the core dimensions of emotion, substantial progress has been made in modelling more complex constructs such as player engagement. For instance, Xue et al. (2017) proposed a Dynamic Difficulty Adjustment framework aimed at maximising player engagement. This system dynamically adjusts the game's difficulty based on the player's skill level and performance, ensuring an optimal balance between challenge and skill, which is critical for maintaining engagement. Huang et al. (2019) introduced a two-stage player engagement modelling approach using Hidden Markov Models. This method involved first predicting the player's engagement state and then modelling the transitions between these states over time, providing a more nuanced understanding of how engagement evolves during gameplay.

Melhart et al. (2020) took a unique approach by using chat logs as a proxy for engagement, training a neural network to predict moment-to-moment engagement levels. This study highlighted the potential of leveraging real-time player interactions and communications to assess engagement dynamically. Pinitas et al. (2023) focused on long-term engagement, employing pretrained CNN models and time-conditioning techniques to predict engagement over extended periods in the game *Tom Clancy's The Division 2* (Ubisoft, 2020). This approach aimed to understand how players' engagement levels change over time and what factors contribute to sustained engagement. Recently, Pan et al. (2023) developed a deep learning model for estimating game streamers' engagement by analysing gameplay footage, audio, and facial expressions. This comprehensive approach integrates multiple modalities to capture the diverse signals that indicate a streamer's engagement level, demonstrating the growing sophistication and accuracy of engagement modelling techniques in gaming contexts.

Unlike traditional player modelling studies that primarily focus on predicting individual player behaviours and emotions, this research extends its scope to dyadic interactions, capturing the dynamic interplay between two individuals. In addition to ex-

amining both multimodal and unimodal settings, it also leverages advanced representation learning techniques such as supervised contrastive learning and learning using privileged information and few-shot representation learning. By focusing on diverse testbeds, this thesis aims to develop more robust and adaptable models for understanding emotions in gaming and social contexts.

2.2 | Contrastive Representation Learning

Contrastive learning has emerged as a powerful approach in self-supervised, unsupervised and fully supervised learning, particularly in the domain of representation learning. By learning to distinguish between similar and dissimilar pairs of data points, contrastive learning frameworks can effectively learn useful representations without requiring labelled data. This section explores key methodologies and applications of contrastive learning in affective computing.

2.2.1 | Contrastive Learning Methodologies

Contrastive learning has evolved significantly, with several methods becoming widely adopted due to their effectiveness in various tasks. Each of these methods contributes unique insights and techniques that have broadened the applicability and effectiveness of representation learning. This section elaborates on ten of the most influential contrastive learning methodologies.

SimCLR is one of the seminal works in the domain of contrastive learning, developed by Chen et al. (2020a). It leverages data augmentation to create positive pairs from the same image and contrasts them with negative pairs from different images. The key innovation in SimCLR is its use of extensive augmentations, such as random cropping, colour distortions, and Gaussian blur, to generate diverse views of the same image. This approach forces the model to learn invariant features that remain consistent across various transformations. SimCLR relies on large batch sizes to provide a sufficient number of negative examples within each batch, which helps the model to learn discriminative features effectively. The method uses a normalised temperature-scaled cross-entropy loss (NT-Xent) to measure the similarity between positive pairs relative to negative pairs. This framework has become a cornerstone in contrastive learning due to its simplicity and effectiveness, inspiring numerous follow-up studies and enhancements.

Momentum Contrast (MoCo) addresses some limitations of SimCLR, particularly the need for large batch sizes. Developed by Chen et al. (2020b), MoCo introduces a

momentum-based encoder to maintain a large and consistent dictionary of negative samples. The core idea is to use a queue of encoded representations from previous batches as negative samples, which decouples the batch size from the number of negative examples. MoCo employs two encoders: a query encoder and a key encoder, where the key encoder is updated using a momentum update with the query encoder's weights. This slow-moving momentum encoder stabilises the dictionary of negative samples, providing a diverse and reliable set of negatives for contrastive learning. The method uses a contrastive loss function similar to SimCLR but with a dynamic dictionary of negative samples that evolves over time. MoCo's ability to handle large dictionaries of negative samples efficiently has made it a critical advancement in the field, and its principles have been integrated into various subsequent models.

Bootstrap Your Own Latent (BYOL), developed by Grill et al. (2020), is notable for its departure from traditional contrastive methods that rely explicitly on negative pairs. BYOL employs two neural networks: a target network and an online network, which predict each other's outputs. The online network is trained to predict the target network's output, and the target network is updated as an exponential moving average of the online network. This approach maximises the agreement between different views of the same data point through a self-supervised learning framework, without the need for negative pairs. BYOL's effectiveness lies in its ability to prevent representation collapse through careful network architecture design and training dynamics. The method has demonstrated that competitive performance can be achieved without negative samples, marking a significant shift in contrastive learning strategies and simplifying the training process.

Swapping Assignments between Views (SwAV) is a method developed by Caron et al. (2020) that combines clustering and contrastive learning. SwAV assigns cluster labels to different augmentations of the same image and swaps these assignments to create more challenging contrasts. This approach enhances the robustness of learned representations by leveraging the relationships between cluster assignments. SwAV uses online clustering to assign each data augmentation to a prototype and then swaps these assignments between different views. This technique helps the model learn representations that are invariant to the transformations applied, capturing more meaningful features. By integrating clustering with contrastive learning, SwAV broadens the scope of self-supervised learning approaches and demonstrates significant improvements in representation quality.

SimSiam, introduced by Chen and He (2021), builds on the Siamese network architecture with the goal of simplifying contrastive learning. SimSiam eliminates the need for negative pairs or momentum encoders by focusing on comparing two augmented

views of the same image. The method uses a stop-gradient operation to prevent collapse and encourages meaningful feature extraction by maximising the agreement between different views of the same image. The loss function is designed to optimise the similarity between the two views while ensuring that the representations do not collapse to trivial solutions. By simplifying the architecture and training process, SimSiam retains effectiveness while reducing the complexity of model training. This method demonstrates that high-quality representations can be learned with a straightforward framework, contributing to the accessibility and usability of contrastive learning techniques.

Prototypical Contrastive Learning (PCL), developed by Li et al. (2020a), integrates clustering with contrastive learning by grouping similar representations into prototypes. PCL refines the learning process by contrasting these prototypes, improving the quality and robustness of the resulting representations. The method focuses on learning representations that are not only distinguishable from negative samples but also aligned with prototype clusters, enhancing the model's ability to capture underlying data structures. PCL uses a clustering algorithm to assign each data point to a prototype and then contrasts the prototypes to learn robust representations. This hybrid approach of combining clustering with contrastive learning offers a nuanced method for representation learning, leveraging the strengths of both paradigms to achieve superior performance.

Supervised Contrastive Learning (SupCon), introduced by Khosla et al. (2020), extends contrastive learning to supervised settings. In SupCon, positive pairs are formed from samples of the same class, while negative pairs are from different classes. This method leverages label information to enhance representation learning, demonstrating the applicability of contrastive principles in supervised learning environments. SupCon modifies the standard contrastive loss to incorporate class labels, allowing for the direct use of labelled data to guide the learning process. This approach has shown significant improvements in various supervised learning tasks, highlighting the versatility and effectiveness of contrastive learning beyond unsupervised settings. By integrating class labels into the contrastive learning framework, SupCon bridges the gap between supervised and unsupervised learning, providing a powerful tool for representation learning in both contexts.

Barlow Twins, introduced by Zbontar et al. (2021), focuses on redundancy reduction by minimizing the cross-correlation matrix between the representations of two views of the same image. This technique emphasises decorrelating features to capture diverse information without relying on negative pairs. Barlow Twins aims to reduce the redundancy between feature representations, promoting the extraction of distinct and informative features. The loss function used in Barlow Twins penalises the off-diagonal

elements of the cross-correlation matrix, ensuring that the learned features are as independent as possible. This method's emphasis on feature decorrelation contributes to learning more comprehensive and robust representations, which can be beneficial in various downstream tasks. The innovative focus on redundancy reduction sets Barlow Twins apart from traditional contrastive learning approaches and highlights the importance of feature diversity in representation learning.

Information Noise-Contrastive Estimation (InfoNCE) is a foundational contrastive learning objective introduced by Oord et al. (2018). InfoNCE aims to maximise mutual information between different views of the same data, serving as the basis for many contrastive learning frameworks. The InfoNCE loss encourages the model to distinguish between positive pairs (views of the same data point) and a large number of negative pairs (views of different data points), thereby learning representations that capture meaningful information. This objective function has proven to be highly effective and has influenced the design of many modern contrastive learning methods. InfoNCE's ability to measure and optimise mutual information has been instrumental in advancing the field, providing a robust and generalisable framework for contrastive representation learning.

Contrastive Language-Image Pre-Training (CLIP), developed by Radford et al. (2021), leverages large-scale datasets of images and their textual descriptions to learn joint visual and textual representations. CLIP aligns visual and textual representations by contrasting matched (positive) image-text pairs with unmatched (negative) pairs. This method demonstrates remarkable generalisation capabilities across various vision and language tasks, underscoring the versatility and power of contrastive learning in multimodal applications. CLIP uses a dual-encoder architecture, where one encoder processes images and the other processes text. The model is trained to maximise the cosine similarity between the encoded representations of matched image-text pairs while minimising it for unmatched pairs. CLIP's ability to bridge visual and textual information enables it to perform well on a wide range of tasks without task-specific fine-tuning. The success of CLIP highlights the potential of contrastive learning to integrate and leverage multiple modalities, paving the way for future research in multi-modal representation learning.

While many contrastive learning methods focus on unsupervised or self-supervised representation learning, this thesis distinguishes itself by exploring supervised representation learning specifically tailored for affective computing. Affective states, such as emotions and engagement, are inherently subjective and prone to biases, making their reliable modelling a persistent challenge. By leveraging supervised contrastive learning techniques, this research aims to mitigate these biases and enhance the robustness of

affective models. This approach enables the learning of representations that align with human-labelled affective data, ultimately improving the generalisation and reliability of emotion recognition systems.

2.2.2 | Contrastive Representations of Affect

Contrastive learning techniques are among the most widely applied methods for learning representations thus it comes as no surprise that there is growing body of research focusing on contrastive representations of affect. Notably, Li et al. (2021) investigated the impact of unsupervised representation learning on unlabelled datasets for speech emotion recognition. Their study demonstrated that the proposed contrastive predictive coding method based on InfoNCE produced representations that achieved state-of-the-art performance across the activation, valence, and dominance dimensions. This work highlighted the potential of unsupervised methods in extracting meaningful features from speech data. Specifically, their method involved generating representations from raw audio signals, which were then used to predict emotional states. The success of this approach underscores the importance of leveraging unsupervised learning techniques to handle the inherent complexities and variabilities in speech data, thereby enhancing the robustness of emotion recognition systems.

Building on these advances, Mai et al. (2022) introduced a novel hybrid contrastive learning framework. This framework integrates intra-modal, inter-modal, and semi-contrastive learning to enable the model to explore cross-modal interactions and preserve inter-class relationships, thereby reducing the modality gap. Their approach demonstrated significant improvements in capturing the complex relationships between different modalities, such as audio and visual data, in emotion recognition tasks. The hybrid framework's ability to process and integrate information from multiple sources allows for a more holistic understanding of emotional expressions, leading to more accurate and nuanced emotion recognition.

Yin et al. (2021) addressed cross-corpus emotion recognition challenges by proposing a two-step framework based on contrastive learning. Their method involved pre-training with contrastive learning within a specific domain, followed by fine-tuning in a similar domain. The results showed that this approach effectively transferred representations across different datasets, improving the robustness and generalisability of emotion recognition models. This two-step framework highlights the importance of domain adaptation in emotion recognition, demonstrating that contrastive learning can be a powerful tool for transferring knowledge between different emotional corpora and enhancing the performance of emotion recognition systems in diverse contexts. Recent

studies have begun to explore the role of supervised contrastive learning (SCL) in affective computing. For example, Yang et al. (2023) introduced a cluster-level supervised contrastive learning approach, reducing the high-dimensional representation space to three dimensions. This method achieved efficient dimensionality reduction while maintaining competitive performance, highlighting the potential of SCL for interpretable and computationally efficient emotion recognition.

Building on these developments, this thesis explores SCL for affective computing while diverging from conventional approaches that rely on predefined emotion labels. Instead of leveraging categorical annotations, it focuses on learning representations that capture the core affective dimensions. Preliminary work by the author Pinitas et al. (2022b) provided early evidence that SCL may enhance the discriminability of arousal-related features, motivating a more systematic investigation presented here. Furthermore, this thesis extends SCL by incorporating it as a pretraining strategy for teacher models within the Learning Using Privileged Information framework, enabling more effective knowledge transfer from privileged to standard modalities. In addition, it investigates few-shot contrastive representation learning methods, aiming to enhance generalisation in scenarios with limited labelled data.

2.3 | Learning Using Privileged Information

Learning Using Privileged Information is an innovative machine learning paradigm introduced by Vapnik and Vashist (2009). The fundamental idea behind LUPI is to utilise additional information available during the training phase that is not accessible during testing. This privileged information helps to improve the learning efficiency and performance of the model by providing richer data representations during the training process. This section presents notable works in this direction and applications to affect modelling.

2.3.1 | Foundational Work

As mentioned above, the LUPI paradigm was first formalised by Vapnik and Vashist (2009), who introduced the SVM+ framework. This extension of the classical Support Vector Machine (SVM) framework allows the incorporation of privileged information to define a corrective function that adjusts the decision boundary, thereby improving the model's generalisation capabilities. SVM+ takes advantage of additional, richer information available during training but not during testing, helping the model to better understand the structure of the data and improve its performance on unseen instances. This foundational work established a new direction for machine learning research, emphasising the potential of using more informative datasets to enhance the training process.

Building on the aforementioned work, subsequent research has explored various extensions and applications of the LUPI paradigm. For example, Lopez-Paz et al. (2015) unified the concepts of knowledge distillation and LUPI. They proposed a framework where a teacher model, trained with privileged information, guides a student model that lacks access to this information during training, thereby significantly enhancing the student model's learning process. This method effectively transfers knowledge from the teacher to the student, leveraging the privileged information to improve the student's performance even in its absence during testing. This approach bridges the gap between training and testing environments, providing a more robust learning process. Knowledge distillation, in itself, aims to compress the knowledge of a larger model into a smaller one, but when combined with LUPI, it gains an additional layer of efficiency by utilizing privileged data to create more informed, compact student models.

Similarly, Lapin et al. (2014) introduced adaptive mechanisms within the SVM+ framework, adjusting the influence of privileged information based on its relevance and quality. Their weighted SVM approach effectively handles noisy or less informative privileged information, ensuring that the model benefits from the additional data without being misled by inaccuracies or irrelevant details. This adaptation allows the model to selectively use privileged information, enhancing its flexibility and robustness in various scenarios. By weighting the privileged information according to its reliability, the approach prevents the degradation of model performance due to potentially misleading data, a critical step towards practical applications where data quality can vary significantly.

Moreover, Feyereisl and Aickelin (2012) extended the idea of LUPI to unsupervised learning, focusing on the use of privileged information they developed the aRi-MAX method to enhance the performance of the K-Means. They also introduced the P-Dot al-

gorithm, which uses an information-theoretic approach to combine privileged and non-privileged data for better clustering. The effectiveness of this method was demonstrated through its application to simple digit recognition tasks. Their work demonstrated enhanced performance in digit clustering applications by using privileged information to better capture the relationships between data points. Clustering methods are fundamental in many machine learning algorithms, and by incorporating privileged information into the corresponding distance function, models can produce more accurate clusters

Further extending LUPI into deep learning architectures, Sharmanska et al. (2013). explored the use of privileged information in learning-to-rank problems. Their methods demonstrated significant performance improvements in ranking tasks by incorporating additional training data that provides context or supplementary insights into the ranking criteria. This application of LUPI in deep learning highlights the paradigm's versatility and effectiveness in different types of machine learning tasks, beyond traditional classification problems. Learning-to-rank is particularly challenging because it involves ordering items based on their relevance, and privileged information can provide critical context that helps models learn more accurate ranking functions.

Shi and Kim (2017) applied LUPI to human activity recognition using deep neural networks. They showed that incorporating privileged information during model training and refinement enhances model adaptability and accuracy. This approach allows the model to continually improve and adjust to new data, making it more effective in dynamic and real-time environments. The use of privileged information in this context helps the model to better understand and predict complex human behaviours, showcasing once again LUPI's potential in practical and real-world applications. Human activity recognition often deals with diverse and ambiguous data, where privileged information can significantly enhance the model's interpretative power, leading to more accurate predictions and better adaptability to new activities.

In addition to applications in deep learning, Wang et al. (2015) extended LUPI to Bayesian models, leveraging privileged information to enhance probabilistic inference and decision-making. This integration of LUPI with Bayesian approaches contributes to more robust learning processes by incorporating additional sources of information into the probabilistic framework. This approach improves the model's ability to make accurate predictions and decisions under uncertainty, demonstrating the broad applicability of LUPI across different machine learning paradigms. Bayesian methods are particularly well-suited for handling uncertainty and incorporating prior knowledge, and the inclusion of privileged information can further refine these probabilistic models, making them more accurate and reliable.

Finally, privileged information has also been integrated into representation learning approaches. Yang et al. (2017) proposed a representation learning approach for face verification and recognition, where privileged information is used to incorporate additional information into the representation learning process. This method leverages the additional features provided during training, enhancing the model's ability to distinguish between different faces from data that is not available during testing. This application of LUPI highlights its effectiveness in handling complex data. The representation learners can capture intricate relationships between data points, and incorporating privileged information into these models can significantly improve their performance, especially in tasks that require a nuanced understanding of relationships, such as face recognition.

In summary, LUPI paradigm has significantly impacted various fields within machine learning by improving model performance through the use of additional training information. From its foundational SVM+ framework to applications in deep learning, Bayesian methods, and graph-based approaches, LUPI continues to provide valuable insights and techniques for leveraging privileged information to enhance learning processes. By integrating richer and more informative datasets during training, LUPI offers a powerful tool for developing more accurate and robust machine learning models, with broad applicability across different domains and tasks. This paradigm shift towards utilizing privileged information represents a significant advancement in machine learning, offering new possibilities for improving model performance and addressing complex challenges in various applications.

While many machine learning approaches for affective computing rely solely on conventional supervised learning, this thesis distinguishes itself by exploring Learning Using Privileged Information (LUPI) as a framework for improving affective representation learning. Affective states, such as arousal and valence, are inherently subjective and context-dependent, making them particularly challenging to model. Traditional models often struggle with incomplete or ambiguous labels, but LUPI leverages additional privileged information available during training that is not accessible at inference time. By incorporating these auxiliary signals during training, this thesis aims to enhance the robustness of affective models. Additionally, this thesis extends LUPI by integrating it with SCL further enhancing its capacity. Through this approach, the thesis advances affective computing beyond standard supervised learning paradigms, enabling more reliable affective computing systems.

2.3.2 | Applications in Affective Computing

Knowledge distillation is a technique of increasing popularity that learns a student model from a teacher model. This concept is mainly used to transfer knowledge from a large, complex teacher model to a smaller, more efficient student model (Gou et al., 2021). The primary goal is to maintain the performance of the larger model while reducing the computational resources required for deployment. Knowledge distillation typically involves training the student model to replicate the output of the teacher model, capturing the knowledge encoded in the teacher's parameters in a more compact form.

For instance, Sun et al. (2020) proposed a novel technique for micro-expression detection that distils knowledge from a pre-trained deep teacher neural network to a shallow student neural network via a teacher-student correlative framework. This approach allows the student network to learn from the rich features extracted by the teacher, achieving high accuracy with a simpler architecture. Micro-expression detection is a challenging task due to the subtle and brief nature of micro-expressions. The correlative framework leverages the detailed feature maps generated by the teacher model to guide the student model, enabling it to capture fine-grained details necessary for accurate micro-expression recognition.

Similarly, Liu et al. (2022) developed a cross-modal consistency modelling-based knowledge distillation framework for image-text sentiment classification of social media data. By aligning the information between images and text, the student model can effectively capture the sentiment expressed in multimodal social media posts. Social media data often contain rich multimodal content where images and text complement each other. The cross-modal consistency approach ensures that the student model learns to interpret and combine information from both modalities, enhancing its ability to accurately classify sentiments in diverse posts.

Additionally, Jeong et al. (2022) trained a student model on the soft labels of a teacher model for multi-task emotion recognition, demonstrating the versatility of knowledge distillation across various affective computing tasks. Multi-task learning involves training a model to perform multiple related tasks simultaneously, which can lead to better generalisation and performance. By using the soft labels generated by the teacher model, which contain more information than hard labels, the student model can learn more nuanced representations that are beneficial for recognising different emotional states.

In traditional knowledge distillation, both the teacher and the student networks have access to the same input space. However, several affective computing problems have asymmetric train and test input distributions. This discrepancy means that the

training phase might have access to richer or more diverse data than what is available during testing. For example, in emotion recognition, training data might include detailed physiological measurements or contextual information that is not present in real-time testing scenarios. Hence, it is not surprising that the Learning Using Privileged Information (LUPI) paradigm has started becoming popular in the affective computing community. LUPI allows the use of additional information during training that is not available during testing, thereby addressing the asymmetry between the training and testing phases.

Makantasis (2021) introduced a ranking model that treats additional training information as privileged information to rank affect states. This approach leverages richer data available during training, such as audio or contextual information, to improve the ranking model's accuracy in predicting affective states. By utilizing privileged information, the model can better understand the underlying patterns and variations in affective states, leading to more accurate and reliable rankings during testing when only the primary data is available.

In another study, Makantasis et al. (2021b) proposed a privileged information framework predicting arousal from gameplay footage while treating telemetry and heart rate as privileged information. By using detailed physiological and telemetry data during training, the model can better understand the nuances of player arousal, leading to more accurate predictions based solely on gameplay footage during testing. This approach demonstrates the practical application of LUPI in scenarios where additional sensors and data sources are available during training but not feasible during deployment.

Furthermore, Zhang and Etemad (2021) proposed a novel LUPI pipeline to distil EEG representations via capsule-based architectures for both classification and regression tasks. This approach utilises privileged information during training to enhance the model's ability to interpret complex EEG data, resulting in more robust and accurate performance in both classification and regression scenarios. The use of capsule-based architectures helps in capturing spatial hierarchies in the data, making the distilled knowledge more effective for the student model. EEG data are inherently noisy and complex, and privileged information such as additional sensory inputs or contextual information can provide crucial insights that improve the learning process.

This thesis investigates the use of the Learning Using Privileged Information (LUPI) paradigm for affective state prediction. The approach leverages fine-grained features and fused modalities as privileged information during training, while relying on raw audiovisual input at test time, with the aim of improving model robustness in real-world conditions. Supervised contrastive learning (SCL) is further explored as a pre-training mechanism to strengthen in-vitro models and enhance their transferability. The

methodology is evaluated on two multimodal affect datasets, providing evidence that LUPI can support more reliable affect modelling when transitioning from controlled laboratory settings to dynamic environments such as games.

2.4 | Few-Shot Learning

Few-shot learning (FSL) is a paradigm in machine learning that focuses on developing models capable of generalising from a limited number of examples. This capability is crucial in scenarios where labelled data is scarce or expensive to obtain. This section presents the most widely applied FSL approaches and moves on surveying applications in affective computing.

2.4.1 | Few-Shot Learning Approaches

Over the last years, various few-shot learning (FSL) methods have been proposed, each addressing the challenge through different mechanisms. Few-shot learning, which aims to enable models to generalise well from only a few training examples, has become popular in scenarios where labelled data is scarce. The existing FSL algorithms can be broadly classified into three main categories based on their learning strategy: *optimisation-based algorithms*, *metric learning approaches*, and *hybrid methods*.

Optimisation-based algorithms focus on training models to quickly adapt to new tasks with minimal updates. A key innovation in this area is the Model-Agnostic Meta-Learning (MAML) algorithm, introduced by Finn et al. (2017) MAML optimises for a set of initial parameters that can be fine-tuned with a few gradient steps on a small number of examples from a new task, thereby achieving rapid adaptation. This approach is notable for its model-agnostic nature, meaning it can be applied across several models without model-specific adjustments. Building on this, Nichol and Schulman (2018). proposed Reptile, which simplifies the MAML framework by aggregating gradients over multiple tasks without the need for task-specific reinitialisation. This modification reduces the computational cost while maintaining the effectiveness of fast adaptation. Another significant contribution in this category is Meta-SGD by Li et al. (2017b), which extends MAML by not only optimizing the initial model parameters but also the learning rates. This dual optimisation allows for more nuanced control over the learning process, facilitating quicker and more efficient adaptation to new tasks. Additionally, methods such as Meta-Learner LSTM proposed by Ravi and Larochelle (2016), use an LSTM-based meta-learner to generate learning updates, which effectively encodes learning strategies for rapid adaptation.

Metric learning approaches aim to map input data into a latent space where samples from the same class are closer together, and samples from different classes are farther apart. This approach is particularly useful for tasks where new classes are frequently introduced. Prototypical Networks, developed by Snell et al. (2017), exemplify this approach by learning a prototype (i.e., a central representation) for each class in the latent space. These prototypes are calculated as the mean of the embeddings of the support set examples, and classification is done by finding the closest prototype to the query example. This method ensures that the model can efficiently classify new instances based on a small number of examples. Relation Networks, proposed by Sung et al. (2018), extend the metric learning framework by learning a deep distance metric. This approach uses a neural network to model the relationship between pairs of examples, thereby learning to compare samples in a way that generalises well to new classes. Another influential method in this domain is Matching Networks by Vinyals et al. (2016), which uses an attention mechanism to weigh the importance of each support example in making predictions. This approach dynamically adjusts its focus based on the similarity between the support and query examples, providing a flexible and effective means of classification. Furthermore, Siamese Networks Koch et al. (2015) utilise twin networks to compute a similarity score between pairs of examples, making them highly effective for tasks such as one-shot recognition. Gidaris and Komodakis (2018), proposed a dynamic few-shot visual learning without forgetting that extends an object recognition system with an attention-based few-shot classification weight generator and redesigns the classifier of a ConvNet model as the cosine similarity function between feature representations and classification weight vectors.

Hybrid methods combine aspects of different FSL approaches to leverage their respective strengths. One promising area of hybrid methods is the integration of contrastive learning (CL) techniques with few-shot learning frameworks. CL focuses on learning to distinguish between similar and dissimilar examples by maximising the similarity between positive pairs and minimising the similarity between negative pairs in the embedding space. Liu et al. (2021) demonstrated the effectiveness of CL in FSL by applying noise contrastive estimation to develop a few-shot embedding model for image classification. This approach enhances the model's ability to learn discriminative features, even from a few examples. Chen et al. (2022) introduced ContrastNet, a CL-based framework designed for few-shot text classification. This framework addresses the dual challenges of learning discriminative representations and mitigating overfitting, which are common issues in few-shot learning scenarios. In the context of video analysis, Zheng et al. (2022) proposed a mixed-supervised hierarchical contrastive learning approach. This method aligns discriminative temporal clips from videos, en-

abling more effective learning of temporal patterns in a few-shot setting. Additionally, Jian et al. (2022) combined supervised contrastive learning with the standard masked language modelling loss in prompt-based few-shot learners. This hybrid approach was applied across 15 different language tasks, demonstrating significant improvements in performance by leveraging both labelled and unlabelled data effectively. Another notable hybrid approach is the Transductive Propagation Network by Liu et al. (2018), which incorporates a transductive inference mechanism to exploit the structure of the test data, significantly improving classification performance by utilising both the labelled and unlabelled samples during training.

These diverse methodologies underscore the rapid advancements and growing interest in the field of few-shot learning. As researchers continue to push the boundaries, there is a clear trend towards integrating various techniques, such as optimisation-based methods, metric learning, and contrastive learning, to develop more robust and flexible FSL systems. These approaches not only enhance the performance of models in low-data scenarios but also expand the applicability of few-shot learning to a broader range of tasks and domains, including image classification, text analysis, and video processing. The ongoing research in FSL promises to further improve the ability of AI systems to learn efficiently from limited data, making it a critical area of study for the future of machine learning.

Unlike previous studies, this thesis investigates the effectiveness of few-shot representation learning within affective computing, addressing the challenge of learning robust affective models with limited labelled data. By leveraging metric learning and contrastive learning techniques, it explores how representations can be learned more efficiently from small data samples. This approach enhances model generalisation, particularly in real-world, dynamic environments where large-scale labelled affect datasets are often scarce or impractical to obtain.

2.4.2 | Few-Shot Learning for Affective Computing

The literature on emotion recognition reveals significant advancements through the use of machine learning approaches, focusing particularly on few-shot learning, and domain adaptation. These methods have been effectively applied to emotion recognition tasks using physiological signals, speech, and facial expressions, each contributing to the broader understanding and technological capability in this field.

For Instance, Zhang et al. (2022b) proposed an innovative emotion recognition algorithm known as Deep Siamese Networks (EmoDSN), which focuses on fine-grained valence and arousal (V-A) recognition. This approach is particularly noteworthy for its

ability to operate with a minimal amount of training data, typically requiring fewer than 10 samples per class. This is achieved by maximising the distance metric between signal segments with different V-A labels, thus addressing the significant challenge of needing large amounts of annotated data. The EmoDSN algorithm was rigorously tested across different environments, including desktop, mobile, and head-mounted display-based virtual reality settings. The results were promising, showcasing the model's effectiveness even with limited training data.

Ning et al. (2021) tackled the challenges associated with the variability and non-stationarity of EEG signals through the development of Single-Source Domain Adaptive Few-Shot Learning Networks (SDA-FSL) for cross-subject EEG emotion recognition. Traditional machine learning methods often assume that training and testing sets come from the same data distribution, a condition that is rarely met in the EEG field. The proposed method, evaluated on benchmark affective computing datasets, demonstrated superior generalisation performance in cross-dataset experiments. This was achieved by designing a CBAM-based feature mapping module to extract common features and employing a domain adaptation module to align the data distribution of two domains. The incorporation of Prototypical Networks with an instance-attention mechanism further preserved domain-specific information, leading to outstanding performance.

In the field of speech-based emotion recognition, Feng and Chaspari (2021) implemented a few-shot learning approach using a metric learning technique through a Siamese neural network. This method is particularly significant for its ability to work with the limited labelled data available in ambulatory studies. By leveraging the abundance of labelled speech data from acted emotions, the approach models the relative distance between samples rather than relying on absolute patterns, which is crucial for recognising emotions from spontaneous speech. Tested on four different datasets, this method showed not only feasibility but also superior performance compared to commonly used adaptation methods, such as network fine-tuning and adversarial learning. This lays a foundation for ambulatory tracking of human emotion in spontaneous speech, which can significantly contribute to real-life mental health assessment.

Ahn et al. (2021) addressed the issue of performance degradation in cross-corpus scenarios by proposing a novel method that combines few-shot learning with unsupervised domain adaptation. This approach is designed to learn class similarity from the source domain samples adapted to the target domain without requiring any labelled samples from the target domain. By utilising multiple corpora during training, this method enhances the robustness of emotion recognition to unseen samples. Experiments conducted on emotional speech corpora in three different languages demonstrated that this method outperformed other approaches in cross-corpus emotion recognition, highlight-

ing its potential for broader applications.

Shome and Kar (2021) proposed a few-shot federated learning framework for facial expression recognition, addressing privacy concerns and dataset bias which are prevalent issues in this field of affective computing. Federated learning allows for training models collaboratively with decentralised private data on user devices, thus maintaining data privacy. Their approach utilises a few samples of labelled private facial expression data to train local models in each training round. These local models are then aggregated in a central server to form a globally optimal model. Additionally, they designed a federated learning-based self-supervised method to update the feature extractor network on unlabelled private facial data, enhancing the robustness and diversity of face representations. Tested on two benchmark datasets for facial expression recognition, the framework achieved performance comparable to state-of-the-art centralised approaches, demonstrating its efficacy and privacy-preserving capabilities.

Chen et al. (2023) developed a privacy-preserving self-supervised ViT for few-shot facial expression recognition. This system is designed to address privacy concerns, inadequate data transfer, and class imbalance in facial expression recognition. The method integrates self-supervised learning (SSL) and few-shot learning (FSL) to train a deep learning model with fewer labelled samples. Specifically, the Vision Transformer (ViT) is pre-trained with four self-supervised pretext tasks—image denoising and reconstruction, image rotation prediction, jigsaw puzzle, and masked patch prediction—to obtain a pretrained ViT encoder. This encoder is then fine-tuned on a labelled FER dataset to extract spatio-temporal features and implement the FER task. The extensive experimental results across 4 datasets show high recognition accuracies, demonstrating the model's effectiveness and robustness.

Ciubotaru et al. (2019) focused on the generalisation ability of low-shot learning methods for facial expression recognition. Their study revisited and compared existing few-shot learning methods, particularly in terms of their generalisation ability via episode-training. The challenge of recognising novel classes with limited data and significant variations within the same semantic category was addressed through extensive experiments, demonstrating the efficacy of low-shot learning methods in enhancing the model's performance and generalisation.

Zou et al. (2022) addressed the challenge of recognising compound facial expressions by proposing an Emotion Guided Similarity Network (EGS-Net) in a cross-domain few-shot learning setting. This method aims to identify unseen compound expressions with a model trained on easily accessible basic expression datasets. The EGS-Net consists of an emotion branch and a similarity branch, which are jointly trained in a multi-task fashion. The regularisation provided by the emotion branch prevents overfitting, while

a two-stage learning framework further improves the inference ability of the similarity branch on unseen compound expressions. Experimental results on both in-the-lab and in-the-wild compound expression datasets demonstrated the superiority of this approach against several state-of-the-art methods.

Finally, Zhang et al. (2022a) contributed significantly to the field of autism diagnosis by developing a few-shot learning method for analysing hour-long Autism Diagnostic Observation Schedule (ADOS) videos. This method addresses the complexity and length of ADOS videos by leveraging well-established computer vision tools for spatio-temporal feature extraction and marginal fisher analysis. The approach also incorporates few-shot learning and scene-level fusion strategies to classify individuals into categories such as Autism, Autism Spectrum, and Non-Spectrum. Achieving high accuracy in the automatic classification of Autism Spectrum Diagnostic (ASD) traits and the benefits of integrating scene-level fusion strategies.

This thesis investigates few-shot learning as a framework for affect and engagement modelling by reformulating classification as a set of domain-specific few-shot tasks, aiming to improve generalisation in data-scarce scenarios. To enable this study, it introduces GVFS (GameVibe Few-Shot), a dataset designed for few-shot evaluation in viewer engagement modelling within video games. A comparative analysis of few-shot methods, including metric-based and contrastive approaches, benchmarks their performance against conventional affect modelling techniques, providing evidence of their potential for domain generalisation. Experiments with multiple pretrained backbones and fine-grained features suggest that few-shot learners can outperform standard baselines in selected classification settings. In addition, this thesis explores a new objective, Silhouette Distance loss, inspired by silhouette scores, as a means to improve intra-class cohesion and inter-class separation in representation learning.

2.5 | Summary

This chapter presented the essential background and notable literature. The review of affect modelling showcased the evolution from traditional unimodal approaches, reliant on manual feature extraction, to contemporary multimodal frameworks powered by deep learning. Multimodal techniques were highlighted as essential for capturing the complexity of emotional states, integrating visual, auditory, and physiological data to achieve more nuanced affect predictions. In parallel, the investigation of engagement and player modelling illustrated the application of these technologies to interactive domains, offering insights into user behaviour and emotional responses in dynamic envi-

ronments. Contrastive learning emerged as a versatile tool, by leveraging contrastive objectives, methodologies like SimCLR, MoCo, and BYOL have enabled the learning of robust representations with limited labelled data. Applications in affective computing revealed the power of contrastive learning to improve model generalisability, particularly for multimodal and cross-corpus scenarios. The LUPI paradigm introduced innovative ways to incorporate additional information during training, enhancing model performance without requiring privileged data during testing. Applications ranged from emotion recognition tasks utilizing physiological signals to gameplay analysis and EEG-based affect modelling, all benefiting from the added layer of contextual understanding provided by privileged data. Finally, FSL methodologies were reviewed as a solution to data scarcity in affective computing. Techniques like Siamese networks, Prototypical Networks, and hybrid approaches have proven effective for tasks with minimal labelled data, such as facial expression recognition and autism diagnostics. By integrating concepts like domain adaptation and cross-modal learning, these methods address challenges in generalisation across diverse datasets and scenarios.

3

Methodology

This chapter presents the methodological framework applied throughout this thesis. The primary goal of these methods is to derive representations that are both expressive and generalisable, enabling them to capture complex structures in data and adapt effectively to new, unseen tasks. The methodologies explored here are crucial for advancing affect modelling via representation learning, where understanding and predicting human emotions and behaviours rely on the quality of the learned embeddings.

We first introduce the general principles of representation learning (Section 3.1). This includes the core objectives of learning embeddings that are useful for downstream tasks such as classification. Following this, we focus on contrastive representation learning (Section 3.2), a highly effective approach that utilises pairs of similar and dissimilar examples to form a latent space that consist of discriminative representations.

Next, we focus into learning using privileged information (Section 3.3), where auxiliary data, available only during training, is employed to accelerate and improve the learning process. This methodology leverages additional knowledge sources during training, resulting in models that are more efficient and robust. Finally, we explore few-shot representation learning (Section 3.4), a paradigm designed to address the challenge of generalising from a small number of labelled examples. Few-shot learning relies heavily on the quality of the learned representations and demonstrates the importance of embedding spaces that can adapt to new tasks with minimal supervision.

3.1 | General Concepts of Representation Learning

Representation learning is a branch of machine learning focused on discovering useful transformations of raw data into a more structured and compact form, enabling models to perform tasks more effectively Bengio et al. (2013). Rather than relying on manu-

ally engineered features, representation learning allows models to automatically learn the most salient aspects of the data, which are crucial for making accurate predictions. These learned representations, or embeddings, are designed to capture the underlying patterns in the data, facilitating tasks such as classification, clustering, and regression.

The main goal of representation learning is to enhance a model's ability to generalise to unseen data. In traditional machine learning workflows, domain experts manually design features, which is often labour-intensive and may overlook important but subtle patterns in the data. Representation learning automates this process, allowing models to discover and encode relevant features that lead to more efficient learning and improved generalisation. Effective representations should be robust to variations and noise, retain essential information, and support transfer learning, where the learned knowledge is applied to related tasks.

There are several approaches to representation learning, each suited to different types of data and tasks. Supervised representation learning uses labelled data to guide the learning process, enabling models to learn features that are explicitly optimised for the task at hand, such as object recognition in images. In contrast, unsupervised learning methods, such as autoencoders and clustering algorithms, do not rely on labelled data. Instead, they seek to capture the inherent structure of the data by minimising reconstruction error or grouping similar examples together. Additionally, self-supervised learning has gained popularity as an approach that leverages the data itself to create learning signals. This technique often involves predicting missing parts of data or contrasting different transformations of the same input, enabling models to learn meaningful representations without the need for labelled data.

Evaluating the quality of learned representations is a key aspect of representation learning. A good representation should exhibit strong discriminative power, meaning it should clearly distinguish between different classes or categories in the latent space (Chen et al., 2020a). Another important criterion is transferability—the ability of a learned representation to generalise to related tasks or domains. Models that produce representations with high transferability can be fine-tuned or adapted for new tasks with minimal additional training (Bengio et al., 2013). Finally, compactness is a desirable property (Goodfellow, 2016), as effective representations often result in lower-dimensional embeddings that capture the most relevant aspects of the data while discarding noise and redundant information. Representation learning has found success across a wide range of applications, from image recognition and natural language processing (Le-Khac et al., 2020) to more specialised domains such as affective computing (Mai et al., 2021; Pinitas et al., 2022b; Yin et al., 2021), where the goal is to infer emotional states from physiological signals or facial expressions. In such tasks, the ability to extract

meaningful features from complex, noisy data is critical for the model's performance.

Despite its successes, representation learning presents several challenges. One major challenge is data efficiency. Many deep learning models require large amounts of data –that may include labels– to learn good representations, which can be a significant obstacle in domains where data is scarce or expensive to obtain. Techniques like few-shot learning and data augmentation aim to address this limitation by enabling models to generalise from a small number of examples Jian et al. (2022); Liu et al. (2021). Another challenge is generalisation: ensuring that representations learned from one dataset or task can be applied effectively to new, unseen data. While transfer learning and domain adaptation offer potential solutions, questions remain about how best to transfer knowledge across different tasks. Finally, there is the issue of interpretability. Deep representations, especially those learned by complex neural networks, are often opaque and difficult to understand, raising concerns in critical fields such as healthcare or autonomous systems, where interpretability and transparency are essential for trust and accountability.

In this thesis, we focus on contrastive learning as the primary approach for representation learning due to its demonstrated effectiveness in producing discriminative, transferable, and compact representations—qualities that align with our research objectives. Unlike autoencoders, which focus on reconstructing the input data in its entirety, contrastive learning emphasises learning representations by distinguishing between similar (positive) and dissimilar (negative) samples in the latent space. This discriminative nature of contrastive learning makes it particularly suitable for tasks where clear class separation is critical, such as affect modelling, as it ensures that representations capture meaningful differences while discarding irrelevant or noisy information.. While autoencoders excel in capturing holistic information through reconstruction, they may encode redundant or task-irrelevant features that can hinder performance, especially in downstream tasks requiring high-level abstractions. By contrast, contrastive learning naturally aligns with our goal of learning robust representations, making it the more appropriate choice for this work.

3.2 | Contrastive Representation Learning

As mentioned above, contrastive representation learning is a technique that aims to learn meaningful embeddings by contrasting pairs of examples. The fundamental principle is to bring similar examples– data samples that have similar attributes such as transformations of the same image –closer in the learned embedding space while push-

ing dissimilar examples (e.g., images of different objects) further apart. This approach relies on designing an effective loss function that guides the model to create a well-structured latent space, where meaningful similarities between data points are captured and preserved. In contrastive learning, the similarity between data points can be defined either through explicit labels, in the case of supervised contrastive learning, or by relying on augmented views of the same data, in unsupervised/ self-supervised contrastive learning. Both approaches follow the same core idea but differ in how positive and negative pairs are constructed.

3.2.1 | Supervised Contrastive Learning

pSupervised contrastive learning builds upon *unsupervised contrastive learning* by incorporating label information to construct sample pairs. While unsupervised contrastive learning focuses on discovering the intrinsic structure of data without labels, supervised contrastive learning explicitly leverages label information to guide the formation of positive and negative pairs. In this framework, for each data point x_i (referred to as the *anchor*), all other data points within the same class are treated as *positives*, while data points from different classes serve as *negatives* (Khosla et al., 2020). This approach encourages the model to structure the embedding space such that representations of samples with the same label are closer together, and those with different labels are farther apart. Compared to traditional supervised learning, which minimises classification error without explicitly organizing the embedding space, supervised contrastive learning fosters a more structured and discriminative representation. Mathematically, the supervised contrastive loss L_{SC} can be written as:

$$L_{SC} = \sum_{i \in I} \frac{-1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(r_i \cdot r_p / \tau)}{\sum_{a \in A_i} \exp(r_i \cdot r_a / \tau)} \tag{3.1}$$

where I is a set that includes all samples and P_i is the set that includes only the samples that are assigned to the same class as i. A_i is a set that contains any element of set I besides element i. With r_i, r_p and r_a we denote the latent representations of the model for the samples i, p and a, respectively. Finally, τ stands for a non-negative temperature hyperparameter that controls the sharpness of the similarity distribution. It should be noted that i, p and a correspond to the index of the current sample, a sample positive to the current sample and a sample different from the current one, respectively.

The supervised contrastive loss works by ensuring that all examples from the same class are clustered together in the embedding space. This is achieved by maximising the similarity between the anchor x_i and all its positive examples $P(i)$ while minimising

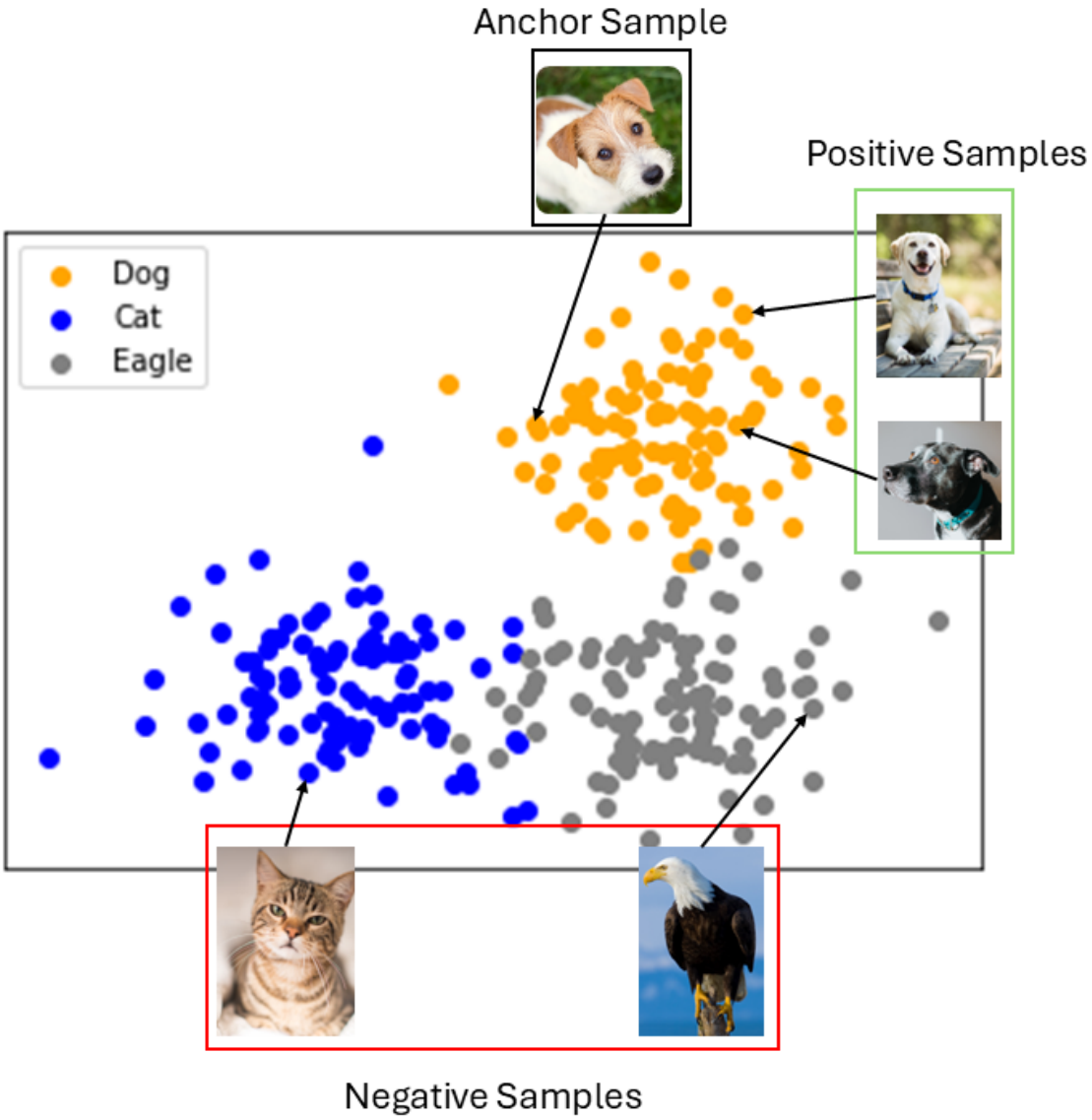


Figure 3.1: Illustration of the Supervised Contrastive Learning (SC) process. The "Anchor Sample" (dog) is paired with "Positive Samples" (other dogs) from the same class, shown in orange, and "Negative Samples" (cat and eagle) from different classes, represented by blue and gray. SC learning maximises the similarity between the anchor and positive samples while decreasing the similarity between the anchor and negative samples, helping create distinct class boundaries.

the similarity between the anchor and all other examples in the batch. The dot products $r_i \cdot r_p$ and $r_i \cdot r_a$ measure the similarity between the anchor and positive examples and the anchor and all the samples, respectively (higher values indicating greater similarity).

The loss is essentially a softmax function that normalises the positive pair similarity by comparing it with the similarity between the anchor and all examples in the batch. The temperature parameter τ plays a crucial role in controlling the spread of the softmax distribution.

3.2.1.1 | Relationship between Supervised Contrastive and Cross-Entropy Loss

Supervised contrastive learning offers a viable alternative to the conventional cross-entropy loss in classification tasks. The cross-entropy loss functions by directly minimising the error between the predicted probability distribution and the true class labels. Specifically, it seeks to maximise the likelihood of the correct class by assigning high probability to the true label and penalising incorrect predictions. While this approach is highly effective for training classification models, its focus is strictly on correct class assignment, which means it does not impose any constraints on how the learned embeddings are structured in the latent space. As a result, models trained with cross-entropy loss often produce embeddings where instances of the same class are not necessarily grouped together, and instances of different classes may not be adequately separated. Hence, while the model may achieve high classification accuracy, the underlying representation space can suffer from poor generalisation, particularly in unseen tasks or domains. By contrast, the supervised contrastive loss explicitly encourages the creation of a well-structured embedding space, where embeddings of examples from the same class are pulled closer together and embeddings of examples from different classes are pushed apart Khosla et al. (2020). This results in more discriminative and transferable representations, which can subsequently be fine-tuned for various downstream tasks. Specifically, after training with supervised contrastive loss, a simple linear classifier can be applied to the learned embeddings to achieve high performance on classification tasks.

3.3 | Learning Using Privileged Information

Learning using privileged information (LUPI) is an advanced machine learning framework that enhances the learning process by incorporating additional information during training that is not available at test time Vapnik and Vashist (2009). This privileged information can take various forms, such as expert knowledge, additional features, or a more detailed view of the data, and is used to guide the learning process to achieve better performance. The key insight behind LUPI is that the additional information available during training can act as a teacher, providing the learner, also called stu-

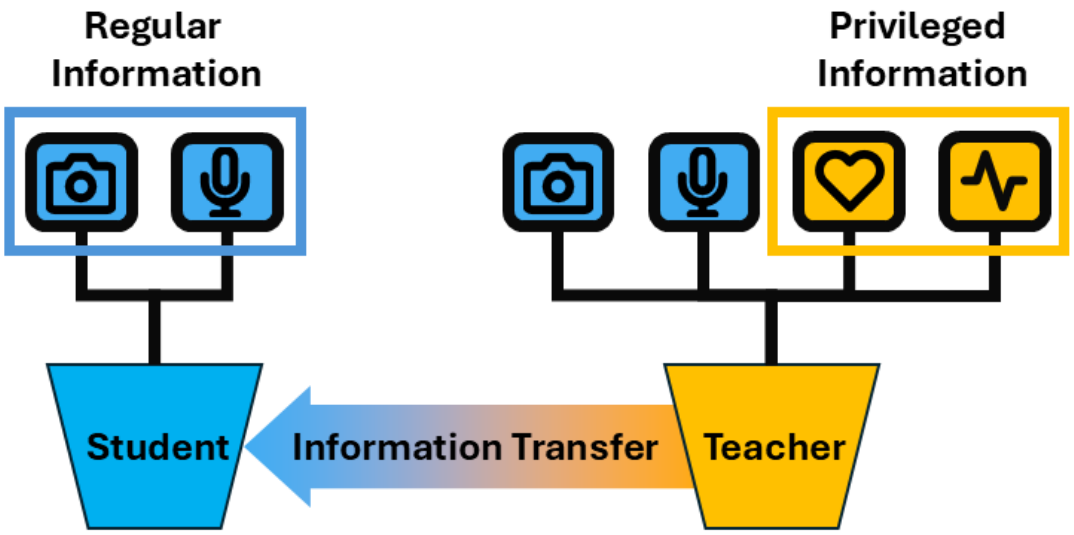


Figure 3.2: Illustration of the LUPI concept, where a teacher model has access to additional privileged information (such as emotional or physiological data, denoted by the heart and waveform icons) that the student model (which only has regular inputs, such as visual and audio data) does not have access to. The teacher uses this privileged information to enhance the learning process and transfers this knowledge to the student model.

dent, with insights that improve it's predictive power, even when this information is no longer available at inference time (Figure 3.2). Hence it becomes evident that the teacher model plays a critical role in the LUPI framework since it is the only model with direct access to privileged information.

Representation learning is fundamental in improving the performance of both teacher and student, as it focuses on creating meaningful and compact embeddings that capture the structure of the data. By ensuring that the teacher model learns high-quality representations from privileged information, the student model can inherit the knowledge of those rich embeddings, thus improving its ability to generalise and perform effectively on unseen data, even without access to the privileged inputs. In this way, representation learning not only enhances the discriminative power of the teacher but also ensures that the student receives better guidance during the learning process.

3.3.1 | Mathematical Formulation of LUPI

In the following section, we explore how knowledge from privileged information can be transferred to a machine learning model. It is important to distinguish this process from the more common techniques of transfer learning employed in deep learning Ng et al. (2015). Transfer learning is primarily used to address challenges in small-sample settings by fine-tuning a pre-trained model on a specific task to perform effectively on a related but distinct task. The key idea is to leverage prior knowledge from a model trained on large datasets to improve performance when data for the target task is limited.

In contrast, the Learning Using Privileged Information (LUPI) framework addresses a different type of problem: one involving an asymmetry in the availability of information between the training and testing phases. Instead of focusing on pre-trained models or small datasets, LUPI utilises additional information—known as privileged information—that is available exclusively during training but not at test time. The model is trained from the ground up using this additional information, with the goal of enhancing its ability to generalise to unseen data, even when that privileged information is absent. This fundamental difference makes LUPI a unique approach, distinct from traditional transfer learning, as it explicitly tackles the challenge of learning under asymmetric distributions of data between the training and deployment phases.

Consider a typical supervised learning setup where we have training data $D_{train} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ represents the input data, and $y_i \in Y$ represents the corresponding labels. In the LUPI framework, we are additionally provided with privileged information $\tilde{x}_i \in \mathbb{R}^p$, which is only available during training. The goal is to use the triplet $(x_i, \tilde{x}_i, y_i)$ to train a model, such that at test time, the model relies only on the input x_i. Let $f(x_i)$ denote the primary learning model (student), which maps the input data x_i to an output space Y. Additionally, let $g(\tilde{x}_i)$ represent the secondary model (teacher) that incorporates the privileged information $\tilde{x}$ during training. The objective is to train $f(x)$ such that it generalises well on unseen data, even though $\tilde{x}$ is not available during testing. The general form of the optimisation objective can be expressed as:

$$L_P = (1 - \alpha) L_D(f(x_i), y_i) + \alpha L_{PI}(f(x_i), g(\tilde{x}_i)) \tag{3.2}$$

, where L_D is the loss that optimises for performance on the downstream class (e.g. cross-entropy for classification) and L_{PI}, is the loss that facilitates the transfer of information between the teacher and the student. Furthermore, α is a hyperparameter that controls the influence of L_{PI} towards the general optimisation direction. Following the

principles mentioned above, in this thesis we considered the approach of probability distribution distillation.

3.3.1.1 | Probability Distribution Distillation

This privileged information distillation approach involves transferring privileged information in the form of probability distribution. In particular, the distribution of the probabilistic predictions of the teacher model $g(\tilde{x})$ are transferred to those of the student model $f(x)$. Consequently, eq 3.2 can be written as:

$$L_P = (1-\alpha)L_{CE}(f(x), y) + \alpha L_{KL}(g(\tilde{x}), f(x)) \tag{3.3}$$

where L_{CE} is the conventional cross-entropy loss and L_{KL} is the Kullback-Leibler (KL) divergence, a measure of how one probability distribution (e.g. student model) differs from a second, reference probability distribution (e.g. teacher model). It is often used in machine learning and information theory to quantify how much information is lost when using an approximation instead of the true distribution. As mentioned, this formulation leverages privileged information by using the probabilistic predictions of the teacher as a soft target to shape the learning process of the student. By providing confidence scores, the teacher helps the student to understand which examples are harder to learn, thereby guiding its training process more effectively. This can help the student model to generalise better by aligning its probability distribution with the more informed teacher model. Hence, the student model can indirectly have access to privileged information during training but not during inference. It should be noted that $L_D = L_{CE}$ is set for the downstream task, which is a classification task in this thesis. This setup ensures that the student model is guided both by the hard labels through cross-entropy and by the additional information from the teacher via KL divergence, enhancing its performance.

3.4 | Few-Shot Representation Learning

Few-shot representation learning is a subfield of machine learning focused on training models that can generalise from very few labelled examples (Wang et al., 2020) thus adapting effectively to new domains. This is in contrast to traditional machine learning paradigms, which require large amounts of labelled data to achieve high levels of performance. The primary objective in few-shot learning is to leverage prior knowledge, learned representations, or auxiliary techniques to achieve high performance despite limited supervision. Few-shot learning (FSL) is especially useful in scenarios where

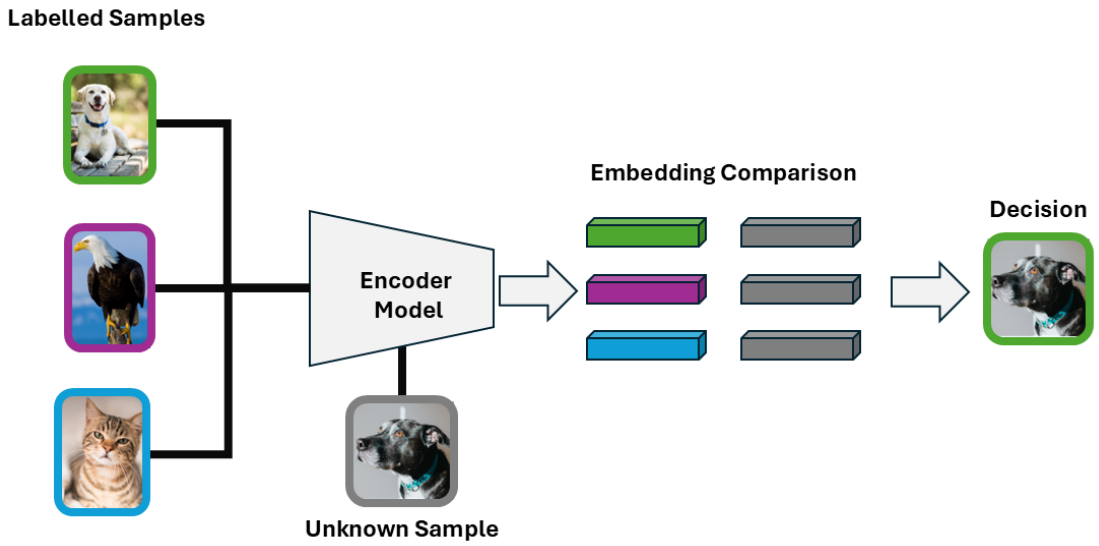


Figure 3.3: Illustration of the FSL process. The encoder model generates embeddings from a set of labelled samples, representing different categories (dog, eagle, cat). When presented with an unknown sample (a new dog image), the encoder extracts its embedding, which is then compared to the embeddings of the labelled samples. Based on these comparisons, the system makes a decision and classifies the unknown sample as a dog.

gathering large labelled datasets is costly or impractical, such as in medical diagnosis Cai et al. (2020), rare event detection Wang et al. (2022), or affective computingZou et al. (2022) .

In the general representation learning context, the goal is to encode raw input data into a latent space where similar examples are close together and dissimilar examples are far apart. This latent space serves as a compressed, more informative view of the data, making it easier to solve downstream tasks. However in FSL, the quality of this latent space becomes even more crucial because the model must generalise from very few examples. If the learned representations are sufficiently rich and discriminative, the model can leverage these embeddings to quickly adapt to new tasks with limited supervision. Thus, few-shot learning depends heavily on pre-training robust representation spaces, where embeddings can be reused across different tasks and domains.

3.4.1 | Mathematical Formulation of Few-Shot Learning

A common approach to few-shot learning involves metric-based learning, where the model learns a metric space that allows for efficient comparison between examples. In

this setup, the model learns a function $f(x;\theta)$, parametrised by θ, which maps input data $x \in \mathbb{R}^d$ to a latent space $z = f_\theta(x) \in \mathbb{R}^p$. The objective is to ensure that in this latent space, data points from the same class are clustered closely, while data points from different classes are well separated.

In a typical few-shot classification problem, given a small support set $S = \{(x_i, y_i)\}_{i=1}^k$, where x_i are the inputs and y_i are the labels, the model must classify a query point x_q by comparing it to the support set. Consequently, the support set consists of a small number of labelled examples that act as the model's reference for learning, while the query point is an unseen, unlabelled sample that needs to be classified. A popular approach is to compute the distance between the query point and each support point in the latent space and classify x_q based on the closest support point. This can be written as:

$$\hat{y}_q = \arg\min_{y_i \in S} d(f_\theta(x_q), f_\theta(x_i)), \tag{3.4}$$

where $d(\cdot,\cdot)$ is a distance metric, such as the Euclidean distance or cosine similarity. In this formulation, the success of the model depends on the quality of the learned representation $f_\theta(x)$, which should ensure that examples from the same class are close in the latent space.

3.4.1.1 | Prototypical Network Loss

The Prototypical The Prototypical Network (PN) loss was first introduced by Snell et al. Snell et al. (2017). Using that loss, the authors aim to learn a metric space in which classification can be performed by computing the distances between the query samples and the prototypes derived for each of the N classes within the support set. The PN loss is defined as

$$L_{PN} = -\frac{1}{|Q|} \sum_{(x_i^q, y_i^q) \in Q} \sum_{n=1}^{N_q} \mathbb{I}(y_i^q, n) \log(p_\theta(y_i^q = n | x_i^q)), \tag{3.5}$$

where $p_\theta(y_i^q = n|x_i^q) = softmax(-d(f_\theta(x_i^q), c_n^s))$ is the probability of a query sample x_i^q to fall into the class n, $c_n^s = \frac{1}{|S_n|} \sum_{(x_i^s, y_i^s) \in S} (y_i^s = n) f_\theta(x_i^s)$ is the prototype of class n, $d(\cdot)$ corresponds to the Euclidean distance, $f_\theta(\cdot)$ represents the learnable embedding functions, $|S_n|$ is the number of samples of class n in the support set, and $|Q|$ is the cardinality of the query set.

3.4.1.2 | Matching Network Loss

Similarly to PN, the Matching Network (MN) Loss Vinyals et al. (2016) aims to map an unlabelled example to a latent space defined by a small labelled set, enabling adaptation

to new classes without fine-tuning. The optimisation objective uses a simple attention mechanism to weight sample distances between the support and query sets. The MN loss is defined as

$$L_{MN} = -\frac{1}{|Q|}\sum_{i=1}^{|Q|}\log p_\theta(y_i|x_i, S) \tag{3.6}$$

where $|Q|$ is the number of query samples, x_i is the i-th query sample, y_i is the true label of the i-th query sample, and S is the support set. Additionally $p_\theta(y_i|x_i, S) = \sum_{(x_j,y_j)\in S} a(x_i, x_j)\cdot \mathbb{I}(y_j = y_i)$ where $a(x_i, x_j)$ is the attention mechanism defined as $a(x_i, x_j) = softmax(f_\theta(x_i)\cdot f_\theta(x_j))$ which is the probability of a query sample x_i to be similar with the support sample x_j when projected on the space $f_\theta(\cdot)$ defined by the support set S.

3.4.1.3 | Supervised Contrastive Loss

The main objective of the Supervised Contrastive (SC) loss is to generate representations that increase the similarity between samples with the same label (positive pairs) while decreasing the similarity between samples with different labels (negative pairs). Minimising this loss function leads to distinct representations for each class. Building on earlier research in few-shot representation learning for image classification Liu et al. (2021), we define SC as follows:

$$L_{SC} = \frac{1}{|S|}\sum_{s\in S}\frac{-1}{|P_s^q|}\sum_{p\in P_s^q}\log\frac{\exp(r_s\cdot r_p/\tau)}{\sum_{q\in Q}\exp(r_s\cdot r_q/\tau)}, \tag{3.7}$$

In this formulation, S represents the support set, while P_s^q refers to the subset of query samples that are assigned to the same class as s, with $q \in Q$ indicating any element from the query set. The latent representations, denoted by r_s, r_p, and r_q, are generated by applying the function f_θ to the samples x_s, x_p, and x_q, respectively. The parameter τ is a non-negative temperature hyperparameter that modulates the distribution of similarity between representations. Lastly, s, p, and q correspond to the indices of the current support set sample, a query sample that is positive to the current support sample, and any sample from the query set, respectively. As can be discerned from Eq. 3.1, SC focuses on creating dissimilar class representations since positive pairs are present in both the numerator and denominator of the loss. However, it does not account for compactness within class representations. This limitation can lead to scattered intra-class representations, which may impact the overall structural quality of the latent space.

3.4.1.4 | Silhouette Distance Loss

In a supervised context, the Silhouette score is adapted to encourage representations that not only consider the separability of classes but also the cohesion within each class. The *Silhouette Distance* (SD) loss, proposed as part of this thesis, evaluates the quality of clustering patterns based on labels by measuring the average Euclidean distance within the same class (intra-class cohesion) and the average nearest-class Euclidean distance (inter-class separation) (Pinitas et al., 2024b). The term “Silhouette Distance" was chosen because the loss function draws inspiration from the silhouette score, providing an intuitive understanding of its operation. While it does not meet the criteria to be classified as a metric distance, it can be considered a non-metric distance, similar to cosine distance. The minimisation of this loss yields representations that consider both the separability of classes and the compactness within each class. Formally, the SD loss is defined as:

$$L_{SD} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1 - Sil(q, S)}{2}, \tag{3.8}$$

where

$$Sil(q, S) = \frac{b(q, S) - a(q, S)}{max_{\delta}\{a(q, S), b(q, S)\}} \tag{3.9}$$

denotes the silhouette score. Given a query sample (x_q, y_q), $a(q, S)$ corresponds to the average distance between the representation of x_q and the representations of all the support set samples x_s that have label $y_s = y_q$. Furthermore, $b(q, S)$ is the average Euclidean distance between the representation of x_q and the nearest support set class. $max_{\delta}\{a(\cdot), b(\cdot)\} = max\{a(\cdot), \delta\}$ when $b(\cdot) \leq a(\cdot)$ and $max_{\delta}\{a(\cdot), b(\cdot)\} = b(\cdot)$ when $b(\cdot) > a(\cdot)$ ensuring that $S(\cdot)$ is differentiable for all samples. δ is a small scalar value guaranteeing numerical stability when $a(\cdot) = b(\cdot) = 0$.

3.5 | Summary

This chapter provided an overview of the methodologies employed in this thesis. These methodologies aim to create informative representations that can effectively capture complex patterns in data. In particular, the general principles of representation learning, which emphasise learning embeddings for downstream tasks like classification were outlined. Contrastive representation learning was then introduced as an effective method for constructing a latent space with discriminative features by leveraging pairs of similar and dissimilar examples. Moreover, this chapter delved into the concept of learning using privileged information (LUPI), where auxiliary data, available

only during training, is used to accelerate and refine the learning process and discussed the impact of the teacher models in the learning process. Finally, few-shot learning was explored as a solution to the challenge of generalising from a limited number of labelled examples, highlighting the importance of well-structured embeddings that allow models to adapt to new tasks with minimal supervision

4

Affect Corpora

This section provides an overview of the datasets utilised throughout this thesis. The experiments conducted in the subsequent chapters rely on two widely recognised and extensively used datasets in the field of affective computing. These datasets offer rich, multimodal information and have been established as benchmarks for studying affective states, facilitating the evaluation of various computational models in the field. Their inclusion ensures that the methodologies proposed in this thesis are tested in a robust and reliable manner, allowing for meaningful insights into their effectiveness and generalisation capabilities. In addition to these established datasets, a new dataset focusing on viewer engagement during video game interactions was introduced specifically for the purposes of this thesis. This dataset plays a critical role in benchmarking the performance of few-shot learning models, which are central to the methodological contributions of this work. By incorporating this novel dataset, this thesis addresses the challenge of limited training data and provides a framework for evaluating models that must learn effectively with minimal labelled samples. The rest of this chapter is organised as follows. In section 4.1 we provide details about the RECOLA database, a dataset of dyadic interactions between french-speaking participants. Section 4.2 surveys the three platformer games of the AGAIN dataset which features gameplay videos & features along with self-reported annotations of player arousal. Finally section 4.3 introduces the GameVibe corpus, a dataset tailored for the task of domain generalisation within the fields of Affective Computing and Digital Games. This corpus consists of 30 diverse FPS games annotated in terms of viewer engagement.

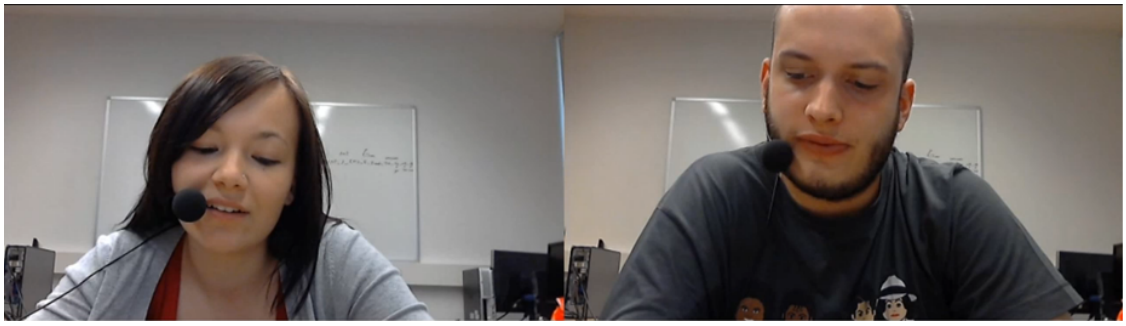


Figure 4.1: Illustration showing two samples from the RECOLA dataset, featuring a male and a female participant conversing while sitting in front of a webcam. Both participants are positioned in a controlled laboratory environment, ensuring consistency in recording conditions. Their facial expressions, body language, and vocal cues are captured for the purpose of emotional and behavioural analysis. The dataset focuses on real-time audiovisual recordings to study and evaluate emotional responses in social interactions.

4.1 | The RECOLA Database

The RECOLA multimodal database (Ringeval et al., 2013) was developed to facilitate the analysis and modelling of spontaneous human behaviours in computer-mediated communication settings (see Figure 4.1). As a rich and highly challenging resource, it comprises 9.5 hours of synchronised multimodal data, including audio, visual, and physiological signals (electrocardiogram and electrodermal activity). The dataset captures online dyadic interactions between 46 French-speaking participants, specifically designed to study spontaneous communication dynamics. Its multimodal nature enables the exploration of human affect and behaviour in a complex, naturalistic environment.

To support researchers in effectively utilising the dataset, the creators have provided high-level feature sets for each modality. The visual component includes 40 features, such as facial action units, head pose estimations, and optical flow features that capture subtle movement and expression changes. The audio modality is represented by 130 features, closely aligned with those used in the COMPARE challenge dataset Schuller et al. (2014), including descriptors of voice intensity, pitch, and mel-frequency spectral coefficients. Physiological signals, represented by 116 features, focus on features extracted by electrocardiogram and electrodermal activity

A notable aspect of the dataset is the annotation of arousal and valence, key components of affective states, which were assessed by six expert annotators, 3 male and 3 female (see Figure 4.2). These annotations, continuous and bounded between the val-

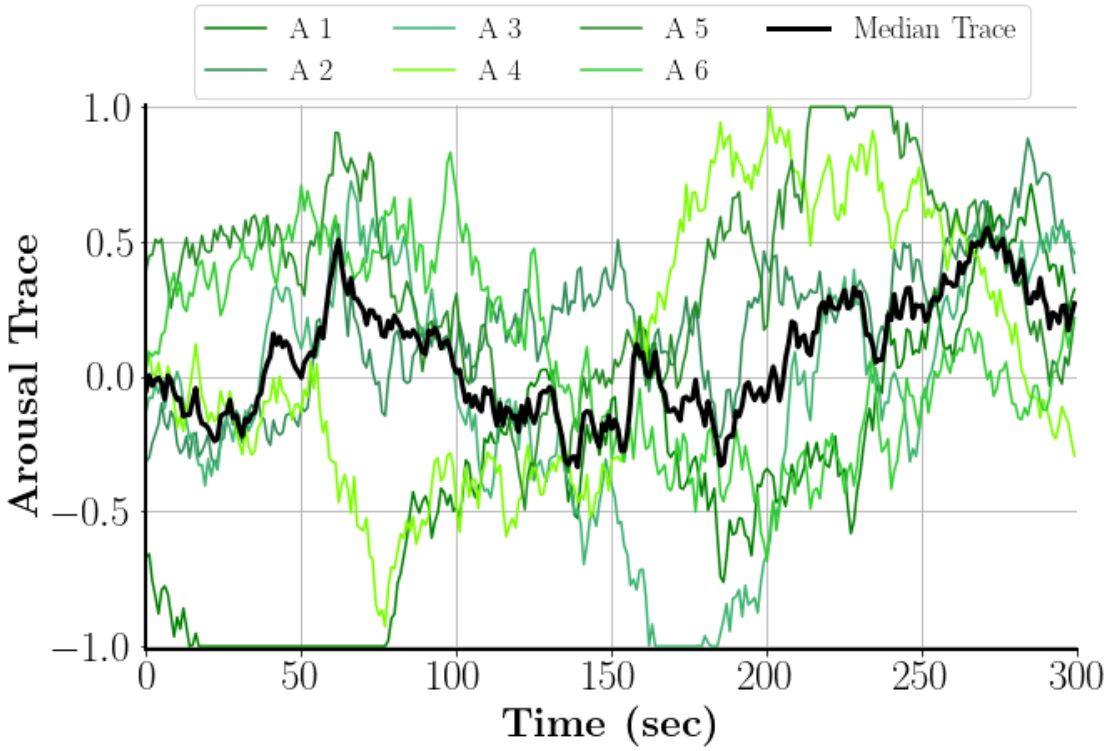


Figure 4.2: Illustration of arousal annotation traces for a randomly selected participant from the RECOLA dataset. The six individual traces (A1-A6) correspond to annotations made by six expert annotators, each evaluating the participant's level of arousal over time. These traces reflect fluctuations in emotional intensity during the recorded interaction. For improved visual clarity, a black trace is overlaid, representing the median value of all annotations.

ues of [-1, 1], are provided at a temporal resolution of 25Hz, allowing for a fine-grained analysis of affective dynamics over time. This comprehensive annotation schema offers a detailed perspective on the temporal fluctuations of emotional states during interaction.

Additionally, beyond the extracted features, raw data from all modalities are also included in the database, ensuring that researchers can reprocess or extract new features as needed. While the dataset initially includes recordings from 46 participants, only 34 gave explicit consent for their data to be shared. Of those, the public dataset includes data from 23 participants, with the remaining 11 withheld for use in evaluation sets in prior competitions, ensuring a consistent benchmark for comparative research. Table 4.1 summarises the key elements of the dataset.

Table 4.1: Overview of the key characteristics of the RECOLA corpus, including participant details, video information, and annotation attributes

Number of Participants	23
Number of Annotators	6
Number of Videos	23
Video database size	1 hour 55 minutes
Video duration	5 minutes
Annotation Perspective	Third-person
Annotation Type	Continuous bounded
Affect Labels	Arousal, Valence

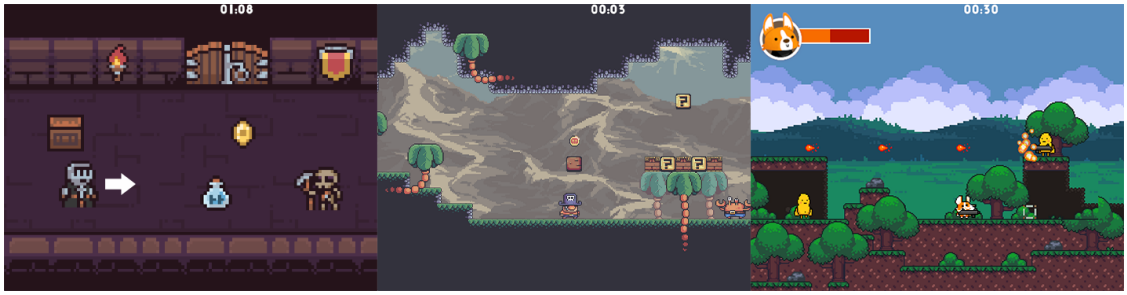


Figure 4.3: Illustration of the three AGAIN games used in this study. From left to right, the games depicted are *Endless*, *Pirates!*, and *Run'N'Gun!*. Each game offers distinct gameplay mechanics and challenges, providing a diverse set of environments for player behaviour and experience analysis.

Table 4.2: Key characteristics of the platformer games in the AGAIN corpus, detailing participant numbers, video data, annotation perspective, and affect labels.

Number of Participants	120 per game
Number of Gameplay Videos	360 (120 per game)
Video database size	12 hours
Number of Games	3 games
Gameplay video duration	2 minutes
Annotation Perspective	First-person
Annotation Type	Continuous unbounded
Affect Labels	Arousal

4.2 | Platformer Games: AGAIN Dataset

The methodologies proposed in this thesis are rigorously evaluated using data from three platformer games included in the Arousal Video Game AnnotatIoN (AGAIN) dataset (Melhart et al., 2022) (see Figure 4.3). These games were specifically chosen to offer a diverse range of gameplay experiences while maintaining computational feasi-

Table 4.3: Summary of the general gameplay features of the AGAIN corpus, outlining various metrics used to describe player actions, environmental elements, and game events.

feature	**description**
time_passed	time counted from the start of the game
score	the player score
input_intensity	number of keys pressed
input_diversity	number of keys pressed
idle_time	percentage of time spent with no key presses
activity	inverse of idle_time
movement	distance travelled
bot_count	number of visible bots
bot_movement	bot distance travelled
bot_diversity	number of visible unique bots
object_intensity	number of objects
object_diversity	number of unique objects
event_intensity	number of events
event_diversity	number of unique events

bility for experimental testing. The first game, *Endless*, is an infinite runner where players must avoid obstacles as they progress indefinitely, providing a continuous challenge that can provoke varying levels of arousal. *Pirates!* is a traditional jumping platformer that closely mirrors the gameplay mechanics of Super Mario Bros (Nintendo, 1985), introducing familiar yet engaging dynamics through level-based progression and obstacle avoidance. Finally, *Run'N'Gun!* adds a layer of complexity by requiring players to move, aim, and shoot simultaneously, combining action and strategy to test the player's reflexes and decision-making abilities under pressure.

These three games, though different in design, all share the characteristic of arcade-style controls, emphasising quick responses and rewarding players with points for successful in-game actions. The contextual diversity across these games makes them well-suited for examining arousal elicitation, as the variance in gameplay mechanics (e.g., jumping, running, shooting) can provoke distinct emotional reactions, from frustration to excitement. This range allows for a broad exploration of how different in-game stimuli influence players' affective states.

The AGAIN dataset was curated via Mechanical Turk, where participants engaged in two-minute gameplay sessions followed by a self-assessment of their arousal levels. After completing the gameplay, participants were asked to annotate their emotional responses using the RankTrace tool Lopes et al. (2017), which is integrated into the PAGAN (Platform for Affective Game ANnotation) environment Melhart et al. (2019). The

annotation process followed a stimulated recall protocol, meaning participants watched a replay of their own gameplay footage and provided a moment-to-moment arousal trace (see Figure 4.4). This method captures their emotional responses in a more reflective manner, as participants can better identify and label the affective states they experienced during gameplay. Table 4.2 summarises the key statistics of this dataset.

Similarly to RECOLA, the creators of AGAIN have provided a high-level feature set for each game. The general feature set (see Table 4.3) includes 14 features that describe player actions, environmental elements and game events. Moreover, the authors have extracted a set of game-specific features for each game (33 for *Endless*, 39 for *Pirates!* and 47 for *Run'N'Gun!*). The game-specific features correspond to the player status (e.g. player_health), gameplay events (e.g. bot_aim_at_player), bot status and the proximal and general game context (e.g bot_player_distance and pickups_visible). It should be noted that beyond the extracted features, raw data from all gameplay videos are also included in the dataset.

4.3 | The *GameVibe* Corpus

While the AGAIN dataset offers valuable insights into player arousal during gameplay, its design presents several limitations. The dataset includes annotations for a small number of games that feature relatively simplistic mechanics. This lack of contextual and gameplay diversity restricts the dataset's ability to capture the complexity and variability of affective responses across different gaming scenarios. To overcome this shortcoming, we introduce *GameVibe*—a dataset specifically curated to address these gaps. The *GameVibe* corpus (Barthet et al., 2024), available for download[1], is a dataset consisting of audiovisual gameplay footage from 30 first-person shooter (FPS) games, annotated for viewer engagement by a total of 20 participants. This dataset is designed to explore how different audiovisual elements of FPS games influence viewer engagement and is a valuable resource for studies in game perception, engagement analysis, and affective computing. The gameplay videos in the corpus are derived from a selection of 30 diverse and popular commercial FPS games, each with distinct visual and auditory styles, as depicted in Figure 4.5 and summarised in Table 4.4.

The selection of games and corresponding gameplay footage was guided by several key criteria. First and foremost, the primary goal was to capture a broad range of audiovisual stimuli that could elicit varying levels of engagement from viewers. To this end, the dataset includes games featuring diverse graphical styles, such as photorealistic en-

[1] `https://osf.io/p4ngx/`

Figure 4.4: Visualisation of arousal annotation for a player's gameplay footage, utilising the time-continuous unbounded RankTrace protocol. This method allows annotators to dynamically assess the player's arousal levels providing a continuous, fluid representation of emotional intensity. The unbounded nature of the RankTrace tool enables annotators to capture subtle fluctuations in arousal without preset limitations, offering a more nuanced understanding of the player's emotional experience throughout the gameplay session.

vironments found in modern AAA titles, retro pixelated graphics reminiscent of older generations of video games, and cartoon-like aesthetics common in stylised or arcade shooters. The release dates of all games present in this corpus are shown in Figure 4.8. Additionally, the selected games span multiple gameplay modes, such as Battle Royale, where players compete to be the last survivor; single-player campaigns, which often involve narrative-driven objectives; and Deathmatch, a fast-paced multiplayer mode focused on combat and scoring points by eliminating opponents. These diverse modes are illustrated in Figure 4.7.

Table 4.4: Summary of the key statistics of the *GameVibe* corpus, including participant count, gameplay video details, and annotation type focused on engagement.

Number of Participants	20 (5 per session)
Number of Gameplay Videos	120 (30 per session)
Video database size	2 hours
Number of Elicitors	30 games
Gameplay video duration	1 minute
Annotation Perspective	Third-person
Annotation Type	Continuous unbounded
Affect Labels	Engagement

Another important consideration in curating the *GameVibe* dataset was ensuring that the videos focused purely on gameplay without extraneous audio distractions, such as commentary from players or streamers. The corpus exclusively includes the original sounds from the games themselves, such as weapon fire, environmental noises, and background music, to provide a consistent and authentic gameplay experience for annotation. This allows researchers to isolate the impact of game audio-visual elements on viewer engagement without the confounding influence of external commentary or player reactions.

Moreover, in order to maintain the dataset's focus on gameplay-driven engagement, all videos were carefully edited to limit the inclusion of non-gameplay content, such as cut scenes, transition animations, or loading screens. Any non-gameplay elements were restricted to a maximum of 15 seconds within each video, ensuring that the bulk of the footage showcases active gameplay moments, which are more relevant for studying engagement patterns. This meticulous selection and editing process ensures that the *GameVibe* corpus provides high-quality, focused gameplay footage, ideal for studying the relationship between game stimuli and viewer engagement.

As mentioned above, the corpus contains two key modalities of game context information: video frames and in-game audio. These modalities provide a comprehensive view of the audiovisual aspects of gameplay. The first modality, video frames, consists of a series of high-resolution and low-resolution videos of in-game footage, sampled at 30Hz. The resolution of the videos varies based on the age of the games: for more recent games, the resolution is typically 1280×720 pixels, reflecting the standard for modern high-definition gameplay footage, while for older games, the resolution is 541×650 pixels, corresponding to the lower graphical capabilities of earlier titles. Each video segment represents a 60-second clip of gameplay, ensuring that a wide range of in-game scenarios are captured, from fast-paced combat to slower, exploratory moments.

Figure 4.5: Screenshots from the 30 different FPS games annotated for engagement. List of game titles: (1) Apex Legends; (2) Battlefield 1942; (3) Blitz Brigade; (4) Borderlands 3; (5) Corridor 7; (6) Counter Strike 2016; (7) Counter-Strike 2018; (8) Counter Strike 2019; (9) Counter Strike: Global Offensive; (10) Doom; (11) Dusk; (12) Far Cry 1; (13) Fortnite; (14) Heretic; (15) Hrot; (16) Insurgency; (17) Modern Combat: Sandstorm; (18) Medal of Honor 2010; (19) Medal of Honor 1999; (20) Medal of Honor: Pacific Assault; (21) Operation Bodycount; (22) Outlaws; (23) Overwatch 2; (24) PUBG; (25) Superhot; (26) Team Fortress 2; (27) Void Bastards; (28) Wolfenstein 3D; (29) Wolfenstein New Order; (30) Wolfram Wolfenstein.

The second modality, auditory information, is extracted directly from the gameplay videos and consists of stereo sound sampled at 44 kHz. The nature of the audio varies between older and more recent games. In older titles, the audio typically consists of MIDI-style background music, which was commonly used due to technological limitations in early video games. This type of music is often simple and looped, providing a rhythmic accompaniment to the gameplay. In contrast, the audio in more recent games tends to feature dynamic sound environments that respond to player actions and in-game events. These modern audio systems can include realistic sound effects, immersive background music, and ambient noise that changes based on the player's location or progress within the game. The combination of dynamic audio and interactive soundscapes in newer games adds another layer of complexity to the audiovisual experience, potentially influencing viewer engagement and emotional response.

The gameplay videos in the *GameVibe* corpus were annotated based on each viewer's level of engagement while watching the footage from a third-person perspective. This

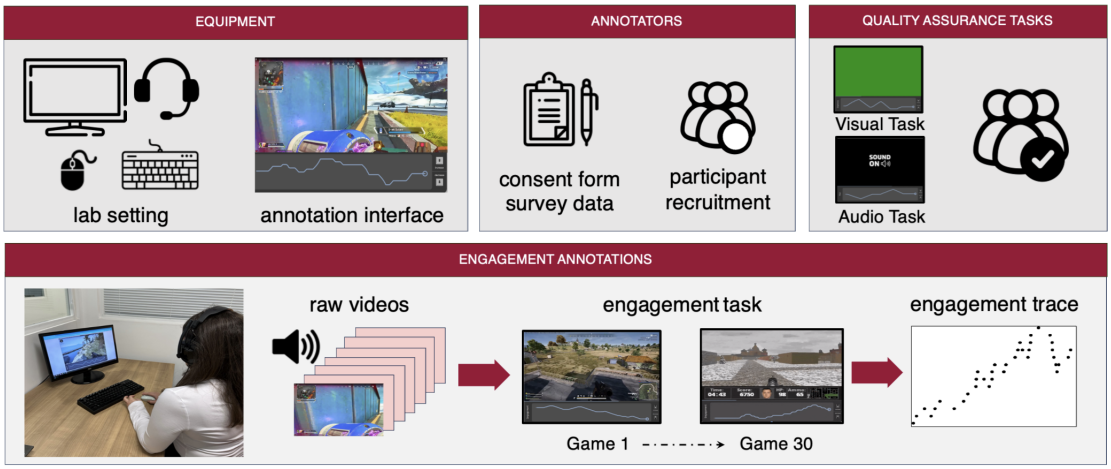


Figure 4.6: Summary of the *GameVibe* engagement annotation protocol. The top figures detail the setup, including the lab environment with the necessary equipment (e.g., annotation interface), the annotators' responsibilities (e.g., consent, survey data, and participant recruitment), and quality assurance tasks (e.g., visual and audio consistency checks). The bottom figure demonstrates the engagement annotation process, starting with the collection of raw videos, followed by annotators performing an engagement task on a series of game videos (30 short gameplay videos), resulting in the generation of a trace per game that quantifies user engagement.

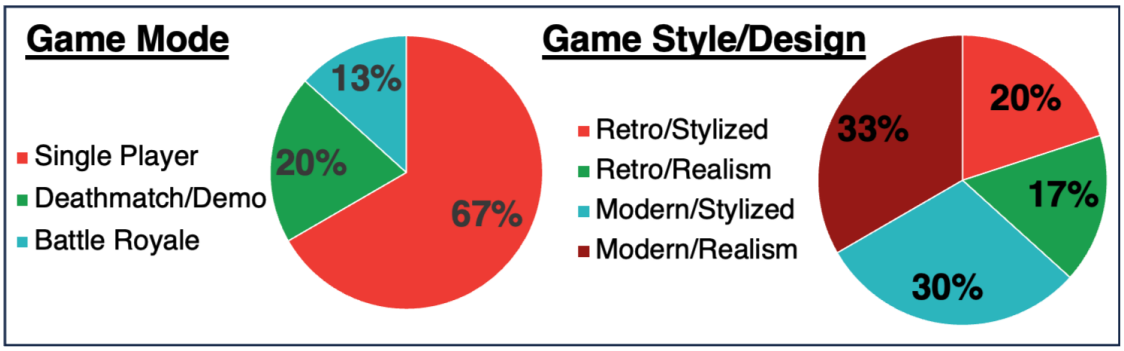


Figure 4.7: The diversity of the *GameVibe* dataset in terms of game modes (left) and game style/design (right). This diversity reflects a broad range of gameplay experiences and visual aesthetics in the dataset.

annotation protocol is considered third-person because the annotator is not the player of the game but instead evaluates their own engagement as an external observer of the gameplay. This annotation task was carried out over four different sessions, with each session being annotated by five different, randomly assigned participants. To ensure consistency, participants were provided with a clear, concise definition of engagement: "A high level of engagement is associated with a feeling of tension, excitement, and

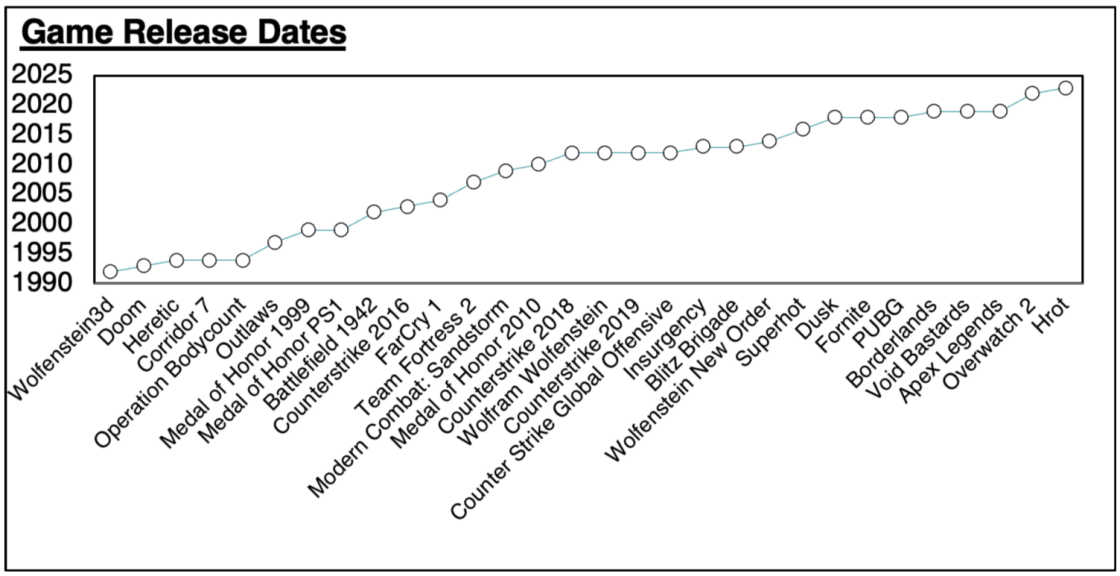


Figure 4.8: The timeline of game release dates within the *GameVibe* corpus, spanning from 1992 to 2024. The games included range from early titles like Wolfenstein 3D (1992) and Doom (1993), to more recent releases such as Apex Legends (2019) and Hrot (2024). This chronological diversity reflects the evolution of gaming over three decades.

readiness. A low level of engagement is associated with boredom, low interest, and disassociation with the game."

This definition helped standardise participants' understanding of engagement, ensuring that their annotations were based on similar criteria. In each session, participants were asked to annotate 30 short FPS gameplay videos, each lasting one minute, with one video representing each of the 30 games. The order in which the videos were presented was randomised to reduce the potential for habituation effects, where participants might become accustomed to the stimuli and exhibit biased responses over time. To further enhance the generalisability of the dataset, a different set of gameplay videos was used for each session. This was intentional, as the corpus is designed to facilitate the study of affect models across varying gameplay contexts. By exposing participants to different videos, the dataset allows researchers to explore how well affective models generalise across diverse audiovisual stimuli and gaming scenarios.

Participants had the ability to pause the annotation process at any time by pausing the video itself, allowing for more thoughtful and accurate annotations without the pressure of time constraints. Each session lasted approximately 30 minutes per participant. The combination of varied gameplay videos, participant randomisation, and clear engagement definitions ensures that the *GameVibe* corpus can be a valuable benchmark for the development of affective and engagement models in game research.

Collecting reliable engagement labels across multiple games simultaneously is impractical due to the high cognitive load of such a task Melhart et al. (2022); Souchet et al. (2022). To address this challenge, we argue that using short gameplay videos as stimuli for affect annotation allows for a balance between annotation reliability and stimulus richness. Short videos reduce the cognitive load on participants while still providing sufficient context for meaningful engagement patterns to emerge. Given the 30 games included in the corpus, each participant's session was limited to 30 minutes—this aligns with the maximum recommended duration for engagement annotation in gaming studies, as reported in the literature (Pinitas et al., 2023).

The 20 annotators who participated in this study were affiliated with the University of Malta, consisting of both research staff and graduate students. Their backgrounds varied, but all participants had a sufficient understanding of the study's objectives and the domain of affect modelling, ensuring that they were capable of providing informed and reliable annotations. Before starting the task, the annotators were briefed on the annotation guidelines to ensure a consistent approach. This briefing included detailed instructions on how to interpret the data and how to apply the annotation criteria accurately. To further ensure consistency and minimise potential biases or external influences on the annotation process, every participant completed the task under identical conditions. All annotators worked in the same controlled environment to reduce the variability that could arise from external distractions or differences in physical surroundings. The room was set up with standardised lighting, neutral wall colours, and consistent temperature settings to create a uniform atmosphere. Additionally, noise levels were controlled to eliminate auditory distractions, and all participants used the same type of computer and monitor setup, with calibrated display settings to ensure consistent visual perception of the data.

Moreover, the annotation process was standardised in terms of equipment and interface. All participants used the same machine and input/output devices, including a computer screen for visual stimuli, headphones for auditory stimuli, and a mouse scroll wheel for the annotation task. This uniform setup ensured that any variability in engagement annotations was due to the content of the gameplay videos rather than differences in hardware or environmental conditions. This carefully controlled environment and structured approach helped to improve the reliability and comparability of the engagement labels collected, ensuring that the resulting dataset offers robust insights into viewer engagement across diverse game contexts.

Data collection for the *GameVibe* corpus was conducted in two phases to ensure the quality and reliability of the engagement annotations. In the first phase, each participant was required to complete two simple yet controlled Quality Assurance (QA) tests—one

visual and one auditory. These QA tests were designed to verify the reliability of the annotators, ensuring that they could accurately perceive and respond to both visual and auditory stimuli in the videos Barthet et al. (2023). This initial step helped to filter out participants who might not meet the required perceptual standards for reliable engagement annotation.

Following the QA tests, participants were given a short break to ensure they were rested before proceeding to the main annotation task. In the second phase, participants were asked to watch 30 randomly ordered gameplay videos, each lasting one minute and featuring one of the 30 games in the corpus. Using the RankTrace annotation tool (Lopes et al., 2017), participants provided continuous annotations of their engagement levels while watching the videos. This continuous annotation approach allowed for real-time tracking of engagement, capturing fluctuations in how participants experienced each gameplay segment. The entire annotation protocol is illustrated in Figure 4.6 It is important to highlight that all data collection and analysis were carried out in full compliance with GDPR regulations and ethical guidelines specific to AI and games research (Melhart et al., 2023). This careful attention to ethical principles ensured that participants' privacy and data security were maintained throughout the study.

4.4 | Summary

In this chapter, we provided a detailed description of the three distinct datasets that form the foundation of the experimental work carried out in this thesis. These datasets, namely the RECOLA database, the AGAIN dataset, and the *GameVibe* corpus were selected due to their established role in affective computing and their relevance to the modelling approaches explored in this work. Each dataset offers a unique perspective, covering different modalities and domains, from real-time dyadic interactions in RECOLA to game player arousal in AGAIN and viewer engagement in FPS gameplay in *GameVibe*. These datasets ensure that the proposed methodologies are tested across a diverse set of contexts, providing robust evaluations of their performance in both well-established and novel benchmarks. Furthermore, the introduction of a new dataset focused on gameplay viewer engagement emphasises the practical applications of the proposed models and their capacity to generalise to new, unseen data. The subsequent chapters build on these datasets to showcase the effectiveness of the approaches presented in this thesis.

5

Contrasting Representations of Affect

Affect modelling has traditionally been approached as the process of mapping measurable manifestations of affect from multiple input modalities to predefined affect labels. This mapping is typically achieved through end-to-end machine learning models that directly learn the relationship between affective expressions and their corresponding labels. However, this thesis explores an alternative perspective: What if we first train general, subject-invariant representations that inherently encode affective information and subsequently utilise these representations for affect modelling? The central hypothesis of this work is that affective labels are not simply a training signal but an integral component of the representation itself. To investigate this, we leverage the emerging paradigm of contrastive learning, adapting it to discover high-level representations infused with affect. In particular, we investigate several supervised contrastive learning approaches for representation training, each incorporating affective information in unique ways. The proposed methods are evaluated in the context of arousal and valence prediction using the RECOLA and AGAIN datasets that provide multimodal user data. Experimental results highlight the potential of contrastive learning to produce representations with superior capacity for capturing affective information. These representations not only improve the accuracy of affect models compared to conventional end-to-end classification but are also general-purpose and subject-agnostic. This research underscores the potential of contrastive learning for affective computing, paving the way for more robust affect modelling frameworks.

5.1 | Motivation

Contrastive learning is a state-of-the-art machine learning paradigm that has shown exceptional success in learning general-purpose data representations across diverse do-

mains (Le-Khac et al., 2020; Saeed et al., 2021). As a self-supervised learning approach, it seeks to map data into a latent space where representations of different views of the same input are brought closer together while representations of dissimilar inputs are pushed apart Le-Khac et al. (2020). This method has seen widespread adoption, particularly in computer vision, where it has been leveraged to achieve robust performance in tasks such as image classification and object detection (Diba et al., 2021; Jaiswal et al., 2020). However, its application within affective computing has only recently gained attention. Emerging research demonstrates its potential for learning representations that are invariant to individual differences, such as subject-specific biases (Shen et al., 2022).

This thesis is motivated by the hypothesis that affective information is inherently tied to its manifestations and can be explicitly encoded into learned representations through contrastive learning. This perspective challenges the traditional end-to-end approaches commonly used in affect modelling, which rely on direct mapping from raw input features to affective states. Instead, we propose leveraging affect annotations to construct contrastive labels, thereby enabling the training of models that encode affective information more effectively. The objective is to design a method for producing models capable of learning general, affect-infused representations.

To validate this hypothesis, this chapter builds upon the Supervised Contrastive Learning framework (Khosla et al., 2020) adapting it for affective computing. Specifically, our approach involves constructing contrastive labels based on affect annotations, enabling the learning of representations that incorporate affective dynamics. These methods are evaluated on the RECOLA and AGAIN datasets (Melhart et al., 2022; Ringeval et al., 2013). As described in chapter 4 the former is a benchmark for multimodal affect modeling that includes continuous annotations for arousal and valence and the latter is a widely-used game dataset for player arousal modelling. The experimental setup involves extracting features from multiple modalities, including audio, video, and physiological signals, within defined temporal windows. These features are then input into a neural network trained via SCL to learn affect-infused representations (see Fig. 5.1). A probe model (Belinkov, 2021) is subsequently employed to assess the quality of these representations through the task of affect state classification (high vs low).

Experimental findings reveal that models trained using SCL consistently outperform traditional end-to-end classification models across a range of experimental settings. Specifically, SCL-based models achieve significantly higher accuracy, demonstrating their ability to learn representations that generalise effectively to downstream tasks. This highlights two key contributions of the thesis: (a) contrastive learning is a powerful framework for learning robust representations in multimodal affect modelling, and (b)

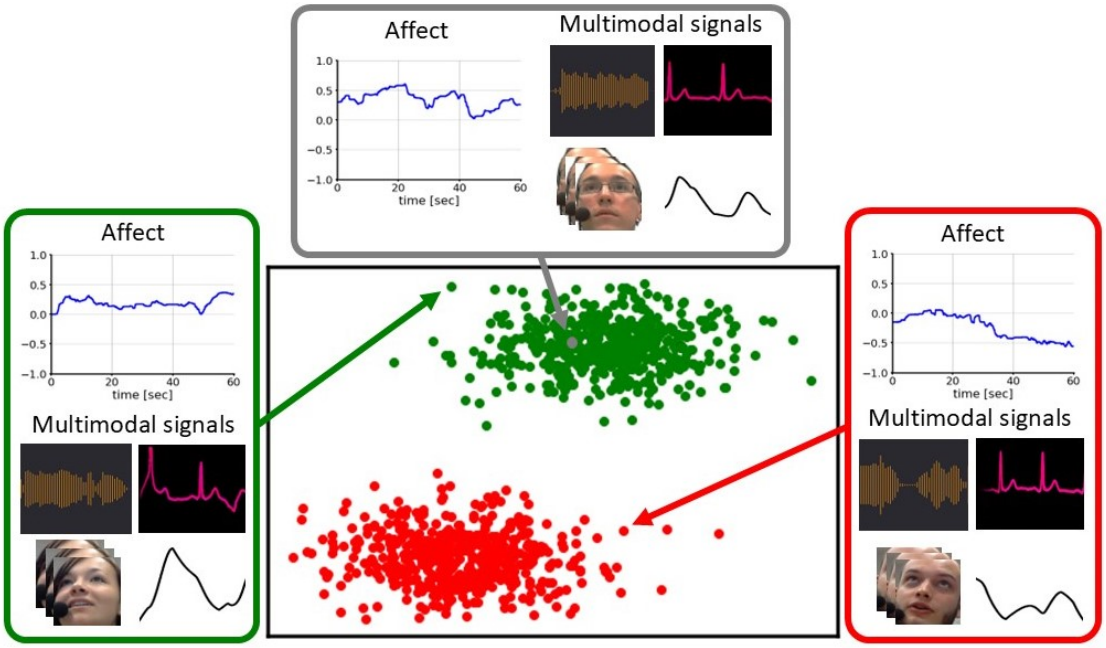


Figure 5.1: A high-level overview of the introduced concept. Supervised Contrastive Learning operates by infusing affect information within the representation, by maximising the similarity between positive embeddings while making negative embeddings dissimilar. The approach assumes that affect is encoded within a multimodal latent space, acting as the defining characteristic that distinguishes data points. Multimodal samples are labelled as positive (green) or negative (red) relative to an anchor affect (grey). Positive pairs share affective patterns similar to the anchor, while negative pairs exhibit contrasting patterns. This framework produces generalised representations, effectively capturing affective patterns across participants.

affective information can be fused into learned representations through CL, resulting in more effective models.

These results provide compelling evidence for the utility of contrastive learning in affective computing and position this framework as a promising approach for advancing the field. By focusing on representation learning, the thesis addresses fundamental challenges in affective computing, such as participant invariance, data sparsity, and multimodal fusion. This work aligns with the broader aim of the thesis to enhance affect modelling through innovative approaches to representation learning.

5.1.0.1 | Contributions

This thesis introduces several novel contributions to the fields of multimodal affect modelling and representation learning. Foremost, it is the first to adapt the Super-

vised Contrastive Learning framework (Khosla et al., 2020) for the purpose of learning affect-infused multimodal representations. By leveraging contrastive learning, this work offered a fresh perspective on affect modelling, enabling the development of representation spaces that encode affective information showcasing that SCL is a versatile method for advancing the understanding of affective phenomena. Another key innovation of this research lies in the exploration of the generalisation capacity of representations learned via SCL highlighting its flexibility in capturing various facets of affective information, thus providing insights into how contrastive learning can be tailored to different affective modelling tasks. Finally, the proposed methodology was rigorously evaluated against end-to-end classification models across a diverse set of modalities within both the RECOLA and AGAIN datasets, focusing on arousal and valence prediction. This comparative analysis not only validates the effectiveness of SCL in this context but also extends the relatively sparse body of literature at the intersection of contrastive learning and multimodal affective computing. By addressing this gap, the thesis establishes a foundation for future work on leveraging contrastive learning to advance affective computing. Through these contributions, this research provides a significant step forward in affect modelling, showcasing how representation learning can be harnessed to create more robust affective models. These advancements align with the broader objectives of the thesis to advance the domain of affective computing and pave the way for generalisable, human-centred AI systems.

5.2 | Modelling Methodology

This section outlines the key components of the algorithms explored in this study. It begins with describing the representation learning modules, followed by a discussion on the appropriateness of SCL for AC. Lastly the techniques used to generate the affect-based supervision signal for contrastive learning were presented. The code for this work is available on GitHub.[1].

5.2.1 | Representation Components

In this section, we first present the main components used in representation learning, namely encoders and probes as depicted in Fig. 5.2. Then, we present a baseline architecture of an end-to-end classifier that we compare against our SCL methodology for assessing the effectiveness of the obtained affect infused representations.

[1] https://github.com/kpinitas/Supervised-Contrastive-Learning-for-Affect-Modeling

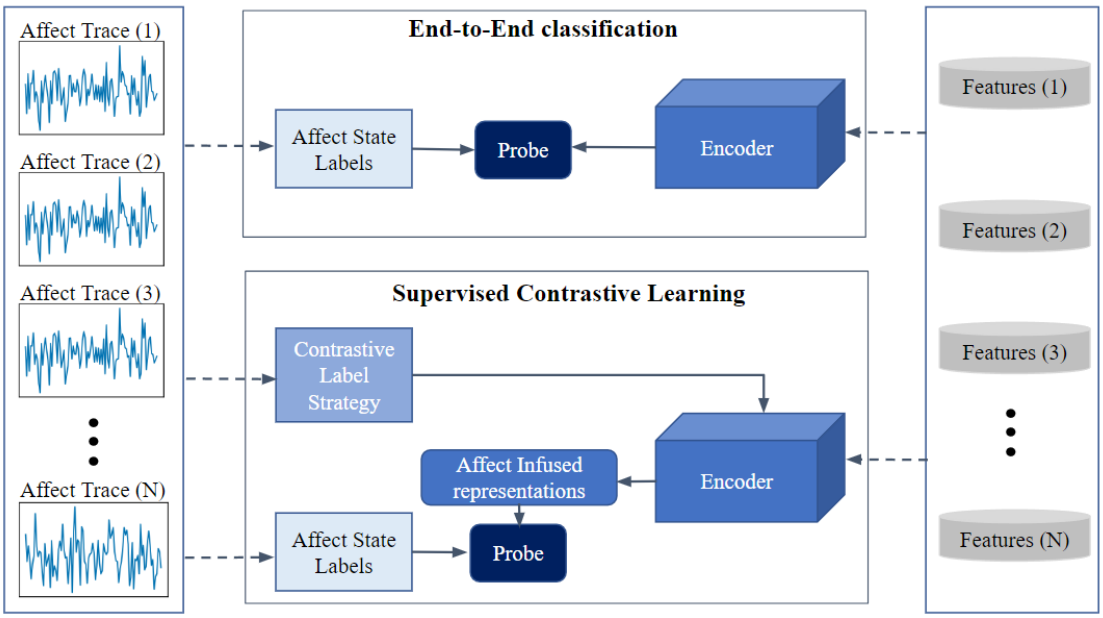


Figure 5.2: Illustration of the training methods employed: the end-to-end classification baseline (top) and the SCL method (bottom). In both learning paradigms, affect labels are derived from participants' annotations such as affect traces from N participants (as depicted on the left of the figure). The corresponding participant's features (depicted on the right of the figure) can be extracted from a single or multiple modalities. SCL first derives affect-infused labels for contrastive encoder pretraining and then trains the probe model based on the representations of the trained encoder.

5.2.1.1 | Encoder

An encoder model E is a model that projects high dimensional data into a latent space of lower dimension, producing high-level representations of the input data. Hence, after training, E is a function that reduces the dimensionality of the data while maintaining essential information about the input space. In this work, we hypothesise that affect information is present on the input space and thus it can be integrated into the latent space via contrastive learning to yield more robust representations. In this work we considered three different encoder architectures namely feature encoder, frame encoder and fusion encoder.

Feature Encoder: The feature encoder used in this work is a simple Artificial Neural Network (ANN) consisting of a single layer of 30 neurons activated by Sigmoid.

Frame Encoder: The frame encoder used in this work is a CNN comprising five convolutional layers of 6, 8, 12, 16 and 20 filters, respectively, followed by a dense layer. The first four convolutional layers are configured with a stride parameter of 2, while

the fifth convolutional layer uses a stride of 1. Each convolutional layer has a kernel size of dimensions 3x3 and employs the ReLU activation function. Finally, a dense layer with 768 neurons, also activated using ReLU, is applied to produce the encoded feature representation. This encoder exploits spatial information by learning filters that operate along spatial dimensions. Moreover, the learned filters implicitly encode the temporal information of the sequence of frames that is passed as input. It should be noted that additional experiments with a small ViT have been performed and the results are reported in the Appendix A.3.

Fusion Encoder: The fusion encoder consists of a frame and a feature encoder. An additional relu-activated layer of 30 neurons is added to the frame encoder in order to match the dimensions of the feature encoder. In this case the feature encoder is a single 30-neuron layer activated by ReLU. The last layer of the fusion encoder is a linear layer of 60 neurons activated by ReLU. The purpose of this last layer is to fuse information from both frames and features in order to produce the final representation.

5.2.1.2 | Probe

Probe architectures are employed to assess the quality of representations learned by a pre-trained encoder E. Specifically, given a known property of the input data (e.g., object categories in object recognition tasks), a probe is trained to evaluate whether this property has been successfully encoded in the latent space. While probe architectures can theoretically include multiple hidden layers, they are typically implemented as a single linear layer with a softmax or sigmoid activation function, commonly referred to as a *linear probe* (Belinkov, 2021).

5.2.1.3 | Baseline

The baseline architecture employed in this work performs end-to-end classification, bypassing the representation learning process based on contrastive labels described in Section 5.2.3, and is denoted as E_b. The E_b model comprises a randomly initialised encoder E followed by a randomly initialised probe architecture. Both the encoder and the probe are trained simultaneously to map inputs directly to affect labels.

5.2.2 | Supervised Contrastive Learning for Affect Modelling

Affective phenomena are continuous, ambiguous, and context-dependent, which makes their representation fundamentally different from discrete objectively-defined tasks such as object recognition. In many cases, affective information is not directly evident in the

raw audiovisual input, but instead emerges from subtle multimodal cues, temporal dynamics, and relative comparisons between samples. Learning affective representations therefore requires methods that go beyond identifying coarse categorical boundaries.

Conventional cross-entropy optimisation is limited in this regard. It enforces separability only at the level of decision boundaries, without shaping the structure of the embedding space. Consequently, subtle affective distinctions can be collapsed into neighbouring decision regions, leading to representations that are less sensitive to graded emotional variation.

Unsupervised contrastive learning offers stronger representation learning by constructing embeddings through instance discrimination. However, its reliance on automatically generated positive and negative pairs typically reflects low-level signal similarities—such as background, speaker identity, or visual appearance—rather than affective similarity. Since affective information is not always explicit in the raw input, unsupervised contrastive learning risks encoding superficial correlations while neglecting the latent emotional structure of the data.

Supervised contrastive learning (SCL) directly addresses these shortcomings by incorporating affective labels into the construction of the embedding space. Samples annotated with similar affective values are treated as positives and pulled closer together, while those with divergent values are pushed apart. This relational inductive bias enforces a geometry that mirrors the graded nature of affective states (e.g., high versus moderate arousal), where relative similarity is often more meaningful than categorical separation. As a result, SCL explicitly encourages embeddings to encode affect-relevant variance that may not be apparent from raw input alone.

From an information-theoretic perspective, SCL can be viewed as maximising the mutual information between learned embeddings and affective annotations, while minimising redundancy across unrelated samples. This ensures that the representation space captures the structure of emotional variability rather than noise or confounding factors. In practice, this translates into embeddings that are more robust under domain shift, more transferable in data-scarce contexts, and more resilient to annotation variability.

The analyses conducted during this doctoral research, though not included in detail in the original submission, supported this theoretical reasoning by showing that SCL embeddings preserved affect-related structure more effectively than both cross-entropy and unsupervised contrastive baselines (Appendix A). These findings motivated the systematic studies presented in Chapters 5–7, providing the binding theoretical context for subsequent investigations into Learning Using Privileged Information (LUPI) and Few-Shot Learning (FSL), both of which rely on discriminative representations.

5.2.3 | Affect-Infused Contrastive Labels

The primary contribution of this chapter lies in leveraging supervised contrastive learning algorithms to derive affect-infused representations that enhance the downstream affect modelling task. Specifically, SCL is employed to pre-train the encoder architecture (refer to Section 5.2.1.1) using affect-based contrastive labels, thereby embedding affective information into the learned representations. Following this pre-training phase, the encoder's weights are frozen, and only the linear probe architecture (outlined in Section 5.2.1.2) is trained to perform the downstream task. This overall training procedure is illustrated in Fig. 5.2.

A critical aspect of contrastive learning lies in the selection of positive and negative samples, as this directly influences the quality of the contrastive supervision signal. When objective annotations, such as class labels, are available, this selection process is relatively straightforward: positive pairs typically consist of samples from the same class, while negative pairs are drawn from different classes (Khosla et al., 2020). However, for subjective annotations such as arousal and valence signals, the pair selection process becomes considerably more challenging. This difficulty arises from the subjective nature of such annotations and the noise introduced by the biases of human annotators (Yannakakis et al., 2018).

This study initially considered several distinct affect measurements derived from any continuous annotation trace: one *absolute* measure reflecting the subject's current emotional state and two *relative* measures capturing fluctuations in the emotional state within a predefined time window (Camilleri et al., 2017; Yannakakis et al., 2018). However, in this chapter, we focus exclusively on the absolute measure, that is the *affect state* (g_s), defined as the mean value of the affect trace within a given time window (Eq. 5.1) (Makantasis et al., 2021a; Melhart et al., 2022).

$$g_s = \frac{1}{w} \sum_{i=0}^{w} v_i \tag{5.1}$$

where w is the window size considered and v_i is the i-th annotation value of the time window. It should be noted that the definitions and results obtained with the remaining relative measures are reported on Appendix A.2. Based on g_s, we explore the following positive/negative sample selection strategy that we detail below. It should be noted that, for the proposed contrastive labelling strategy, we train the SCL models using the same loss function: i.e. the supervised contrastive loss function L_{SC} of Eq. 3.1.

5.2.3.1 | Contrasting Affect: High vs. Low

Intuitively, contrastive labels can be generated by pairing windows with similar affect states as positive pairs and those with dissimilar affect states as negative pairs. To establish affect state similarity, we binarise the affect states g_{s_i} into "high" and "low" categories, considering windows with the same (or different) states as similar (or dissimilar), respectively. In RECOLA, the binarisation criterion is determined by the median ground truth value of the entire set of affect annotation traces ($\tilde{g}_s$) and a threshold ϵ. Specifically, a time window i is labelled as "high" if $g_{s_i} > \tilde{g}_s + \epsilon$ and as "low" if $g_{s_i} < \tilde{g}_s - \epsilon$. The median is used for RECOLA because each participant's data has been annotated by the same six expert annotators, making the median a robust measure of central tendency. For RECOLA, ϵ is set to 0.1, ensuring a precise exclusion of ambiguous annotations.

For the AGAIN dataset, the binarisation criterion is based on the mean value of the annotation traces ($\bar{g}_s$) for each session, as this better reflects the session-specific affect distribution. Each gameplay session in AGAIN is annotated via self-reporting, where the mean is more representative of the central tendency. A time window i is labelled as "high" when $g_{s_i} > \bar{g}_s + \epsilon$ and as "low" when $g_{s_i} < \bar{g}_s - \epsilon$. For AGAIN, ϵ is set to 0.2, accommodating the variability in self-reported annotations. These thresholds were empirically validated to ensure a balanced exclusion of ambiguous annotations while maintaining sufficient data for effective contrastive learning. Such ambiguous values could compromise the stability of the SCL models and, consequently, the effectiveness of the learned representations.

5.3 | Data Preprocessing

The methodology proposed in this chapter is tested on two challenging datasets. The first dataset, RECOLA is a multimodal dataset of online dyadic interactions between 23 participants annotated by 6 experts in terms of arousal and valece. The second dataset, consists of the platformer games of the AGAIN dataset. In this dataset the participants first played the game and annotated their own gameplay. This section describes the preprocessing steps followed for both RECOLA (Section 5.3.1) and AGAIN (Section 5.3.2) datasets.

5.3.1 | Processing RECOLA

As mentioned above the RECOLA database provides both arousal and valence annotations. The same preprocessing approach is followed for both affective dimensions.

Table 5.1: Summary of the number of samples in the RECOLA dataset before and after binarisation of the arousal and valence annotations, along with the percentage of the majority class for each time window size.

RECOLA	**Arousal**			**Valence**		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Before Binarisation	16750	16692	16635	16750	16692	16635
After Binarisation	13752	13612	13444	10827	10750	10630
Majority Class (%)	50.28	50.39	50.82	55.75	55.46	55.99

Specifically, we segment each participant's session (features and frames) into overlapping time windows using a sliding step of 400 ms and window lengths of 1, 2, and 3 s. These hyperparameters—the sliding step and window length—influence the size of the dataset and the temporal granularity of information contained in each window. Since the features and annotations are already synchronised, there is no need to account for the reaction time between the stimulus and the emotional response. After segmentation, each time window contains a sequence of feature vectors and a sequence of frames. To reduce computational complexity, we compute the average value for each feature within the window, representing the time window with a single feature vector. This ensures that the dimensionality of the feature vector is independent of the window length. For frame sequences, we retain 5 grayscale frames with dimensions 224×224 per second. For example, the input for a 3 s time window consists of a single feature vector (via averaging) and a frame tensor with dimensions $15 \times 224 \times 224$. When it comes to affect annotation, we use the median annotation values per time window in order to mitigate inter-annotator disagreement Grewe et al. (2007). The arousal state score g_a is computed based on Eq. 5.1 using the median arousal trace. Similarly the valence state score g_v is calculated based on Eq. 5.1 using the median valecne trace. Table 5.1 presents the number of samples and the corresponding majority class percentage per window length.

5.3.2 | Processing AGAIN

Each game of the AGAIN dataset provides self-reported arousal annotations, which are used in this work with a consistent preprocessing approach applied separately to each of the three games. Similarly to RECOLA, we segment each participant's session (features and frames) into overlapping time windows using a sliding step of 500 ms (the features are provided in intervals of 250 ms) and window lengths of 1, 2, and 3 s. Unlike the RECOLA dataset, synchronisation is required for AGAIN. To account for the reaction time between stimulus and emotional response, the arousal annotations are shifted

Table 5.2: Summary of the number of samples in the AGAIN dataset before and after binarisation of arousal annotations, along with the percentage of the majority class for each game and window length.

AGAIN	**Run'N'Gun!**			**Pirates!**			**Endless**		
	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec	1 sec	2 sec	3 sec
Before Binarisation	29548	29310	29072	25672	25454	25236	29041	28803	28565
After Binarisation	12179	11806	11433	11528	11296	11070	12792	12557	12274
Majority Class (%)	50.13	50.03	50.13	50.29	50.00	50.09	50.02	50.12	50.45

backwards by 1 s Melhart et al. (2021b); Pinitas et al. (2022a). Furthermore, the trace of each participant is normalised within $[0, 1]$ due to the unbounded nature of the annotations in this dataset. Once again, after preprocessing, each time window contains a sequence of feature vectors and a sequence of frames. To reduce computational load, we compute the average value of each feature within the window, resulting in a single feature vector that represents the entire window. For frame sequences, we retain 5 grayscale frames with dimensions 224×224 per second. In the case of arousal annotation, there is no need to mitigate the disagreement between annotators since each participant provides a single trace per game per session (self-reporting annotations). For each session, the arousal state score g_a is computed based on Eq. 5.1 using the corresponding arousal trace (only one trace per session per game). This preprocessing pipeline is repeated separately for each of the three games in the dataset. Table 5.2 presents the number of samples and the corresponding majority class percentage per window length.

5.4 | Results

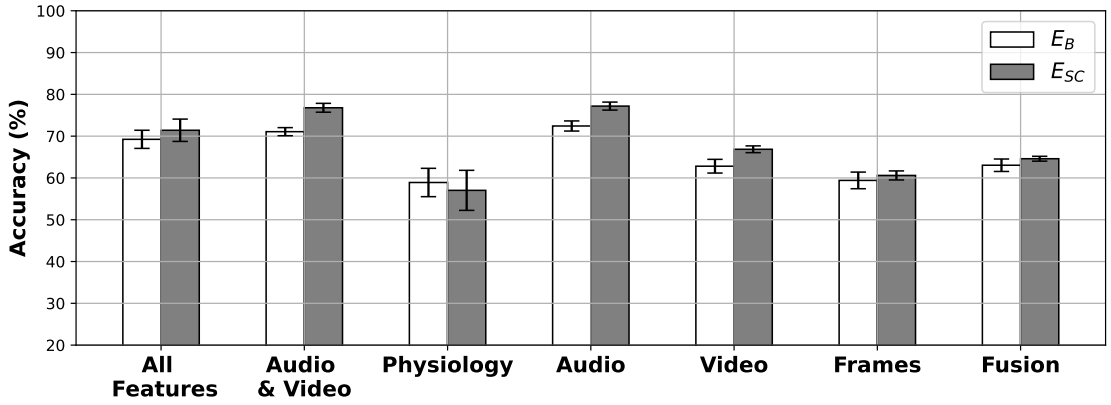
This section first outlines the experimental protocol we use for the evaluation of the methods and then presents the key experimental results obtained for both datasets. It should be noted that RECOLA and AGAIN capture fundamentally different constructs—expert annotated continuous traces vs. self-reported single annotations—with different scales, reliabilities, and normalisation strategies. Accordingly, no direct aggregation across these datasets was performed. Instead, results are reported separately for each dataset (e.g., games in AGAIN, affect dimensions in RECOLA), and comparisons are meant to highlight consistency in methodological trends within each dataset rather than equivalence of absolute values.

5.4.1 | Experimental Protocol

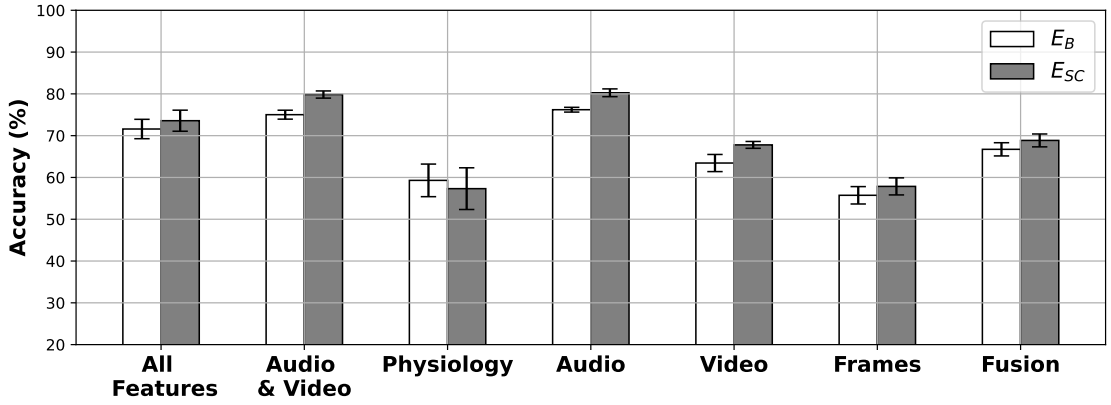
In this chapter, we evaluate the proposed method on the downstream task of affect state classification, where the model learns to classify features within a time window as either a low or high affective response. The encoder E outputs a high-level representation corresponding to the given time window, serving as the input for classification. Each model is trained using the Adam optimiser with a learning rate of 0.001 and a batch size of 128. Additionally, we set the temperature parameter τ in Eq.3.1 to $\tau = 0.1$ To generate the corresponding class labels (e.g. "high" vs. "low" arousal), we follow the procedure outlined in Section 5.2.3.1. For performance evaluation, we adopt a five-fold cross-validation strategy, ensuring that data from different participants are used for training validation and testing, thereby maintaining non-overlapping sets. The validation set size is defined as the 10% of the dataset which corresponds to data from approximately 2 and 10 participants for RECOLA and AGAIN respectively. The models are trained following a validation protocol that employs early stopping, terminating the training process after 5 epochs without improvement in validation loss. The best-performing model is then returned. The class-splitting criterion and the arbitrary threshold ϵ (0.1 for RECOLA and 0.2 for AGAIN) ensure that both the datasets remain balanced. This balance facilitates performance evaluation in terms of accuracy. Experiments are conducted for time window lengths of 1, 2, and 3 seconds to assess the robustness of the proposed method across varying temporal resolutions. It should be noted that experiments were performed independently within each dataset. For AGAIN, a separate model was trained per game (within-game experiments) as the dataset contains a single affective dimension. For RECOLA, distinct models were trained for each affective dimension (arousal and valence). In both cases, the focus was on demonstrating generalisation across participants while maintaining evaluation within each distinct use-case

5.4.2 | Contrastive Learning for Affect Modelling on RECOLA

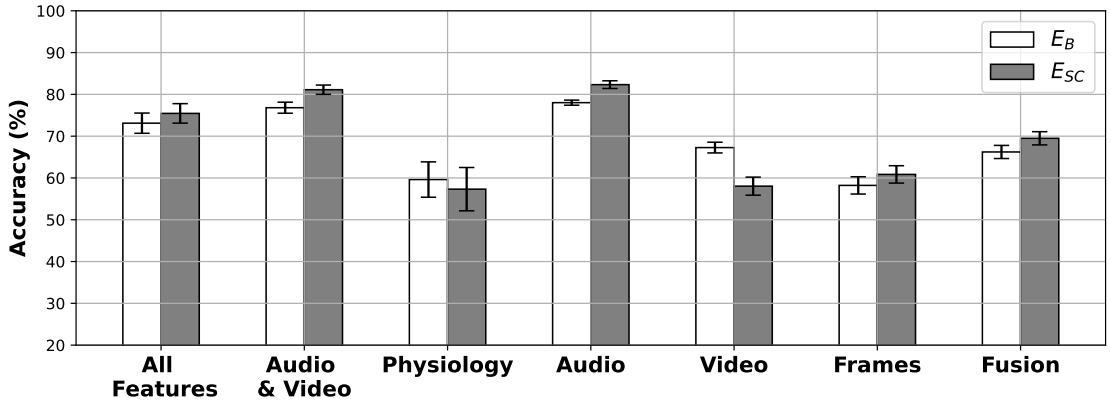
We aim to examine how representations learned through contrastive learning perform in the downstream task of classification between high and low affect states. To this end, we train encoders using supervised contrastive learning with affect state labels corresponding to high and low arousal and valence (E_{SC}). For each encoder, a probe model is trained as outlined in Section 5.2. The baseline model, E_b, conducts end-to-end arousal classification. Our investigation includes seven modality configurations for each affect dimension in the RECOLA dataset: individual modalities (audio, video, and frames), the bimodal physiological signals (ECG and EDA), and the bimodal audio-video combi-



(a) Arousal (1 second time windows)

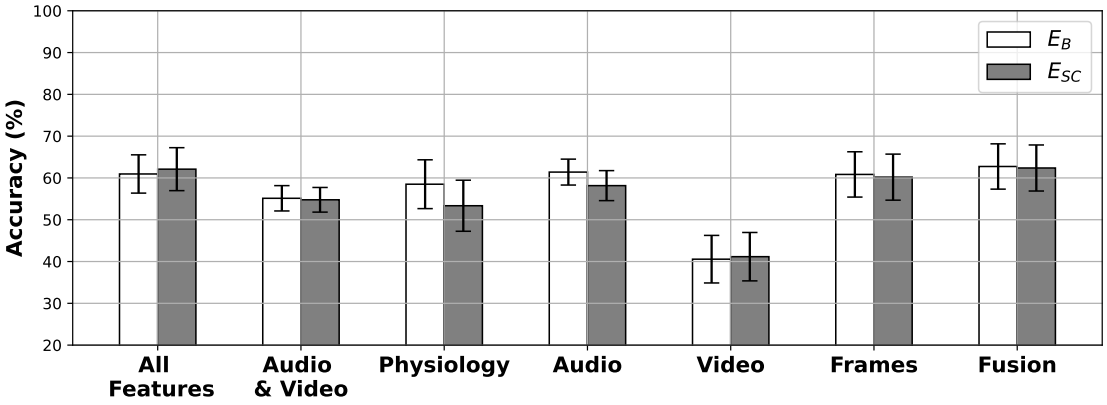


(b) Arousal (2 second time windows)

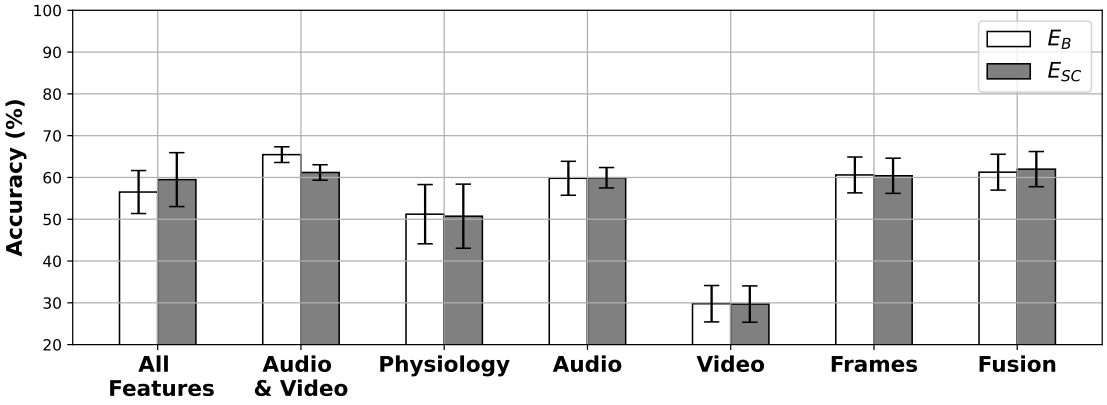


(c) Arousal (3 second time windows)

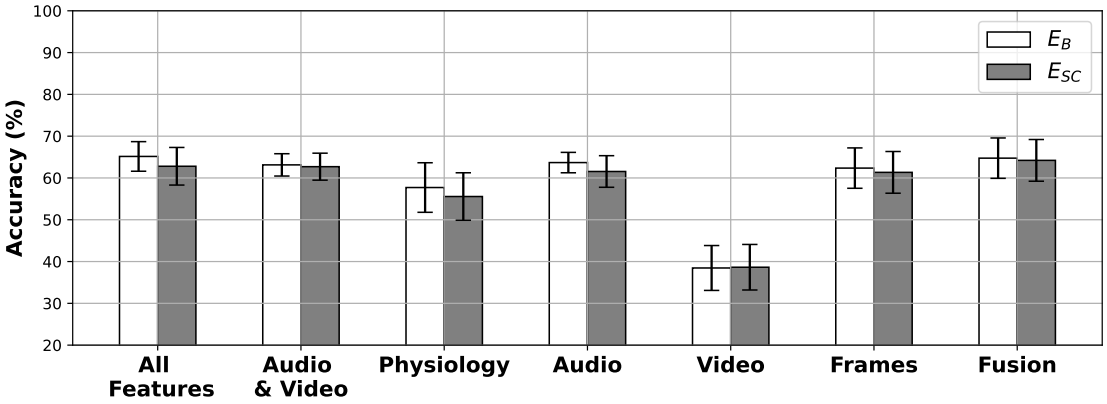
Figure 5.3: **RECOLA Dataset** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) Valence (1 second time windows)

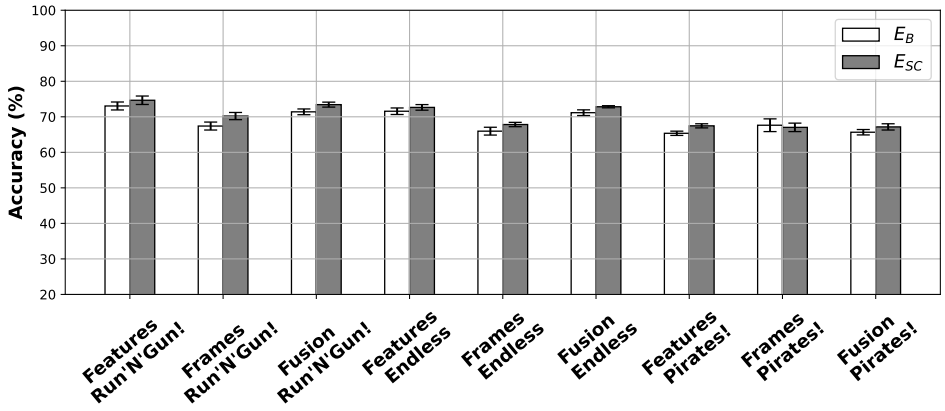


(b) Valence (2 second time windows)

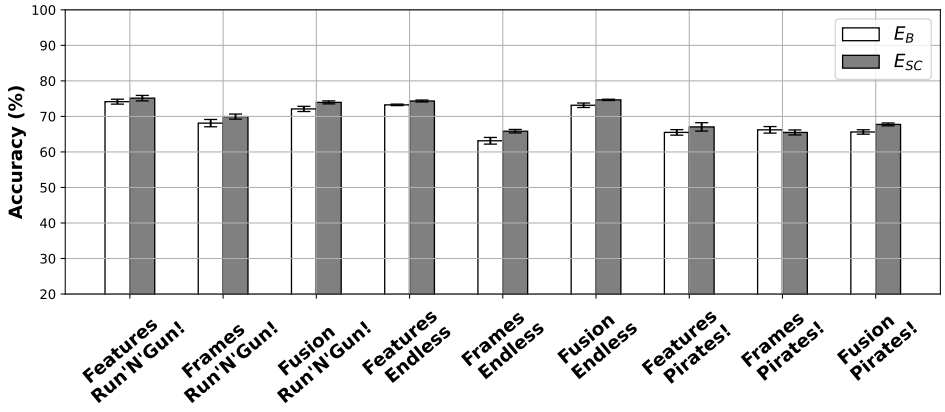


(c) Valence (3 second time windows)

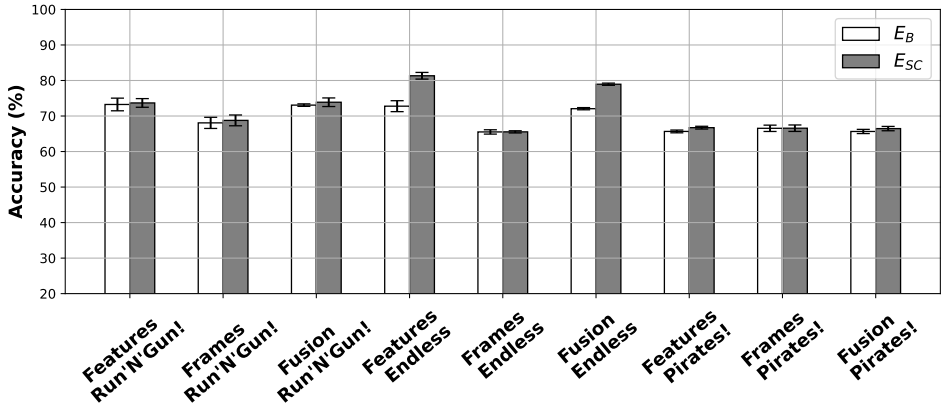
Figure 5.4: **RECOLA Dataset** Average 5-fold validation accuracy scores (%) for high-low valence classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) 1 second time windows



(b) 2 second time windows



(c) 3 second time windows

Figure 5.5: **AGAIN Dataset** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.

nation. Additionally, we explore two multimodal setups: all features combined and late fusion of all features with frames. The experiments cover two modelling approaches, three-time window lengths, and the seven modality configurations, as illustrated in Figures 5.3 and 5.4 for arousal and valence respectively. It is worth noting that the performance is derived from 5-fold cross-validation accuracy, averaged across folds over five independent runs. Statistical significance is determined using an one-tailed paired t-test, with a significance level of $p < 0.05$.

Figure 5.3 showcases the peroformance of the proposed approach for the downstream task of arousal classification within the RECOLA dataset. In particular, E_{SC} results in the best-performing model. Specifically, the E_{SC} model outperforms the baseline end-to-end classifier (E_b) yielding higher accuracy scores across all window lengths and modality configurations except physiology. The pairwise significance tests showcase that E_{SC} performs significantly better than E_b in 17 out of 21 experimental settings (modality configurations and time window lengths). In the case of physiology both modelling approaches perform statistically on par. Comparing across RECOLA modalities, the models achieve the highest accuracy when arousal modelling relies on audio features across all time window lengths. High accuracy scores are also obtained by the models when the audio features are fused in the input space with video features (Audio & Video) and with video and physiology (All Features). Although video features can yield robust arousal predictors, resulting models are inferior to the audio-based models. In contrast, all models underperform when only physiological signals or frames are considered, regardless of training method. It appears that arousal is not well captured by physiology or frames in the RECOLA dataset. Interestingly, while fusing all features with frames (Fusion) improves accuracy compared to frames alone, it performs worse than multimodal fusion without frames (All Features).

In the case of valence, all models perform comparably, with significantly lower accuracy scores compared to arousal. This observation highlights the inherent difficulty of modelling valence using the modalities available in the RECOLA dataset. Specifically, valence appears to be poorly captured by the individual and combined features derived from audio, video, and physiological signals. The consistent underperformance of valence models across all configurations and time window lengths suggests that the features extracted from these modalities do not adequately encode information related to the emotional polarity of a given state. This limitation is further supported by evidence from the Hall of Fame [2] of the AVEC competition, where valence prediction has historically lagged behind arousal prediction across various methodologies and modal-

[2]Hall of Fame models: `https://diuf.unifr.ch/main/diva/recola/news.html`

ities. This suggests that valence-related signals are either less robust or more context-dependent, making them inherently more challenging to model. For example, physiological signals such as ECG and EDA, which capture changes in autonomic nervous system activity, are more closely linked to arousal than to valence. Similarly, while audio and video features may contain cues relevant to valence, these signals are often subtle and may be influenced by external factors such as context or speaker intent, further complicating their interpretation.

5.4.3 | Contrastive Learning for Affect Modelling on AGAIN

In the case of the AGAIN dataset, designed specifically for player affect modelling, we investigate the performance of representations learned via SCL in the downstream task of high vs low arousal classification. Similar to our RECOLA experiments, we train encoders using SCL with arousal labels (E_{SC}). Additionally, a baseline end-to-end classifier, E_b, is trained to provide a direct comparison. Our experiments span three diverse platformer games, two modelling approaches per game (SCL and baseline), three time window lengths, and three modality configurations unique to this dataset: game features, game frames, and their fusion. Performance is evaluated using 5-fold cross-validation accuracy, averaged over five independent runs. Statistical significance is determined via one-tailed paired t-tests, with a significance threshold of $p < 0.05$.

Figure 5.5 illustrates that E_{SC} achieves higher accuracy than the baseline E_b in the majority of experimental settings, demonstrating its efficacy across games, modality configurations, and time window lengths. Specifically, E_{SC} significantly outperforms E_b in 20 out of 27 settings, while in the remaining seven, both models perform statistically on par. An analysis of the modality configurations reveals game-specific trends. In *Endless!* and *Run'N'Gun!*, arousal modelling based solely on game features consistently yields more robust models compared to frame-based or fusion-based approaches. This suggests that the dynamics of gameplay metrics in these games provide richer and more reliable arousal-related information than visual cues. In contrast, for *Pirates!*, there is no clear best modality configuration. Here, frame-based and fusion-based models achieve the highest accuracies, with feature-based models trailing slightly behind. This variability may reflect differences in the way arousal is manifested or captured across the three games, potentially influenced by the distinct gameplay mechanics and visual styles.

5.5 | Discussion

The findings presented in this chapter highlight the potential of SCL for learning representations that retain affect-infused information, The resulting models were particularly tested in the classification of high versus low affect states within two distinct case-studies: dyadic interactions (RECOLA) and player modelling (AGAIN). Despite the promising results, the presented results have several methodological limitations and considerations. A significant methodological simplification is the use of grayscale frames for visual input. It can be argued that the use of grayscale frames may limit the richness of visual features available for learning, particularly in contexts where colour information could play a role in affect recognition. For example, in the AGAIN dataset, colourful game environments might convey arousal-related cues that are muted or lost in grayscale representations. However, the use of grayscale frames instead of full-colour images was a deliberate choice to reduce computational complexity and memory requirements, enabling the experimentation with multiple games, modalities, and modelling approaches within the limited computational resources available during this project. Incorporating full-colour images could have increased the computational burden, potentially limiting the breadth of the experiments conducted. It should be noted that the use of grayscaled frames and late fusion are two widely used approaches in affective computing Makantasis et al. (2021a); Pinitas et al. (2023).

Additionally, the absence of data augmentation techniques could restrict the generalisability of the models to unseen data. Techniques such as random cropping, flipping, brightness adjustments for frames, or perturbations of numerical features could introduce variability during training and improve the robustness of the learned representations. This is particularly relevant for datasets with constrained diversity, such as AGAIN. the absence of data augmentation techniques stems from the need to prioritise foundational experiments that align with the primary focus of this thesis—evaluating the effectiveness of SCL for affect representation learning. While augmentation could enhance the robustness of the models, implementing and systematically testing various augmentation strategies for both numerical and visual inputs was deemed outside the core scope of this work, which centres on methodological contributions rather than dataset engineering.

Furthermore, the decision to focus on binary classification (high vs. low) instead of exploring regression or multi-class classification is consistent with the objectives of this thesis, which aim to investigate whether contrastive learning can effectively differentiate between affective extremes. While more nuanced affect modelling would be valuable, such extensions would require substantial adjustments to the experimental

setup and evaluation framework, diverting focus from the primary research questions. It should be noted that this work focused on learning affect-infused representations building on previous works that had already demonstrated the capacity of unsupervised contrastive learning in modelling basic emotions (Li et al., 2021; Roy and Etemad, 2021). The benefits of formulating affect modelling as a binary classification task are two fold. First the majority of the representation learning approaches is tailored for classification tasks (Le-Khac et al., 2020; Saeed et al., 2021). On the other hand, modelling affect states as binary classification is an approach extensively used in the literature (Makantasis et al., 2021b; Ng et al., 2015; Zhang et al., 2021). Another limitation lies in the use of fixed time window lengths (1, 2, and 3 seconds), which assume that affective changes occur uniformly across these intervals. This may not accurately reflect the dynamic nature of arousal, especially in contexts like gameplay, where affective fluctuations can occur rapidly. Although adaptive or event-based time windowing could provide a more realistic representation of affective dynamics, implementing these techniques would have introduced considerable complexity to the temporal modelling pipeline. Such an investigation would have required extensive optimisation and evaluation, which was outside the scope of this thesis. In summary, while the identified limitations offer avenues for future research, they were not addressed in this work due to resource constraints, and the need to maintain focus on the core research contributions. These limitations, however, provide a solid foundation for future extensions and refinements of the methods developed in this thesis.

Finally, It is worth noting that this thesis does not conduct cross-dataset experiments. All models were trained and evaluated within each dataset independently, and results are reported separately. This choice reflects the fact that RECOLA and AGAIN rely on fundamentally different annotation protocols: RECOLA uses expert annotations bounded in the range $[-1, 1]$ (median of six annotators, $\varepsilon = 0.1$), whereas AGAIN relies on self-reported annotations from a single annotator, normalised to $[0, 1]$ ($\varepsilon = 0.2$). These schemes capture distinct constructs with different reliability characteristics, and therefore direct comparability is neither assumed nor tested in this work.

5.6 | Summary

This chapter introduced a novel approach to affect modelling by leveraging supervised contrastive learning to derive affect-infused representations. Moving beyond traditional end-to-end classification models, we demonstrated how contrastive labels, derived from affect annotations, enable the pre-training of encoder architectures to capture

affective dynamics effectively. The resulting representations were evaluated through downstream tasks, specifically high-low affect state classification, across two diverse datasets: RECOLA and AGAIN. The experimental findings highlight the advantages of SCL-based models, which consistently outperformed baseline models in terms of accuracy, especially for arousal prediction. By utilising various input modalities and time window lengths, we showcased the flexibility and robustness of the proposed methodology. Additionally, the work underscored the role of appropriate contrastive labelling strategies in improving representation quality, particularly when handling subjective annotations. The chapter concludes that contrastive learning is a promising framework for affective computing, providing general-purpose and participant-agnostic representations that address key challenges such as data sparsity and multimodal fusion. These advancements contribute to the broader goals of the thesis by offering a solid foundation for developing robust and scalable affect modelling systems.

6

Learning from Missing Modalities

Affective Computing has made significant progress with the advent of deep learning, yet a persistent challenge remains: the reliable transfer of affective models from controlled laboratory settings (*in-vitro*) to uncontrolled real-world environments (*in-vivo*). This chapter addresses this challenge by introducing the Learning Using Privileged Information paradigm into affective modelling. The LUPI framework leverages fine-grained features and fused data (frames and features) as privileged information during training, while relying solely on image frames during testing. Additionally, Supervised Contrastive Learning is employed to enhance the robustness of *in-vitro* models, whose knowledge is subsequently transferred to *in-vivo* models for improved performance in dynamic, real-world scenarios. Experiments conducted on two benchmark datasets, RECOLA and AGAIN, demonstrate that models trained using privileged information consistently outperform those trained solely on image data. Remarkably, in many cases, these models achieve performance comparable to models trained with access to all modalities during both training and testing. The findings underscore the potential of the LUPI paradigm to bridge the gap between *in-vitro* and *in-vivo* affective modelling, offering a scalable and practical solution for real-world applications.

6.1 | Motivation

Affective Computing has witnessed substantial progress in recent years due to the development of powerful deep learning algorithms (Botelho et al., 2017; Li et al., 2020b; Makantasis et al., 2021a; Toisoul et al., 2021; Trigeorgis et al., 2016; Tzirakis et al., 2021). These advancements have significantly enhanced our ability to model and analyse affective states. However, a critical challenge persists: the reliable transfer of affective models trained on laboratory-collected data (*in-vitro*) to uncontrolled, real-world scenarios

(*in-vivo*). While laboratory settings enable the collection of high-quality data with precise multimodal measurements, these conditions are often far removed from the complexities and unpredictability of real-world environments. In the wild, affective data collection is hampered by numerous environmental and experimental limitations such as privacy concerns, specialised hardware requirements, or limited accessibility. These challenges result in noisy or incomplete data, further widening the disparity between *in-vitro* and *in-vivo* affective modelling. While this gap restricts the deployment of affect models in real-world applications, the *Learning Using Privileged Information* paradigm attempts to mitigate this issue by leveraging additional modalities available during training to guide the learning process. However, it is important to note that LUPI does not directly address the problem of noisy data but instead focuses on bridging the environment gap through the use of privileged information. Addressing these challenges requires innovative approaches to enhance the generalisability and applicability of affective models in dynamic and noisy environments.

This chapter seeks to address the challenges of affect sensing in real-world settings by introducing the concept of privileged information into affect modelling. Specifically, it leverages the LUPI paradigm Vapnik and Izmailov (2015); Vapnik and Vashist (2009), which is particularly well-suited for tasks where the amount and nature of information differs between the training and testing phases of a model. In the context of AC, privileged information plays a critical role in bridging the gap between the development (in-vitro) and deployment (in-vivo) stages of affect models. This work is based on two core hypotheses; a) the LUPI framework can effectively mitigate the limitations of affect sensing in the wild, enabling models to achieve comparable performance in both controlled and real-world environments. b) improving the quality of in-vitro models directly contributes to the development of more effective in-vivo models. To test these hypotheses, experiments are conducted using two widely applied multimodal affect datasets: RECOLA and AGAIN (Melhart et al., 2021a; Ringeval et al., 2013)

In particular, to address the first hypothesis, we evaluate the LUPI paradigm by utilising fine-grained features and fused data (combining frames and features) as privileged information during the training phase (in-vitro), while relying solely on frames as the prevalent information during testing (in-vivo). This choice is justified by the fact that in real-world scenarios, the collection of additional features (such as physiological signals or privileged modalities) is often impractical due to privacy concerns, and hardware limitations. Additionally, this approach allows us to explore how the inclusion of richer, high-dimensional data during training enhances the model's ability to generalise to real-world scenarios where only frames are available. For the second hypothesis, we employ SCL, as described in chapter 5, to enhance the performance of in-vitro models.

The knowledge of these improved in-vitro models is then transferred to in-vivo models, enabling them to leverage the enriched representations and perform more effectively in dynamic, real-world conditions. Our experiments demonstrate that incorporating privileged information during training leads to models that consistently outperform those trained solely on image data. On top of that boosting the performance of the privileged models (models that use solely privileged information) through SCL leads to in-vivo models of higher capacity. Notably, in many scenarios, models trained using the LUPI paradigm achieve performance comparable to models that rely on the fusion of all available modalities (images and privileged information). These results hold for both player modelling (AGAIN) and dyadic interactions (RECOLA).

6.1.0.1 | Contributions

This chapter makes several key contributions to advancing affect modelling in real-world settings. First, we introduce the LUPI paradigm into the domain of AC as a solution to mitigate the challenges associated with transferring models from controlled (*in-vitro*) to uncontrolled (*in-vivo*) environments. Specifically, we employ fine-grained features and fused data (frames and features) as privileged information available exclusively during training, while using raw footage frames as the prevalent information accessible during testing. This approach enables the models to leverage richer representations during training, enhancing their generalisability to real-world conditions. Second, to further strengthen the *in-vitro* models, we utilise Supervised Contrastive Learning to improve their capacity for learning robust and transferable representations. These refined *in-vitro* models are subsequently transferred to the *in-vivo* models, enabling better performance, especially in dynamic environments such as games.

Third, we validate our approach using two established multimodal affect datasets: RECOLA and AGAIN. Both datasets provide annotations of affective dimensions such as arousal and valence, along with multimodal data. By treating raw footage as prevalent information and the remaining modalities as privileged information, we systematically evaluate the impact of the LUPI paradigm on affective modelling. This demonstrates the robustness of the proposed approach in both datasets. These findings highlight the critical role of privileged information in bridging the gap between *in-vitro* and *in-vivo* affect modelling, offering a pathway to creating more robust, accessible, and practical models for real-world applications.

6.2 | Modelling Methodology

This section first describes the use of privileged information with neural network models in a concise way and moves on to outline the model architectures employed.

6.2.1 | Learning Using Privileged Information

As outlined in Section 3.3, Learning Using Privileged Information Vapnik and Izmailov (2015); Vapnik and Vashist (2009) addresses problems characterised by an asymmetric distribution of information between training and test time. LUPI provides the means to *transfer knowledge* from all the available modalities to a machine learning model that makes predictions using only a subset of these modalities Lopez-Paz et al. (2016); Sharmanska et al. (2013). As far as the RECOLA database is concerned, we treat as privileged the information that corresponds to physiology and audiovisual features provided by the database creators (see Section 4.1). For AGAIN, we consider fine-grained gameplay features as privileged information (see Section 4.2). Our choice is justified by the fact that capturing physiology requires specialized sensors while constructing physiology and audiovisual features implies the employment of specific software algorithms. In the same vane, information about the state of the game (number and actions of enemies) requires access to the game engine itself. On the contrary, information that comes from raw footage frames is considered prevalent information due to the fact that it can be captured using conventional cameras that can be available both at training and test times.

It is worth clarifying that transferring knowledge using LUPI is different from the transfer of learning techniques used in deep learning (Ng et al., 2015). Transfer of learning targets small-sample setting problems by finetuning a model trained for a specific task such that it performs well in a similar task. On the contrary, using LUPI focuses on problems with asymmetric distribution of training/testing information and trains the models from scratch. This study explores the use of privileged information with neural network-based models of affect. Following Hinton et al. (2015); Lopez-Paz et al. (2016); Vapnik and Izmailov (2017), we represent that knowledge within the output of a neural network that has been trained and makes predictions based on all available modalities or on privileged information only. This model is called *teacher*. Having a teacher model trained, we can transfer knowledge from privileged information to another model called *student*. The transfer of knowledge can be achieved by feeding the model with only those modalities of information that are available in the wild and forcing it during training to balance between the learning task's loss and learning prediction

Table 6.1: Information modalities used for training and testing the different models. ✓ indicates available modalities, ✗ indicates modalities that are unavailable and - corresponds to modalities that do not exist in the corresponding dataset.

	Pixel		Audio		Visual		ECG		EDA		Game	
RECOLA	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline Model	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	-	-
Fusion Teacher	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-
Privileged Teacher	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	-	-
Student Model	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	-	-

	Pixel		Audio		Visual		ECG		EDA		Game	
AGAIN	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline Model	✓	✓	-	-	-	-	-	-	-	-	✗	✗
Fusion Teacher	✓	✓	-	-	-	-	-	-	-	-	✓	✓
Privileged Teacher	✗	✗	-	-	-	-	-	-	-	-	✓	✓
Student Model	✓	✓	-	-	-	-	-	-	-	-	✓	✗

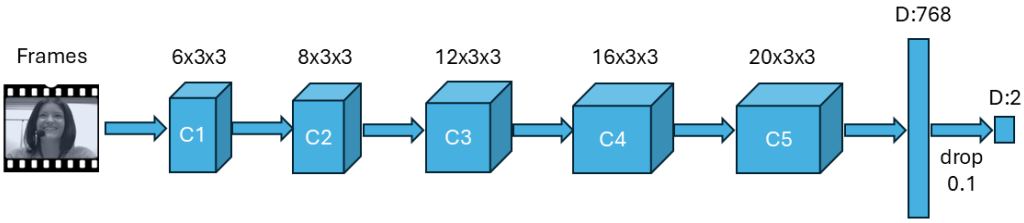
distributions that match those of the teacher model. After training, the student model makes predictions based only on the information that is available in the wild, without any dependence on the teacher model or privileged information. Hence the resulting optimisation objective is the same as the one described in equation 3.2

6.2.2 | Model Architectures

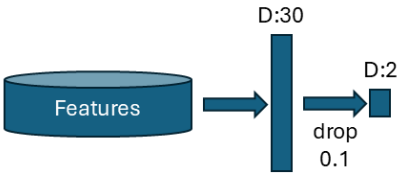
In this section, we first present the main components used in LUPI, namely student and teacher models. Then, we present the baseline architectures that we compare against our method for assessing the effectiveness of the obtained models. It should be noted that both the student and the teacher architectures are illustrated in Figure 6.1

6.2.2.1 | Student Model

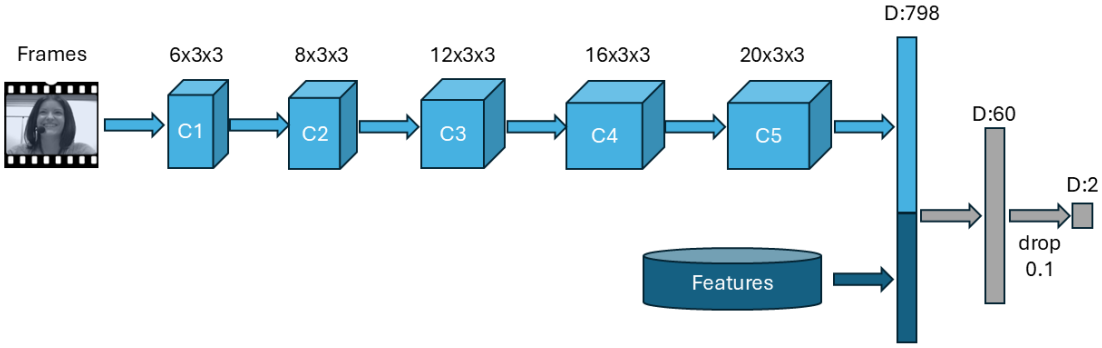
The student neural network S is a predictive model that learns to estimate the target outcomes by utilising the information from frames. During the training phase, the student model leverages knowledge distilled from a teacher model T, which has access to privileged information. This privileged information, available only during training, enhances the learning process by providing auxiliary guidance. The goal of the student model is to effectively generalise to new, unseen data, where privileged information is not available, using only the primary feature set. By distilling the teacher's insights, the student model is expected to achieve improved performance compared to learning



(a) Student



(b) Privileged Teacher



(c) Fusion Teacher

Figure 6.1: Illustration of the design of the student model and two teacher variants (Privileged Teacher and Fusion Teacher) used in this chapter. The Student Model (a) processes input frame sequences through a series of convolutional layers (C1 to C4) extracting visual features. These features are subsequently passed through a fully connected layer D, incorporating dropout regularisation, to produce the final output vector for prediction. The Privileged Teacher (b) passes the precomputed feature vectors through fully connected layers D, which include dropout to produce the output. Finally, the Fusion Teacher (c) follows a similar structure to the student model but includes an additional pathway for processing feature vectors. The visual information is fused with the precomputed feature vectors via late fusion. This combined representation is further processed through fully connected layers, and dropout, to yield the final output.

solely from the primary features. In this chapter, we employ a CNN model of five convolutional and one dense layer activated bye ReLU. The hyperparameters of the layers are the same as the ones of the Frame Encoder used in Section 5.2.1.1. The last layer (decision layer) of the model is a dense layer of 2 neurons activated by SoftMax. It should be noted that a dropout of 0.1 is applied before the decision layer.

6.2.2.2 | Teacher Models

The teacher model T is a predictive model that is trained with access to both the primary feature set (prevalent information) and an additional set of privileged information. The teacher model is typically designed to achieve high performance by leveraging this richer information. Once trained, the teacher model's knowledge is distilled and transferred to the student model S, guiding S to improve its predictions based on the primary feature set alone. The teacher model plays a crucial role in the LUPI framework by acting as a source of enhanced supervision during training. In this work, we consider two teacher architectures.

Privileged Teacher: The privileged teacher P considers only feature information (privileged information). Consequently, it is a simple ANN consisting of two layers. The first layer has 30 neurons and it is activated by Sigmoid (see **Feature Encoder** of Section 5.2.1.1). The output of this layer is fed to the decision layer which is a simple 2-neuron dense SoftMax-activated layer. Once again a dropout of 0.1 is applied before the last layer.

Fusion Teacher: The fusion teacher F has access to both privileged and prevalent information (features and frames). It follows the same architecture as the **Fusion Encoder** of Section 5.2.1.1. The output of this architecture is fed to a 2-neuron dense layer which is activated by SoftMax. A dropout of 0.1 is applied before the output layer.

6.2.2.3 | Baselines

There are two baseline architectures employed in this work corresponding to student models that do not consider privileged information at all. The first one performs end-to-end affect state classification from frames, bypassing the representation learning process and is denoted as E^B. The second baseline E^{SC}, first trains the latent dimension of the student model via SCL and then probes the frozen learned embeddings with a SoftMax-activated layer as the one used in all models of this chapter.

6.3 | Data Preprocessing

This section presents the datasets used for experimentally validating our proposed methodology and the data preprocessing steps.

6.3.1 | Preprocessing RECOLA

As described in the previous chapter, the RECOLA database provides both arousal and valence annotations. The same preprocessing pipeline is applied to both affective dimensions. Each participant's session is segmented into overlapping time windows using a sliding step of 400 ms and window lengths of 1, 2, and 3 s. These hyperparameters determine the dataset size and the temporal granularity of the information. Features are averaged within each window to create a single feature vector, while frame sequences are downsampled to 5 grayscale frames per second, with dimensions 224×224. For example, a 3 s time window includes a single feature vector and a frame tensor of size $15 \times 224 \times 224$. Median annotation values are used for arousal and valence to mitigate inter-annotator disagreement. The arousal and valence state scores, g_a and g_v, are computed based on the median traces and are bounded within $[-1, 1]$, reflecting the original affect scale. The detailed preprocessing pipeline can be found in Section 5.3.1.

6.3.2 | Preprocessing AGAIN

As outlined in the previous chapter, the AGAIN dataset provides self-reported arousal annotations, which are processed separately for each of the three games. Each participant's session is segmented into overlapping time windows using a sliding step of 500 ms and window lengths of 1, 2, and 3 s. The arousal traces are normalised within $[0, 1]$ to address the unbounded nature of the PAGAN annotations (Melhart et al., 2019). Following segmentation, each time window contains a sequence of feature vectors and a sequence of frames. To reduce computational complexity, features are averaged within each window, yielding a single feature vector. For frame sequences, 5 grayscale frames per second with dimensions 224×224 are retained. As the annotations are self-reported, there is no need to mitigate inter-annotator disagreement. The arousal state score g_a is computed using the corresponding arousal trace for each session based on Eq. (5.1). This preprocessing approach is repeated consistently for all three games in the dataset. Once again, the detailed preprocessing pipeline has been described in Section 5.3.2.

6.4 | Results

This section presents the framework for evaluating the impact of privileged information on affect modelling and the experimental results obtained. RECOLA and AGAIN should be regarded as measuring distinct affective phenomena, given their divergent annotation protocols: RECOLA employs expert-provided affect traces, while AGAIN

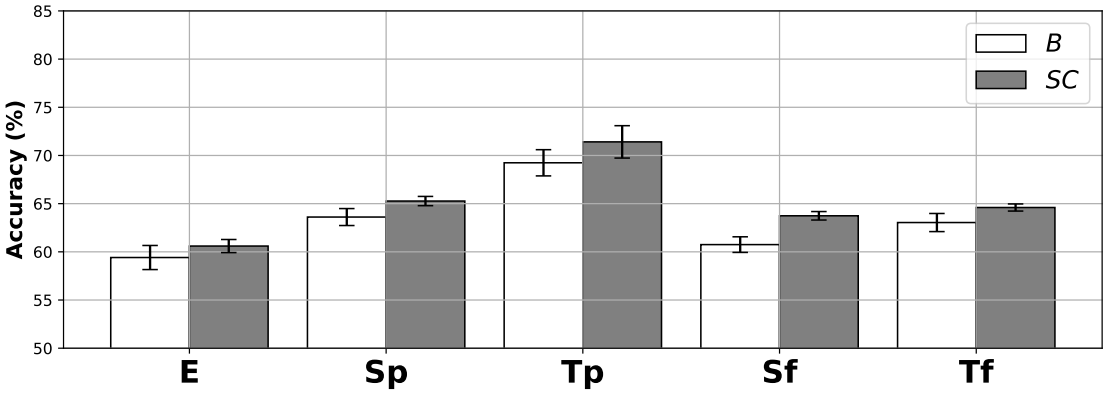
relies on single self-reports. These differences extend to scale, reliability, and normalisation procedures. For this reason, datasets were always analysed independently, with models trained per game in AGAIN and per affective dimension in RECOLA. Reported comparisons are therefore intended to demonstrate methodological robustness within each dataset, not to imply equivalence of absolute annotation values across them.

6.4.1 | Evaluation Framework

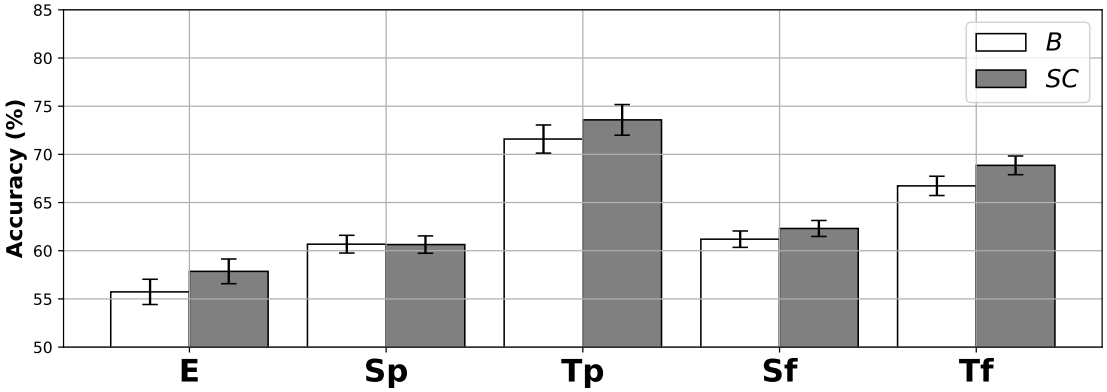
In this chapter, we follow the same protocol as the one described in Section 5.4.1. As mentioned earlier, we treat affect modelling as a classification task. Hence our models attempt to predict high and low affect (arousal or valence) states. The methodology for producing affect state labels is described in Section 5.2.3.1 and the size of the resulted datasets is shown in Tables 5.1 & 5.2. To evaluate the models' performance we follow a 5-fold cross-validation scheme. When splitting the dataset, we do not include data from the same participant in both training and test sets. before every train/test split, we hold out 10% of the participants as a validation set to apply early stopping criteria and avoid model overfitting; training stops after 5 epochs without loss improvement on the validation set. All employed models of affect are evaluated using precisely the same data, i.e., the training, the validation and the test sets are the same for all models. Finally, for affect classification, we report the models' performance in terms of binary classification accuracy. Arguably one of the most important elements of the LUPI optimisation objective is to set a proper value for the hyperparameter α that controls the influence of the teacher to the training process (Eq 3.3). In this study, we followed a grid-search protocol to tune the hyperparameter α (Appendix B) and then used the best values for the remaining experiments. It is worth noting that any claims regarding significance are based on a pairwise t-test $p < 0.05$.

6.4.2 | The Importance of Privileged Information

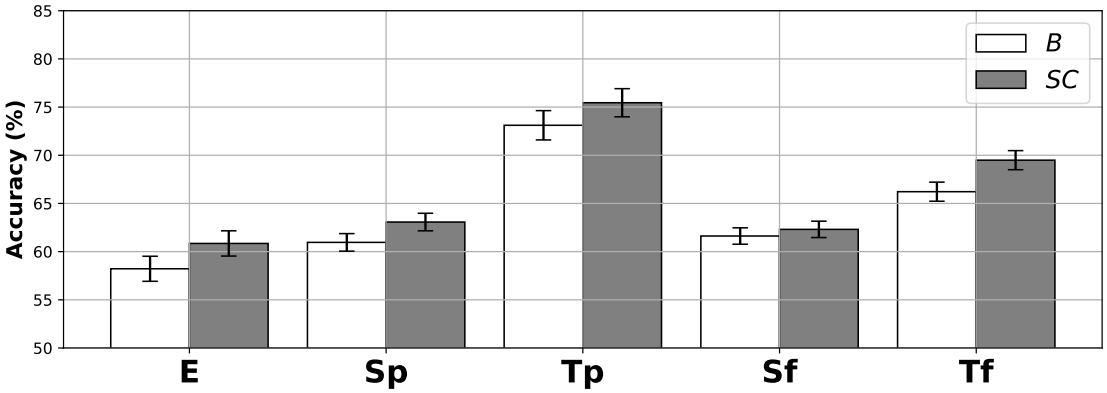
As mentioned above, the student models make predictions using solely information that is available in the wild; in our case the raw footage frames The student models trained via LUPI are denoted as S_p and S_f and have access to privileged information through the Privileged Teacher and Fusion Teacher respectively. We compare the student models' performance against the performance achieved by the Fusion Teacher (F) model that uses all modalities (features & frames) for training and testing, and privileged teacher (T_p) that makes predictions using only privileged information. For the following investigation, we use non-zero α parameter values that yield the most accurate student



(a) RECOLA: Arousal 1s

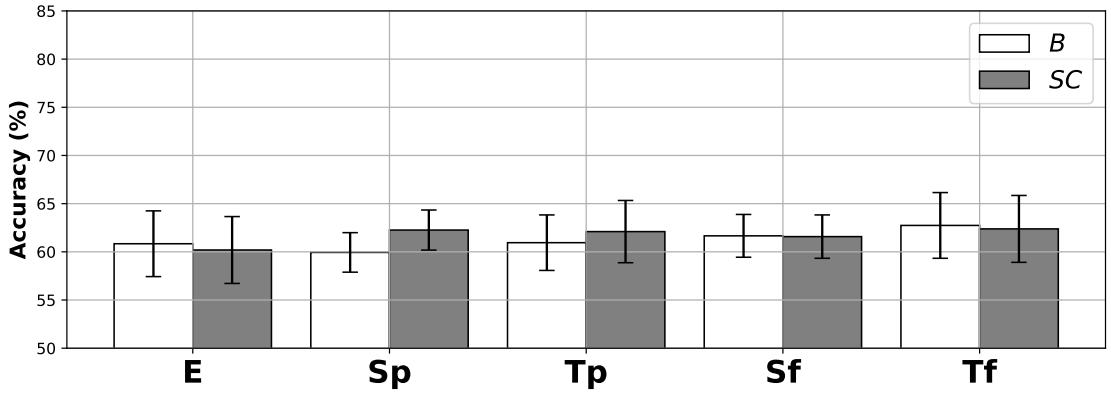


(b) RECOLA: Arousal 2s

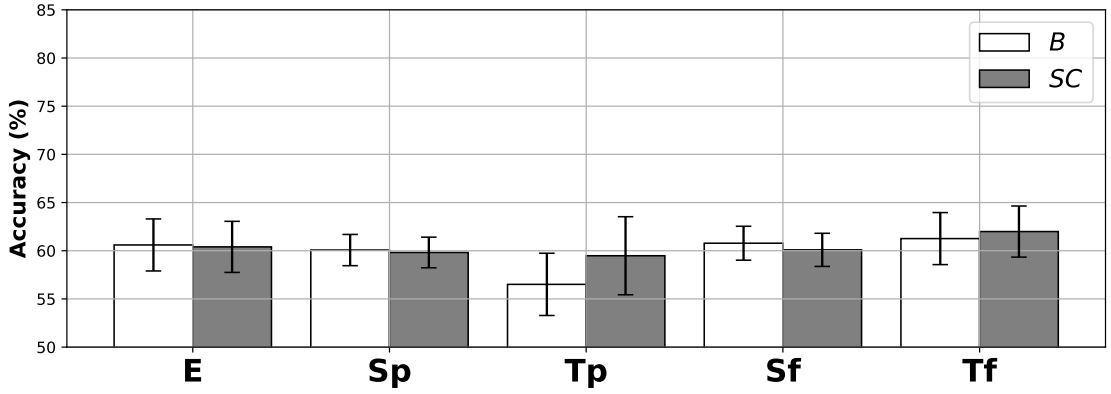


(c) RECOLA: Arousal 3s

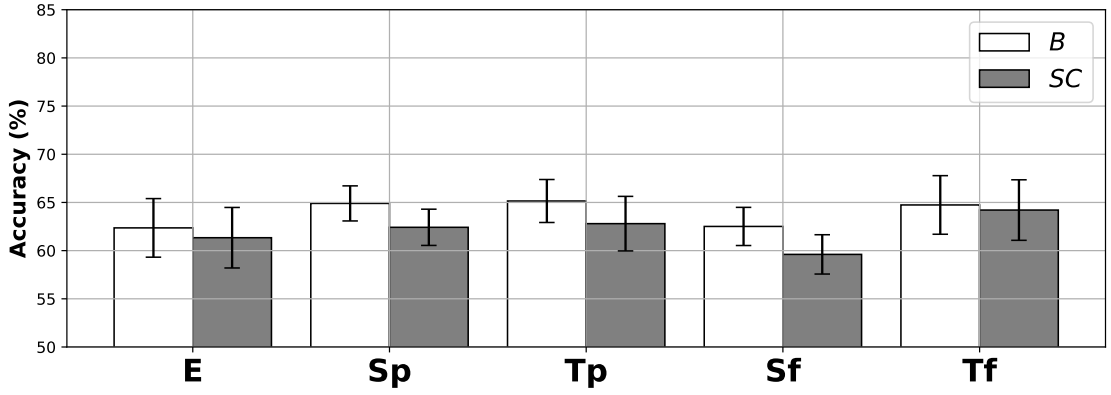
Figure 6.2: **RECOLA Dataset** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) RECOLA: Valence 1s

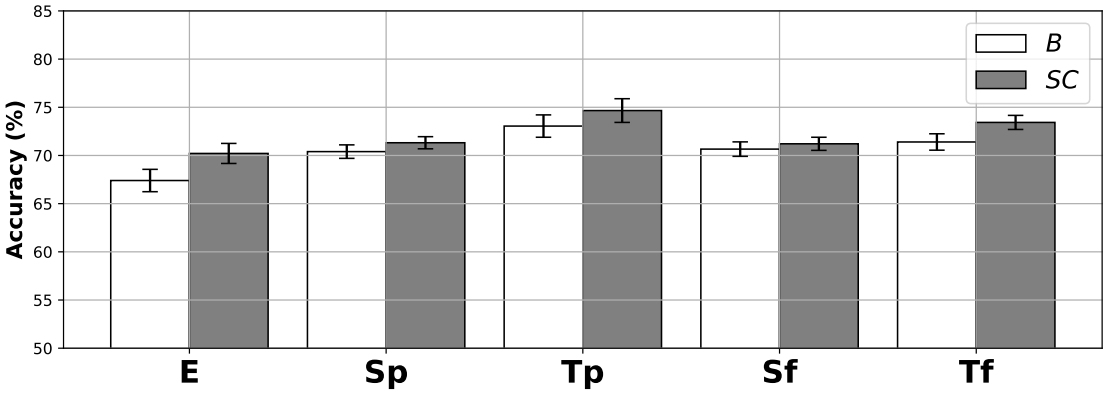


(b) RECOLA: Valence 2s

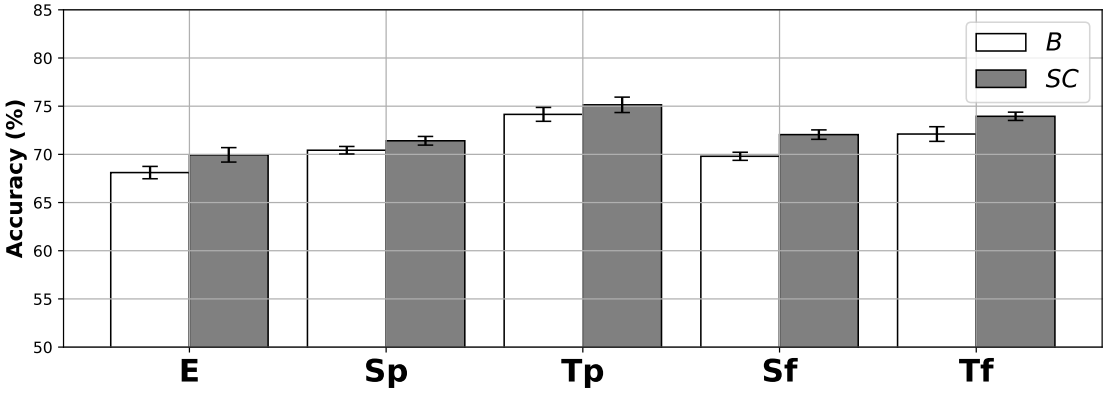


(c) RECOLA: Valence 3s

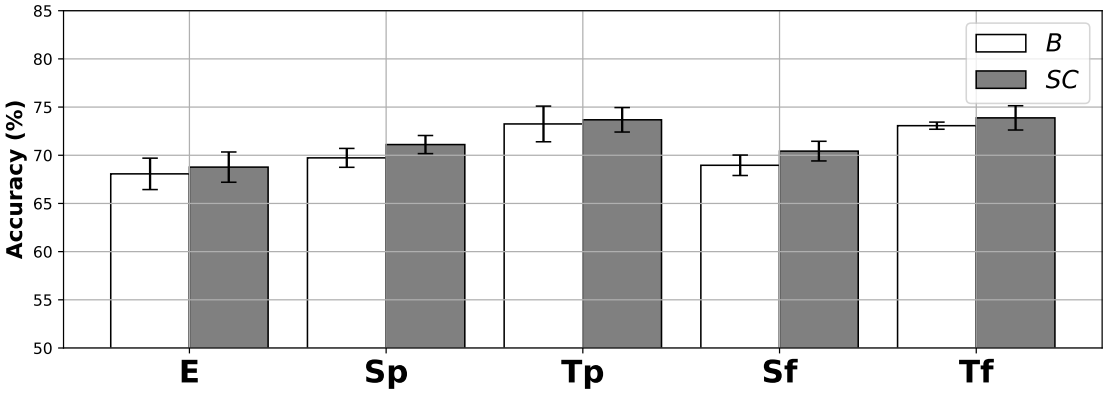
Figure 6.3: **RECOLA Dataset** Average 5-fold validation accuracy scores (%) for high-low valence classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) *Run'N'Gun!* 1s

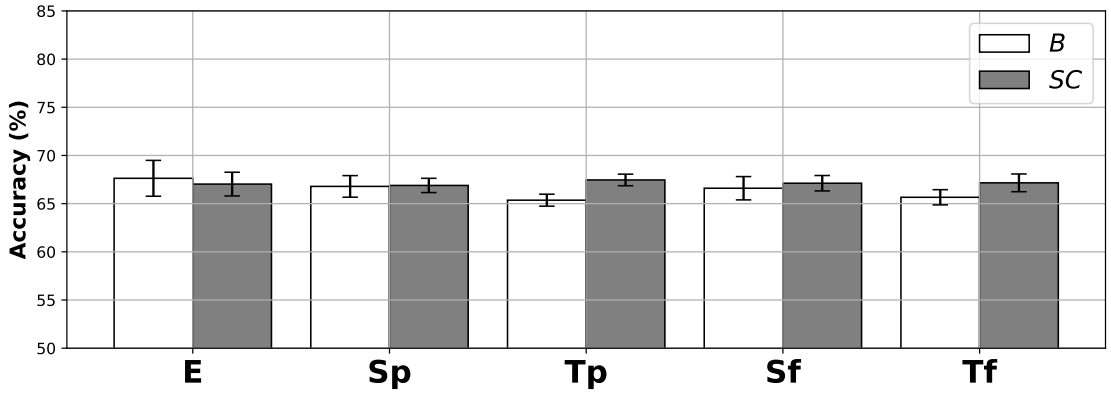


(b) *Run'N'Gun!* 2s

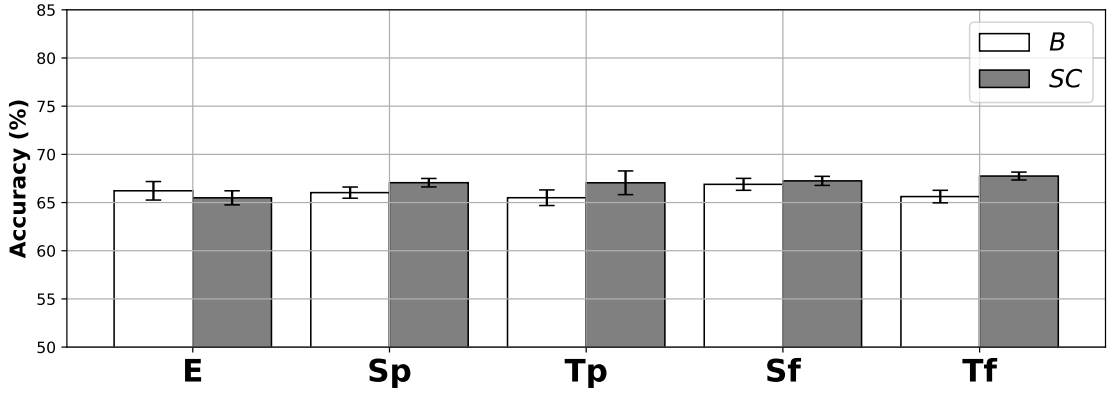


(c) *Run'N'Gun!* 3s

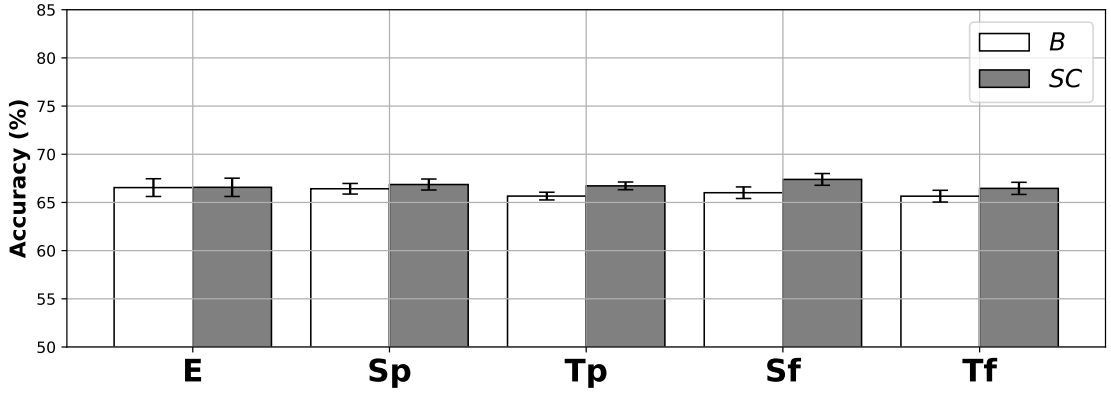
Figure 6.4: **AGAIN Dataset *Run'N'Gun!*** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) *Pirates!* 1s

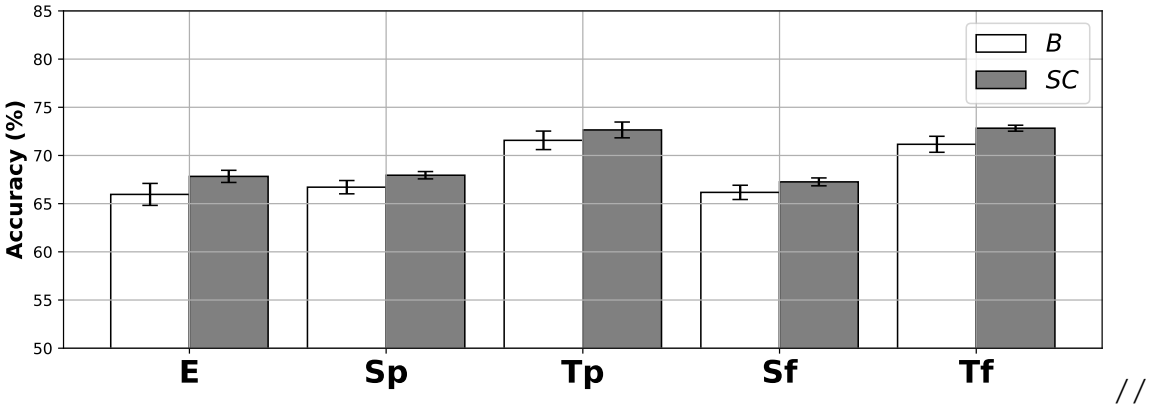


(b) *Pirates!* 2s

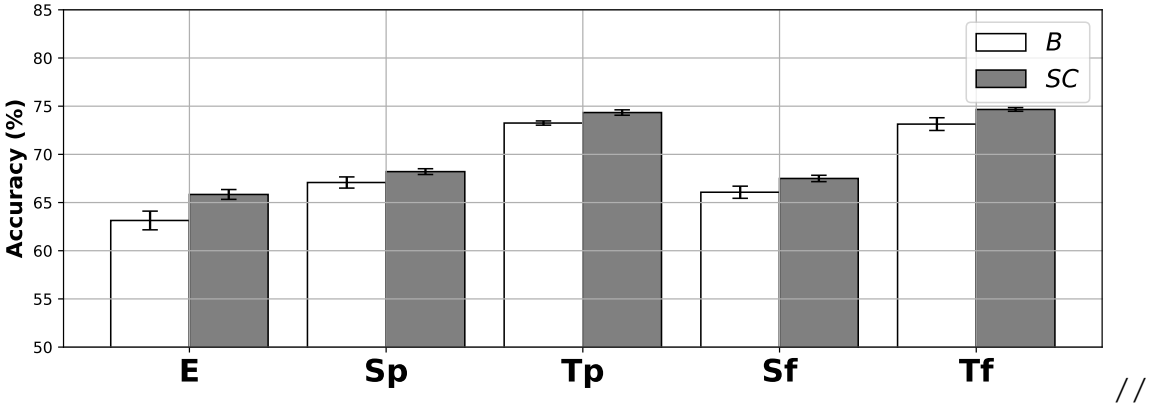


(c) *Pirates!* 3s

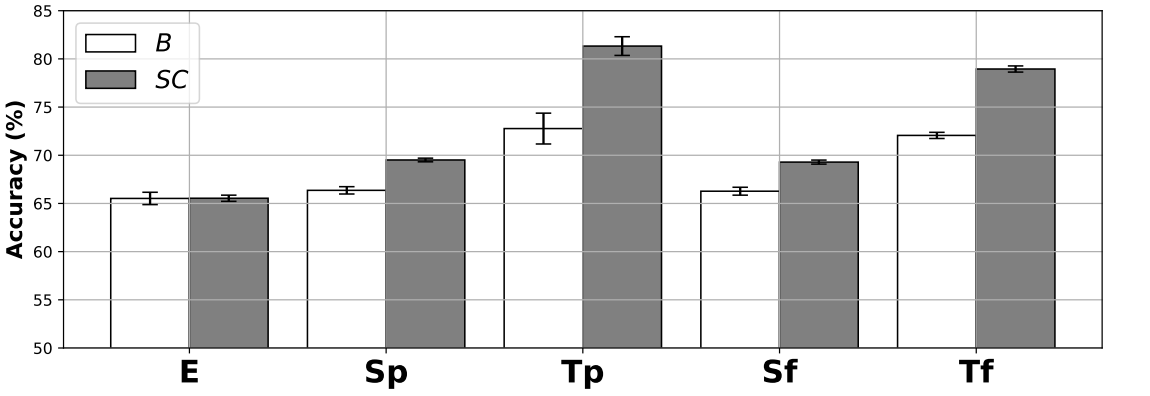
Figure 6.5: **AGAIN Dataset *Pirates!*** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.



(a) *Endless* 1s



(b) *Endless* 2s



(c) *Endless* 3s

Figure 6.6: **AGAIN Dataset *Endless*** Average 5-fold validation accuracy scores (%) for high-low arousal classification. Values are averaged across 5 independent runs; 95% confidence intervals are displayed as error bars.

models based on the sensitivity analysis covered in Appendix B. Moreover, we repeat the 5-fold cross-validation scheme (see Section 6.4.1) five times after reshuffling the participants, for evaluating the significance of our results. For the sake of completeness, we also present the performance of baseline models (E) that do not utilise privileged information at all and correspond to the frame models of the previous chapter. Figures 6.2- 6.6 showcase the performance of the models in high vs low affect state classification. The white bars correspond to models that have been trained in an end-to-end manner while the gray bars indicate that SCL has been a part of the training process. In the case of E^{SC}, ${T_p}^{SC}$, ${T_f}^{SC}$ SCL directly trains the representations of the models and then a linear probe on top of the resulting representations acts as the affect state classifier. The ${S_p}^{SC}$ and ${S_f}^{SC}$ correspond to the student models that are trained via the LUPI framework (Eq 3.3) under the guidance of the SCL-trained teachers ${T_p}^{SC}$, ${T_f}^{SC}$, respectively.

Figures 6.2 and 6.3, report the average 5-fold validation accuracy for high-low arousal and valence classification tasks on the RECOLA dataset. The results are analysed across different time windows (1s, 2s, 3s) and model types. The models evaluated include baseline models, teacher models (Fusion and Privileged Teachers), and student models trained via the LUPI framework. The impact of SCL is also investigated. The baseline models, denoted as E, represent frame-based classifiers trained without privileged information. Across all six RECOLA subplots, baseline models consistently achieve low accuracy scores, indicating that models relying solely on raw frame data are unable to achieve competitive performance. This highlights the limitations of training models without incorporating privileged information or additional learning signals. The teacher models, namely the Fusion Teacher (T_f) and the Privileged Teacher (T_p), achieve significantly higher accuracy compared to the baseline models. The Privileged Teachers (T_p), which considers only privileged features consistently achieve the highest accuracy across the arousal classification tasks, establishing the upper bound for model performance. On the other hand, the Fusion Teachers (T_f), which use privileged information and frames, achieve competitive results but perform slightly worse than the T_p. The performance between T_p and T_f highlights the importance of integrating both privileged information and raw frames during model training for achieving optimal accuracy.

The LUPI-trained student models S_p and S_f demonstrate substantial improvements over the baseline models. The S_f model, which is trained under the supervision of the Fusion Teacher (T_f), performs on par and in some cases outperforms the S_p model, which is guided by the Privileged Teacher (T_p). This result suggests that the inclusion of frames in the Fusion Teacher can allow for better information transfer during the LUPI training process. SCL further enhances the performance of the evaluated models. In particular, models trained with SCL, consistently outperform their end-to-end counter-

parts. For the teacher models, ${T_f}^{SC}$ and ${T_p}^{SC}$ achieve higher accuracy compared to ${T_f}^{B}$ and ${T_p}^{B}$. A similar trend is observed for the student models trained. Both ${S_p}^{SC}$ and ${S_f}^{SC}$ perform on par and in some cases outperform E^B,

While significant improvements are observed for arousal classification, valence classification results remain relatively stable between models, training strategies, and time window lengths, as shown in Figures 6.2 and 6.3. This can be attributed to the inherent challenges of valence prediction, which relies on subtle and ambiguous cues that are less dynamic and harder to capture than arousal. Valence-related expressions often appear visually similar across emotional contexts, making it difficult for models using raw frames to learn discriminative features. Additionally, while privileged information provides strong supervisory signals for arousal, it may fail to capture sufficient valence-specific information, leading to weaker guidance from the Privileged Teacher (T_p) and Fusion Teacher (T_f). This suggests that valence recognition requires additional contextual information, such as semantics or interaction dynamics. Overall, the lack of improvement highlights the limitations of current approaches in learning valence cues from visual modalities and the need for richer, context-aware information to improve performance.

Figures 6.4-6.6 present the arousal classification results for the AGAIN dataset, which includes three platformer games: *Run'N'Gun!*, *Pirates!*, and *Endless*. Each game is evaluated for time windows of 1s, 2s, and 3s. In *Run'N'Gun!* (Fig. 6.4), T_p achieves the highest performance, establishing an upper bound. Among the LUPI-trained student models, ${S_f}^{B}$ and ${S_p}^{B}$ perform on par, outperforming the baseline models E^B. SCL further improves the performance of the student models. In particular, the best student model (bold values) outperforms the baseline models in all experimental settings while performing on par with or surpassing the remaining end-to-end students. For *Pirates!* (Fig. 6.5), performance remains relatively stable over all time windows (1s, 2s, and 3s), with limited improvements as the temporal context increases. This stability suggests that arousal cues are less dynamic and harder to capture due to subtler changes in gameplay. In this case, both teachers perform on par, as do S_f and S_p among the student models. Furthermore, the gains from SCL are smaller compared to the previous game, with ${S_p}^{SC}$ and ${S_f}^{SC}$ significantly outperforming E^B and E^{SC} at 2-second time windows. Notably, in *Pirates!*, the teacher models perform on par with the baseline models, meaning there is no additional information gain to be exploited by the LUPI framework.

Finally, in *Endless* (Fig. 6.6), the end-to-end teachers perform on par while ${T_p}^{SC}$ performs significantly better than ${T_f}^{SC}$ at 3-second windows. Furthermore, the LUPI-trained student models (S_f and S_p) perform on par, while ${S_p}^{B}$ and ${S_f}^{B}$ underperform compared to their SCL counterparts. The ${S_p}^{SC}$ and ${S_f}^{SC}$ models significantly outper-

form E^B at all time windows. In the same vein, ${S_p}^{SC}$ and ${S_f}^{SC}$ outperform E^{SC} for 2- and 3-second windows, while performing on par at 1 second. The results from RECOLA and AGAIN demonstrate the effectiveness of the LUPI framework and SCL for arousal classification. In RECOLA, significant improvements are observed for arousal, while valence remains challenging due to its reliance on subtle and ambiguous visual cues. In AGAIN, performance varies by game, but student models benefit from privileged information regardless of teacher and consistently outperform the end-to-end baseline E^B, in some cases reaching or exceeding the performance of the SCL baseline E^{SC}. Similarly, students trained under contrastive teachers ${T_f}^{SC}$ and ${T_p}^{SC}$ perform on par with, and at times better than, E^{SC}. These findings highlight the critical role of privileged supervision and contrastive learning in addressing the challenges of affect state classification.

6.5 | Discussion

The findings presented in this chapter underscore the potential of the LUPI framework and SCL for leveraging privileged information in affect modelling. By focusing on the binary classification of high versus low affect states, the models demonstrated improved robustness and performance within both the RECOLA and AGAIN datasets. However, several methodological limitations and considerations must be acknowledged. The study employed a grid-search protocol to tune the hyperparameter α, which controls the influence of the teacher in the LUPI framework. While this approach ensured optimal performance for the chosen datasets, it may not generalise to other datasets or applications. The fixed hyperparameter tuning protocol also assumes a one-size-fits-all solution, potentially overlooking dataset-specific nuances. Moreover, the hyperparameter tuning was performed for time windows of 1 second, assuming that the optimal α value remains the same for all time windows. Although advanced techniques such as Bayesian optimisation or population-based training could offer more efficient and potentially more effective approaches to hyperparameter tuning, computational complexity and time requirements associated with these methods were deemed impractical given the resource constraints and scope of this project. Future work could investigate the integration of adaptive hyperparameter optimisation methods to balance efficiency and performance for different datasets.

On top of that the manual selection of privileged features represents a key limitation of this study. While the selected features were thoughtfully aligned with domain knowledge and task-specific objectives, this manual approach may have overlooked complex, latent structures within the datasets that automated methods could exploit.

Techniques such as automated feature selection or deep learning-based joint feature extraction could uncover richer and more nuanced representations of privileged information, enhancing the model's performance and robustness. For example, attention mechanisms or feature-ranking algorithms could dynamically identify and prioritise the most relevant features, revealing subtle patterns that manual curation might miss. However, implementing such methods would have introduced additional computational complexity and resource demands, potentially detracting from the study's primary focus on rigorously evaluating the benefits of the LUPI paradigm. Thus, while manual selection ensured interpretability and alignment with human understanding, future work could explore automated techniques to further advance the framework's scalability and effectiveness.

This study integrates the LUPI paradigm within the supervised affect modelling and representation learning framework, providing a foundation for enhancing affective computing models with privileged information. While the current work focuses on binary classification tasks, a natural extension would be to explore the potential of LUPI in semi-supervised and self-supervised learning paradigms. Such extensions could unlock the ability to learn powerful general-purpose representations that are better suited for diverse affect modelling tasks, especially in scenarios with limited labelled data. Semi-supervised approaches could leverage unlabelled data alongside privileged information to enhance representation learning, while self-supervised methods could enable models to discover intrinsic data structures without explicit labels. These paradigms would be particularly beneficial for scaling affective computing systems in real-world applications where data labelling is often costly and time-consuming. Another promising direction involves extending the application of LUPI to ranking and preference learning paradigms, which are essential for ordinal affect modelling tasks. Unlike binary classification, learning-to-rank models aim to capture the ordinal nature of affective states, such as levels of arousal or valence, as seen in preference-based affect modelling studies (Makantasis, 2021; Yannakakis et al., 2017, 2018). The incorporation of LUPI into these paradigms could enable models to better exploit privileged information to infer subtle affective gradients, thereby improving their ability to model subjective human experiences. However, such extensions would require significant methodological adaptations. These adaptations, while outside the scope of the current study, represent exciting opportunities for advancing the field and addressing more complex affective computing challenges in future research.

Finally, it is worth clarifying that this thesis does not include cross-dataset evaluations. Each dataset was analysed independently, with models trained and tested within its own scope and results reported separately. This decision reflects the distinct annota-

tion protocols employed: RECOLA is based on expert traces bounded in $[-1, 1]$ (median of six annotators, $\varepsilon = 0.1$), whereas AGAIN relies on single self-reports normalised to $[0, 1]$ ($\varepsilon = 0.2$). As these protocols capture different constructs and exhibit varying reliability characteristics, direct comparability between datasets is neither assumed nor attempted.

6.6 | Summary

In this chapter, we investigated a methodology for affect modelling in real-world scenarios by leveraging privileged information and teacher pre-training via supervised contrastive learning. Our central hypothesis posits that privileged information can facilitate the reliable transfer of affect models from controlled environments, where abundant high-quality data is available, to real-world settings. To evaluate this hypothesis, we utilised the RECOLA and AGAIN datasets, which include both raw visual data and high-level handcrafted features. We considered all handcrafted features as privileged information, meaning they are accessible solely during model training, while raw visual data corresponding to video frames remains available during both training and testing. Within this framework, affect modelling was treated as a classification task, focusing on predicting high vs low arousal and valence. To enhance model performance, we pretrained teacher models—those with access to privileged information during both training and testing—via SCL and subsequently transferred their knowledge to student models, which operate without privileged information during testing. This approach assumes that teachers with a higher predictive power can produce more robust student models. Our findings demonstrate that affect models trained with privileged information perform as well as, or even surpass, their teacher models. Crucially, for the field of affective computing, these models do not rely on costly, intrusive, or impractical data modalities during real-world deployment. The proposed knowledge transfer methodology, which aligns student models with teacher predictions, has broad applicability to affective modelling tasks involving multimodal data, particularly when the system must operate in real-world environments or with missing modalities.

7

Affect Modeling with Limited Data

Affective computing, particularly affect modelling with limited data, presents a significant challenge as traditional AI models typically rely on large, labelled datasets to predict emotional states accurately. This challenge is particularly pronounced in dynamic environments such as video games, where user emotions are complex and context-dependent, and in multimodal datasets which require diverse contextual understanding. The difficulty in obtaining sufficient labelled data in such scenarios impedes the development of generalisable models across varied domains and interactions (Section 7.1). This chapter introduces a novel framework that decomposes affect modelling into simpler, domain-specific tasks solvable using few-shot learning (Section 7.2). Through experiments conducted on both the *GameVibe* and *RECOLA* datasets, we show that few-shot learning models outperform traditional approaches, offering a more effective and scalable solution for affective computing with limited data (Section 7.4).

7.1 | Motivation

Affect modelling involves creating systems that infer and predict emotional and cognitive states of humans, playing a critical role in human-computer interaction, user experience research, and intelligent system design (Picard, 2000). A key challenge in affect modelling is ensuring that models generalise across diverse contexts and domains, where emotional states are influenced by factors such as environment, task, or individual differences. This challenge becomes particularly pronounced when labelled affective data is limited, as affective states are inherently subjective, context-dependent, and difficult to annotate in large quantities. The complexity of labelling arises from the fact that human emotions lack clear boundaries and are often perceived differently depending on situational and personal factors.

Domain generalisation seeks to address these challenges by developing models that can generalise across multiple domains within the same task—such as emotion recognition across different games or environmental settings—while maintaining robust performance under varying conditions (Zhou et al., 2022). However, traditional machine learning methods, which rely heavily on large quantities of labelled data drawn from a single distribution, often struggle when faced with shifts in data distributions between training and testing. When the underlying domains differ significantly, these models experience performance degradation. This limitation is especially pronounced in affect modelling, where data typically spans diverse, context-rich environments with significant variability.

Furthermore, the process of collecting vast amounts of labelled affective data is highly impractical due to the extensive cognitive effort required for annotation. The subjective nature of affect further complicates this process, as inter-rater agreement can be low, necessitating careful curation and validation of labelled datasets. As a result, developing approaches that mitigate the need for large labelled datasets while ensuring strong generalisation capabilities is crucial. Few-Shot Learning (Wang et al., 2020) has emerged as a promising solution to this problem, offering techniques that enable models to learn effectively from a limited number of labelled examples. By leveraging meta-learning, metric-based approaches, and task adaptation strategies, FSL facilitates generalisation even in low-data scenarios, making it particularly relevant for affect modelling applications where labelled data is scarce.

Although FSL techniques have led to significant advancements in objectively-defined tasks such as object recognition Li et al. (2017a), their application in affect modelling remains underexplored. Affect modelling involves handling a wide range of emotional and cognitive states, making it a more complex and nuanced challenge. Furthermore, the generalisation problem becomes even more pronounced in dynamic, multimodal environments where affective responses are influenced by numerous interacting factors. Video games represent one such application area, offering a rich platform for studying affective states through diverse stimuli, including audio, visuals, narrative elements, and gameplay mechanics (Yannakakis and Melhart, 2023). They provide unique opportunities to examine user experiences such as engagement while simultaneously posing significant challenges in terms of data collection and model generalisation (see Figure 7.1). These characteristics make video games a suitable testbed for few-shot learning approaches, as they naturally encompass a wide range of contextual factors that are crucial for studying affective responses across multiple domains.

Motivated by the limited research on few-shot domain generalisation for affect modelling, this chapter proposes a framework for multidomain few-shot affect modelling.

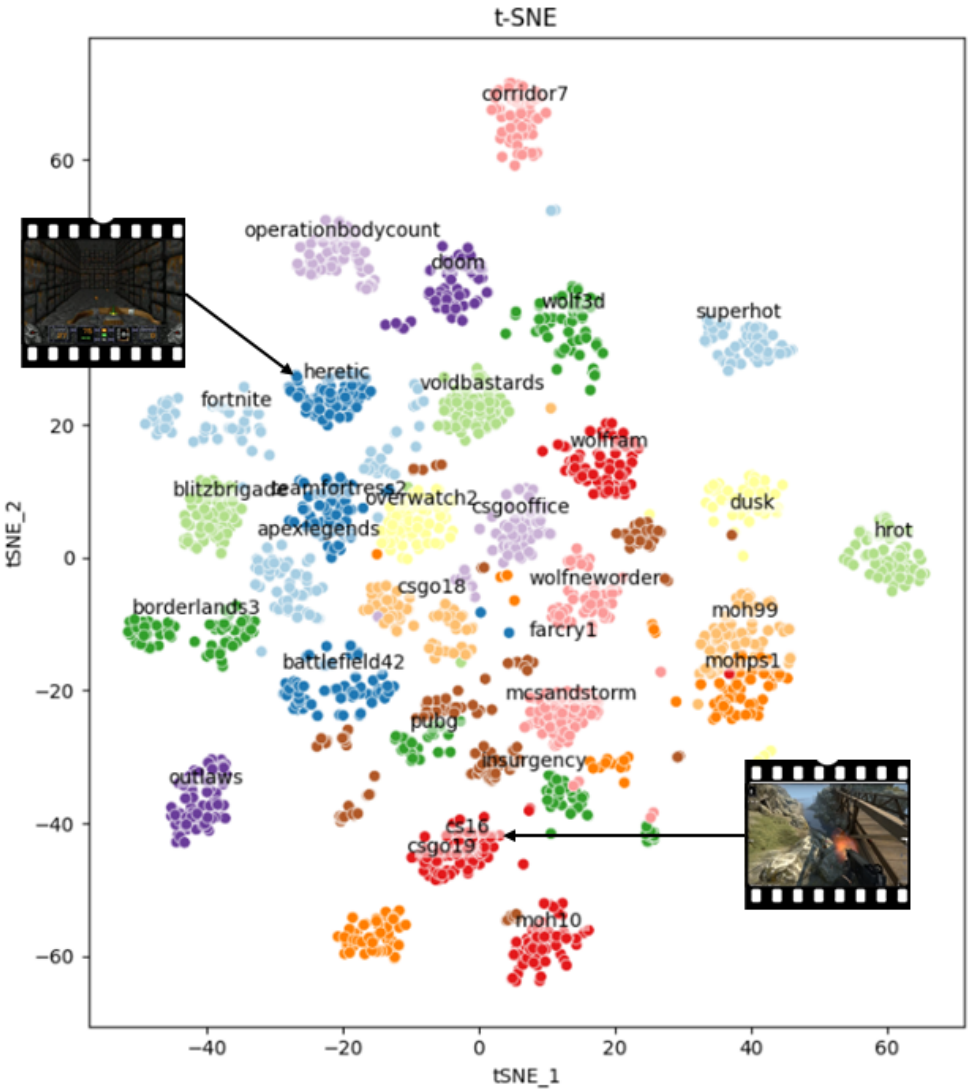


Figure 7.1: t-SNE visualisation of the MVD latent space embeddings for the GameVibe dataset. Each point represents an annotated instance of viewer engagement across various First-Person Shooter (FPS) games. Different clusters correspond to different game titles, demonstrating how the embeddings capture game-specific contextual information. The figure highlights the separation between games like Heretic and CSGO19, with thumbnails illustrating gameplay examples from these two games.

We approach the problem by treating each domain (such as different games and participant interactions) as a separate affective context, thus decomposing the multidomain problem into several domain-specific tasks. Our methodology leverages few-shot learning to ensure that the model can generalise across these domains, using minimal labelled data.

To validate our framework, we utilise video games and dyadic interactions as application areas, introducing the RECOLA Few-Shot (RECOLAFS) and *GameVibe* Few-Shot (GVFS) datasets (Barthet et al., 2024; Ringeval et al., 2013). The GVFS dataset contains data from 30 First-Person Shooter (FPS) games annotated for viewer engagement, while RECOLAFS provides data of dyadic interactions from 18 participants in a controlled environment and it is annotated in terms of both arousal and valence. These datasets provide practical testbeds for evaluating the model's ability to generalise across contextual factors within the task of affect modelling. Specifically, we focus on *participant arousal*,*participant valence*, and *viewer engagement*, capturing both cognitive and affective states, such as attention, interest, and enjoyment. (Yannakakis and Togelius, 2018).

7.1.0.1 | Contributions

This chapter presents several key contributions to the field of affect modelling and few-shot learning. First, we propose an innovative approach to transform multidomain classification tasks into multiple domain-specific few-shot learning problems. This formulation enables models to generalise effectively across domains where data is scarce and diverse, particularly in the context of affect modelling. Building on this, we introduce the GVFS (GameVibe Few-Shot) dataset, a new variation of a publicly available dataset specifically designed for few-shot learning. The GVFS dataset is tailored for learning models of viewer engagement across varying contextual factors within video games, offering a valuable resource for future research in affective computing and affect experience modelling.

In addition, we perform a comparative analysis of several few-shot learning techniques, including metric learning and contrastive learning, and benchmark them against conventional end-to-end engagement modelling methods, which are commonly used in affective computing. This analysis provides insights into the strengths of few-shot learning methods in addressing the challenges of domain generalisation for affect modelling. To further validate our approach, we conduct extensive experiments using various pretrained backbone architectures, evaluating the models across different few-shot classification scenarios, such as 5-way and 10-way, and 1-shot and 5-shot tasks. The results consistently show that few-shot learners outperform the conventional domain-agnostic baseline in most experiments, demonstrating the effectiveness of our proposed method.

Moreover, we introduce a novel loss function, the Silhouette Distance (SD) loss, which is inspired by silhouette scores and aims to create cohesive and separable clusters. The behaviour of this loss is further tested on synthetic data. Although our experiments focus on affect modelling, the proposed formulation is highly versatile and can

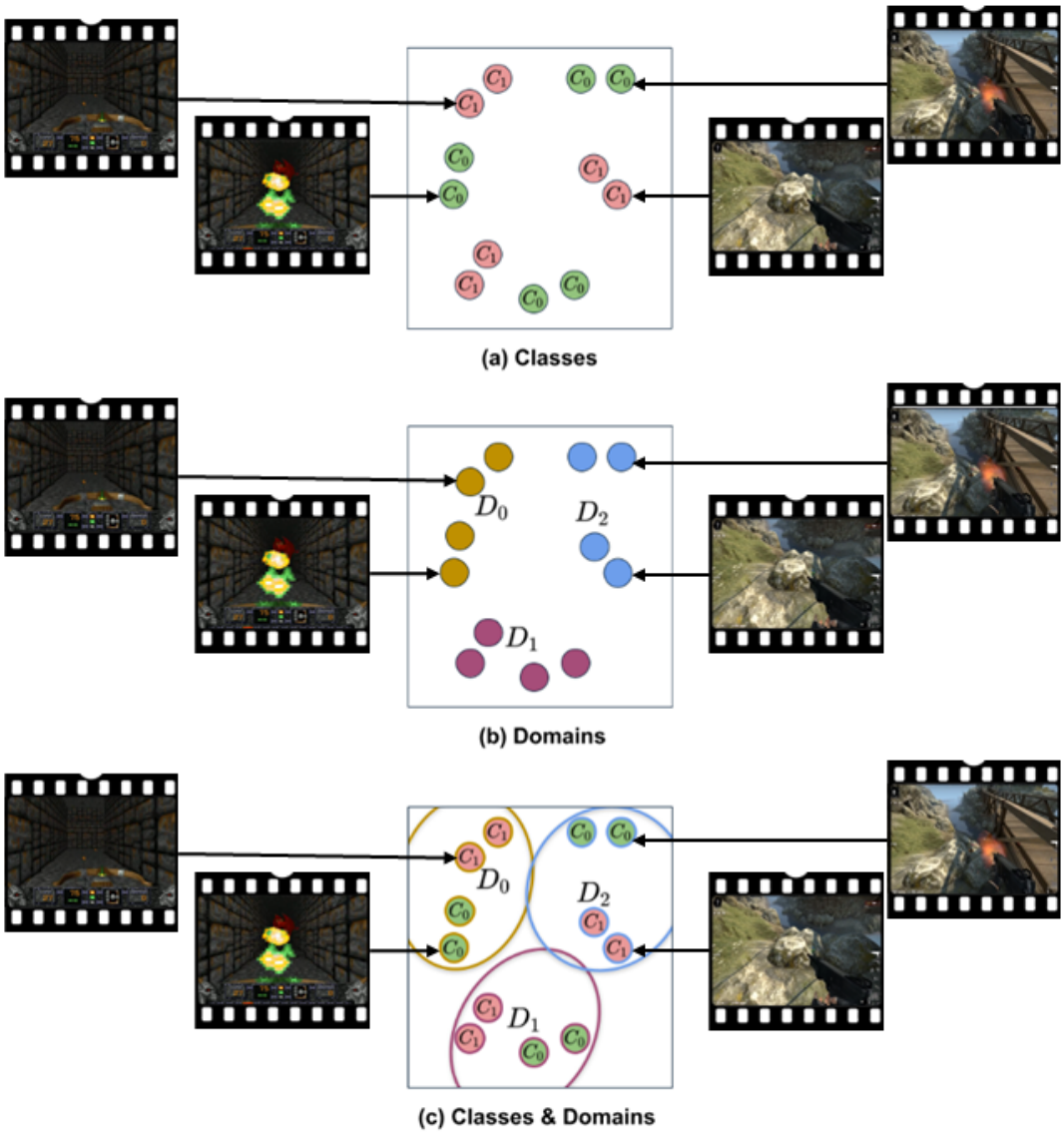


Figure 7.2: Illustration of a classification task represented as 2D embeddings with three domains (yellow, magenta, blue) and two classes (green, pink). The task is to predict user engagement levels (high vs. low) across multiple games (domains). Plot (a) shows that the two classes are not easily separable when considering the entire dataset across all domains. In plot (b), instances are clustered by domain, with points from the same domain grouped closely, but class separation remains unclear within each domain. Plot (c) illustrates the proposed method, which leverages both class and domain information. This approach results in more distinct domain clusters, allowing for clearer class separation within these more homogeneous groups.

be extended to a wide range of multidomain problems, offering a framework that can be applied to other areas of machine learning and AI. These contributions advance the state-of-the-art in affect modelling by addressing the challenges of data scarcity using few-shot learning techniques.

7.2 | Modelling Methodology

This section outlines the proposed framework for handling multidomain problems with a limited number of examples (Section 7.2.1). It describes the model architectures (Section 7.2.2) and the learning objectives (Section 7.2.3) used to learn representations in a few-shot manner. Finally Section 7.2.4 provides a straightforward analysis of the SD loss using synthetic data.

7.2.1 | Problem Setting

A primary contribution of this work is introducing a framework that improves learning from limited data with a high domain gap. Conventional end-to-end modelling methods struggle to generalise amid varying contextual factors within the same task, even in binary classification, especially when labels are noisy due to the subjective nature of engagement Yannakakis et al. (2018). Our framework exploits domain knowledge to decompose the classification problem into non-overlapping domain-specific sub-problems, with different classes falling under different domains. This ensures each domain has unique classes, which simplifies the learning process by reducing the variability and complexity within each sub-problem. Hence, Few-Shot Learning techniques can more effectively learn from the limited data available within each domain, as the model can focus on the specific features and patterns relevant to each unique domain-specific class.

7.2.1.1 | Modified Classification Objective

Let f be a function that projects data from multiple domains D_n with $n \in \{0...N\}$ into a common space of lower dimensions. We can decompose $f(\cdot)$ in the following manner:

$$f(x) = \sum_{n=1}^{N} f_n(x)\mathbf{1}_{D_n}(x) \tag{7.1}$$

were $f_n(x)$ corresponds to the function within n-th domain D_n and $\mathbf{1}_{D_n}(x)$ is the indicator function. It should be noted that $f_n(x)$ retains the core properties of $f(x)$ such as its invariance to specific transformations of x.

Our objective is to learn a function $g(\cdot)$ that maps $f(\cdot)$ to the probability of discrete categories $y \in Y$, where Y is the set of all possible classes. Consequently, the probability of class y given $f(x)$ can be defined as follows:

$$g(f(x)) = p(y|f(x)) = \sum_{n=1}^{N} p(y|f_n(x))\mathbf{1}_{D_n}(x) \tag{7.2}$$

It is evident that the predicted probability distribution depends on the domain D_n. Thus we define y_n to be the event $y|f_n(x)$ and consequently $p(y|f_n(x)) = p(y_n)$ is the probability of class y given the domain D_n. It is important to note that y_n cannot occur outside of D_n and due to that $p(y_n) = 0$ for all $D_i \neq D_n$. As a result, the new classification objective is to learn a function $g(\cdot)$ that maps $f(\cdot)$ to discrete categories $y_n \in Y_{D_n}$, for $n = 1, \cdots, N$, where Y_{D_n} is the set of all possible y_n classes within domain D_n.

In practice, we essentially need to define a set of distinct domain-specific classes with each class to exist only within its corresponding domain. To achieve this we define the relabelling function R_Y as follows:

$$R_Y(y, n) = |Y|n + y \tag{7.3}$$

with Y being the set of all possible labels of the domain-agnostic classification labels (e.g., $\{0, 1\}$ for binary classification as illustrated in Fig. 7.2.a), $y \in Y$, n a unique identifier of each domain and $|Y|$ is the cardinality of Y. For a binary classification problem with 3 domains (Fig. 7.2.c), $|Y| = 2$ and $n \in \{0, 1, 2\}$. Consequently, the relabelled classes for these domains would be $\{0, 1\}$, $\{2, 3\}$, and $\{4, 5\}$ respectively. This ensures that the same class label in different domains is treated as a different label, thus creating non-overlapping domain-specific tasks. Finally the domain identifier n can be derived either by problem-specific knowledge (e.g., each game constitutes a different domain) or by applying a clustering algorithm on top of the initial projection function f.

It is worth noting that this approach is particularly beneficial in the context of affective computing as it accounts for domain-dependent variations in affective states. Given that emotional expressions are highly context-dependent, treating class labels as domain-specific helps mitigate discrepancies arising from differences in data distributions across domains.

7.2.1.2 | Learning From Limited Data

A major limitation of conventional machine learning methods is their reliance on large-scale labelled datasets for training. In many scenarios, particularly those involving diverse domains, collecting extensive labelled data for each domain is impractical or infeasible. Additionally, even if such data were available, training an end-to-end classifier

on all domains would not necessarily yield a model capable of generalising to novel, unseen domains.

Few-shot learning (FSL) addresses this challenge by enabling models to learn and adapt to new categories using only a few labelled examples. To facilitate this, we define three distinct and non-overlapping datasets: D^{train}, D^{val}, and D^{test}, corresponding to training, validation, and testing phases, respectively. The learning process is structured into multiple episodes, where in each episode, data is sampled from one of these datasets. Each episode consists of a classification task involving N unique classes (termed *N-way*) and K labelled samples per class (termed *K-shot*). Within an episode, the dataset is partitioned into a support set (S) and a query set (Q), where the support set provides labelled examples, and the query set contains unlabelled samples that the model must classify using the information extracted from the support set. This episodic formulation ensures that the model is repeatedly exposed to new tasks, thereby improving its ability to generalise from limited data.

The purpose of using D^{train}, D^{val}, and D^{test} across different domains in FSL is to assess the model's ability to generalise effectively when encountering previously unseen categories. The evaluation process examines how well the model classifies new query samples based on the support set. Formally, the i-th sample in the support and query sets is denoted as (x_i^s, y_i^s) and (x_i^q, y_i^q), respectively. Two key hyperparameters influence the difficulty of the few-shot classification problem: N (the number of classes per episode) and K (the number of support samples per class). Increasing N raises the complexity of the classification task by expanding the number of possible categories the model must distinguish within an episode. Conversely, reducing K limits the amount of information available in the support set, making generalisation more challenging.

7.2.2 | Representation Components

The overall methodology employed in this chapter is illustrated in Figures 3.3 and 7.3. In representation learning, an encoder refers to a neural network model that, after training, is capable of transforming input data into low-dimensional, high-level representations that capture meaningful patterns while reducing redundancy. This study explores the effectiveness of these learned representations in predicting user engagement from raw gameplay footage under a FSL setting, which aims to generalise across limited data samples. Specifically, we assess the performance of different neural network architectures as backbones for representation learning in this context.

In the case of the RECOLAFS dataset, we leverage both handcrafted features and deep learning-based representations. In particular, the handcrafted features capture

low-level statistical properties across multimodal signals. In this work we focus on audiovisual and a combination of audiovisual with physiological features. Furthermore, we employ InceptionResNet (Szegedy et al., 2017), a convolutional neural network pretrained for face recognition on the VGGFace2 (Cao et al., 2018) dataset, to obtain an additional set of robust high-level facial embeddings.

For GVFS, we leverage four distinct backbone models: one based on convolutional neural networks and three Transformer-based architectures. The CNN model employed is the Inflated 3D ConvNet (I3D) Bertasius et al. (2021), a well-established architecture that uses 3D convolutions to capture spatio-temporal features from video data. The embedding size for I3D is set to 512, ensuring a compact yet expressive representation of the input data.

For the Transformer-based models, we employ the base versions of MVD Wang et al. (2023b), VideoMAE Tong et al. (2022), and its successor, VideoMAEv2 Wang et al. (2023a). These models operate on the principle of dividing the video input into patches, with each patch size set to 16. The Transformer encoders process these patches by modelling long-range dependencies and interactions between them, making them particularly suitable for video data, where temporal dynamics play a crucial role. The embedding size for each Transformer backbone is 768, providing a higher-dimensional space that may allow for more nuanced feature extraction compared to the CNN-based I3D.

All four backbone models are pretrained on the Kinetics dataset Kay et al. (2017), a large-scale collection of video clips labelled for human action recognition. This pretraining ensures that the models have learned generalisable video features that can be adapted for the FSL tasks in this study. Each model accepts a tensor comprising 16 RGB frames as input, enabling the capture of short temporal contexts in gameplay footage.

In addition to the backbone models and fine-grained features, a fully connected layer with ReLU activation is appended on top of the input representations. This layer is optimised according to the FSL objectives described in Section 7.2.3 and is responsible for fine-tuning the representations to better align with the affect prediction task. The dimensionality of this trainable layer matches that of the output embedding from the corresponding backbone, ensuring smooth integration between the frozen backbone and the FSL-specific task. Finally, we apply $L2$ normalisation to project the resulting embeddings onto the unit sphere, which has been shown to improve stability in downstream tasks by normalising feature magnitudes across samples (Trivedi et al., 2022).

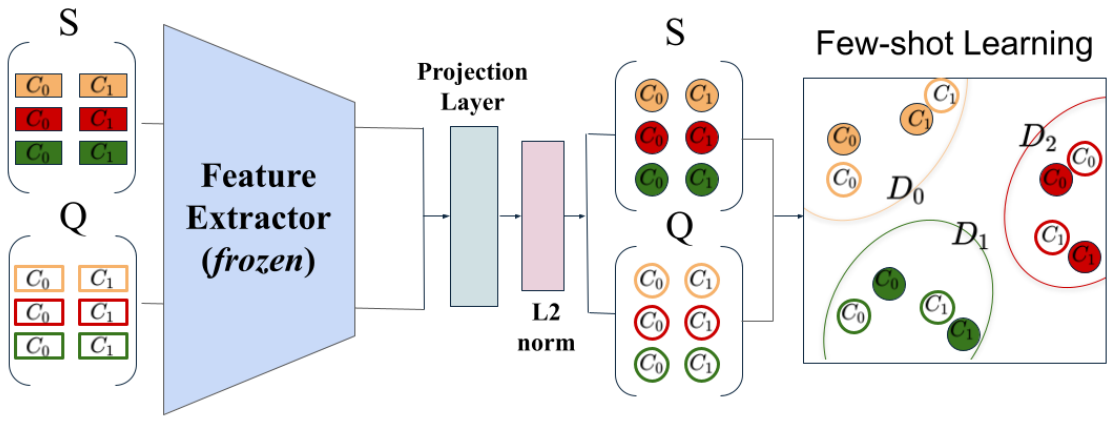


Figure 7.3: Illustration of a few-shot learning problem across three domains (yellow, red, green), each containing two classes (C_0 and C_1). The sets S and Q represent the support and query sets, respectively. First, embeddings are extracted using a pre-trained frozen feature extractor. These embeddings are then passed through a trainable projection layer, followed by $L2$ normalisation. The final step involves optimizing few-shot learning losses using the normalised embeddings from S and Q, enabling the model to learn from limited examples across the domains.

7.2.3 | Few-Shot Learning Objectives

As described in section 3.4 several few-shot learning objectives have been tested in order to enhance the generalisation ability of models in affective computing, particularly when working with limited data. These objectives include the Prototypical Network (PN) Loss, Matching Network (MN) Loss, Supervised Contrastive (SC) Loss, and Silhouette Distance (SD) Loss, all of which play crucial roles in learning robust and transferable representations. (Chen et al., 2022; Pinitas et al., 2024b; Snell et al., 2017; Vinyals et al., 2016)

The Prototypical Networks operate by representing each class with a prototype, which is essentially the mean of the support samples in the latent space. During classification, new query examples are compared against these prototypes based on a distance metric, cosine similarity in our case, with the closest prototype determining the predicted class. This method is particularly effective in few-shot learning because it efficiently captures class-specific patterns with minimal data.

In contrast, Matching Networks use an attention mechanism to compare the query sample to each support sample directly. This network leverages a weighted combination of the most similar support samples to make predictions, which is beneficial in situations where the data is noisy or highly variable. Matching networks perform well under conditions where a clear similarity between support and query examples can be

established.

The Supervised Contrastive loss extends traditional contrastive learning by incorporating label information. This method clusters examples from the same class close together while pushing apart examples from different classes, leading to better separation of class representations. However, it is important to note that this approach does not explicitly optimise for intra-class cohesion but instead relies on achieving high inter-class separation. The key advantage of supervised contrastive learning is that it structures the latent space in a way that facilitates the generation of discriminative representations, making it ideal for few-shot and domain generalisation tasks.

Lastly, the Silhouette Distance loss, introduced as a novel objective in this thesis, focuses on maximising the separation between clusters of different classes while maintaining the cohesion of clusters within the same class. By optimising this metric, the learned representation space permits cohesive yet separable representations, allowing the model to distinguish between classes more effectively, even with limited training data.

7.2.4 | Analysing the Behaviour of the Silhouette Distance Loss

This section provides a detailed analysis of the behaviour of the Silhouette Distance (SD) Loss during the optimisation process, with an emphasis on the two components that make up the loss: the inter-class and intra-class distance metrics. To offer deeper insights into how these components interact and contribute to the formation of well-defined class representations, we construct two synthetic classification scenarios. These scenarios differ in their degree of class separation, allowing us to observe the impact of the SD loss in various contexts. Specifically, the first scenario represents a situation where the classes are clearly distinguishable, while the second scenario introduces more ambiguity, with little to no separation between the classes.

For each of these scenarios, we train a simple ReLU-activated layer that utilises $L2$-normalised activations. The primary goal is to explore the effects of focusing on either one or both components of the SD loss during optimisation. By isolating the contributions of the intra-class and inter-class distances, we aim to better understand their individual and combined effects on the representation space. The synthetic data generation process creates clusters of points that are normally distributed with a standard deviation of $\sigma = 1$. These points are positioned on the vertices of a 20-dimensional hypercube. A key parameter in the data generation process is the class separation factor, denoted as h, which acts as a scaling factor to control the distance between the cluster centroids. In this experiment, we set h to two distinct values—0.9 for the *No Separa-*

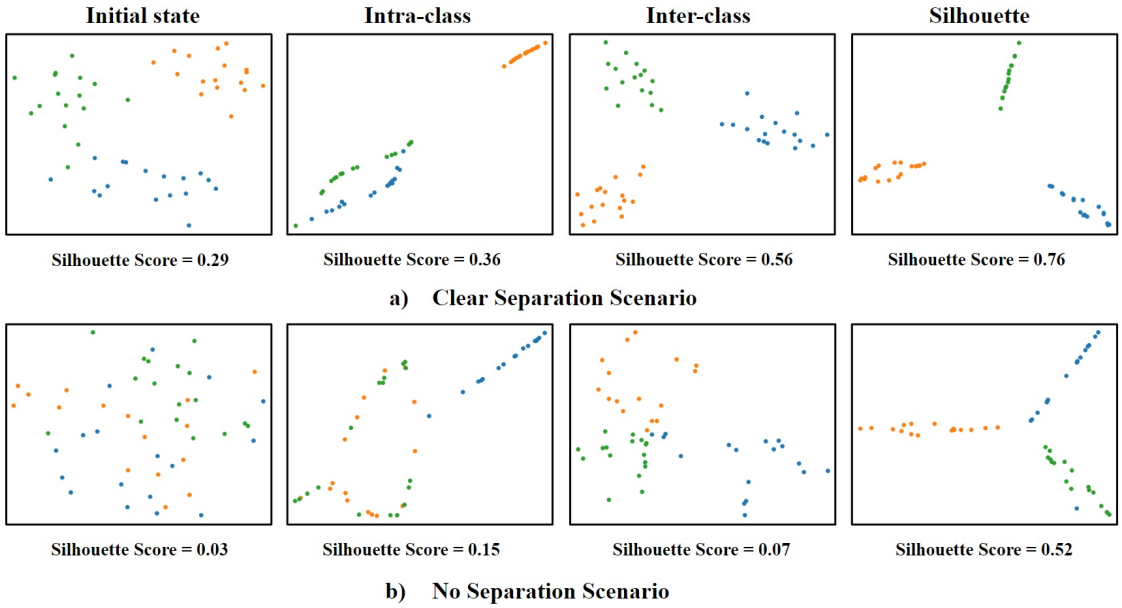


Figure 7.4: Analysis of the SD loss components, visualised with t-SNE plots, along with their respective silhouette scores in two scenarios. *Initial state* corresponds to the dataset input space. *Intra-class* and *inter-class*, respectively, refer to the projection of data optimised for intra-class distance minimisation and inter-class distance maximisation. *Silhouette* refers to a latent space that minimises the silhouette distance. It can be observed, both visually and through silhouette scores, that using individual components as loss functions slightly improves the clustering of different classes within the embedding space. However, when combined into the SD loss they exhibit notable enhancements (rightmost t-SNE plots).

tion scenario, where classes overlap significantly, and 2 for the *Clear Separation* scenario, where the classes are distinctly separated in the feature space.

Figure 7.4 visually demonstrates the influence of each component on the latent space properties across both scenarios. In the *clear separation* case, the data generation process results in well-separated clusters. This separation is preserved throughout the training process, with both inter-class and intra-class distances contributing to maintaining clear class boundaries. The effectiveness of the SD loss in retaining class separation is confirmed by both qualitative observations and quantitative metrics, such as silhouette scores, which highlight the robustness of the representations.

Conversely, the *no separation* scenario reveals the distinct roles of the two components of the SD loss. When focusing on the intra-class distance, the resulting clusters become tighter and more compact, but at the cost of class separation, making the clusters less distinguishable from one another. On the other hand, the inter-class distance promotes better separation between the clusters, but the intra-class cohesion diminishes,

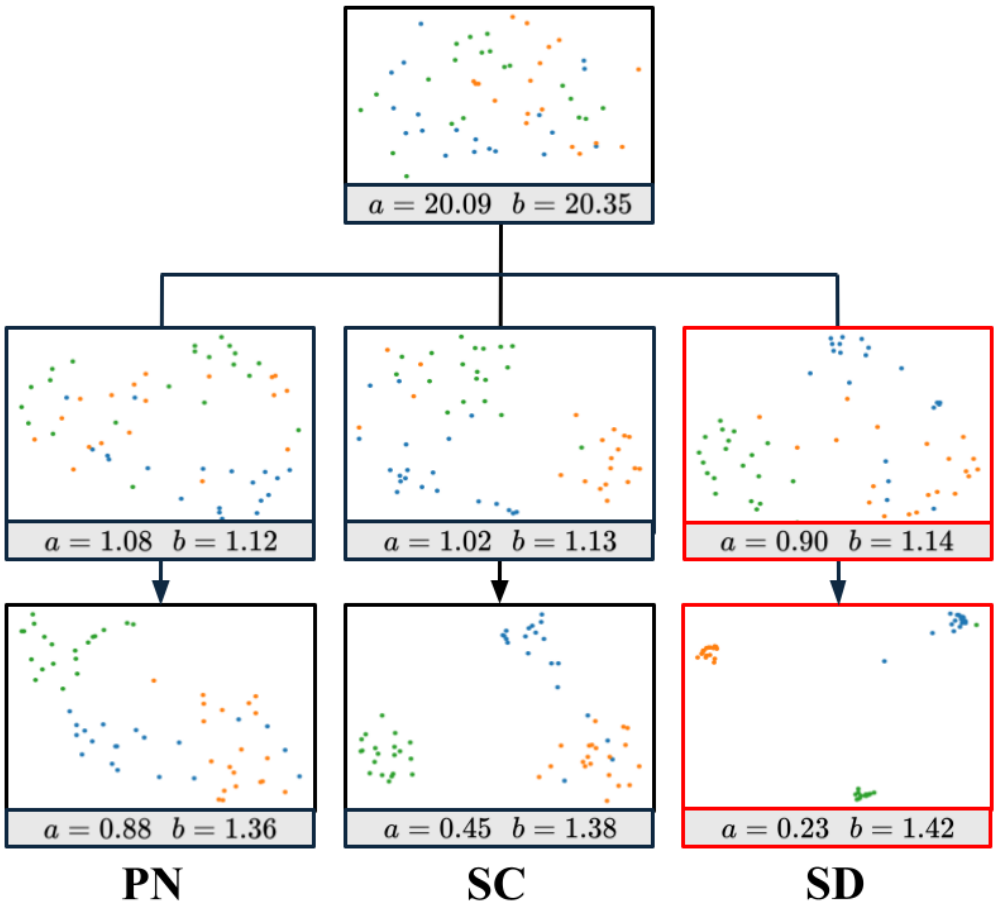


Figure 7.5: Illustration of t-SNE visualisations of embedding spaces learned by optimising three different loss functions: PN (Prototypical Network), SC (Supervised Contrastive), and the proposed SD (Silhouette Distance) loss. The top row represents the initial embeddings from a synthetic dataset with 20 input dimensions and three classes (colour-coded). The middle row shows the optimisation state at epoch 50 using a single-layer perceptron, while the bottom row depicts the final embeddings at convergence (epoch 100). Parameters a and b denote intra-class distance (lower is better) and inter-class distance from the nearest class (higher is better), respectively.

leading to more diffuse and less tightly grouped points within each class. This trade-off between cohesion and separation becomes especially apparent in scenarios where the classes are naturally overlapping, as evidenced by the relatively modest improvements in silhouette scores compared to the clear separation case.

Moreover, the experiment underscores some limitations of relying solely on either inter-class or intra-class distance. While each component brings improvements in certain aspects of the clustering process, neither is sufficient on its own to produce optimal representations. In the *no separation* scenario, although there is some improvement in

the clustering of points relative to the initial feature space, as indicated by improved silhouette scores, the overall performance is not as substantial as in the clear separation scenario.

An additional analysis examines the differences in representations learned by optimising SD against PN and SC loss functions on synthetic data. Following the same data generation process, we set the class separation parameter to $h = 1.1$ in this case. As shown in Figure 7.5 all loss functions enable the learning of discriminative representations in a supervised setting. However, we argue that L_{PN} primarily focuses on defining a decision boundary between the query sample and its corresponding prototype, functioning similarly to a cross-entropy minimisation objective. In contrast, L_{SC} emphasises the formation of well-separated clusters for each class, whereas L_{SD} optimises the overall clustering structure, promoting both cohesion within classes and separation between them. This is further illustrated in Figure 7.5, where SD achieves the lowest intra-class distance and the highest inter-class distance.

Ultimately, the SD loss optimisation leads to higher-quality clusters than those found in the initial latent space in both synthetic scenarios. However, the specific characteristics of the resulting clusters—such as their shape, cohesiveness, and degree of separation—are strongly influenced by the geometrical properties of the input space, such as the arrangement and separation of the points in the high-dimensional hypercube. This suggests that while the SD loss is effective at refining representation quality, the structure of the input data plays a crucial role in determining the ultimate success of the clustering process.

7.3 | Data Preprocessing

This section presents the datasets used for validating our proposed methodology and the data preprocessing steps.

7.3.1 | The GameVibe Few-Shot Dataset Preprocessing

The GameVibe Few-Shot (GVFS) dataset is derived from the GameVibe corpus, a publicly available dataset featuring gameplay footage from 30 diverse commercial first-person shooter (FPS) games, annotated for viewer engagement. The corpus consists of four subcorpora, each containing gameplay clips from the 30 FPS games, and annotations are provided by five randomly assigned participants to capture varied perspectives. Engagement annotations were provided for 30 one-minute gameplay videos (one per game) using a definition of engagement linked to feelings of tension and excite-

ment (high engagement) versus boredom and disinterest (low engagement). To ensure reliability, the games were presented randomly to the annotators to avoid habituation, and short videos were chosen to balance between rich engagement stimuli and reliable annotation. It should be noted that a detailed description of the GameVibe dataset is provided in Section 4.3

When it comes to data preprocessing, each session video is divided into 1-second time windows (non-overlapping). To account for reaction time between gameplay and annotation, the input is shifted by 1-second relative to the annotation time window. Video segments are converted into RGB frames (30 per second). To reduce computational load, 16 RGB frames (224×224×3 pixels) are sampled at regular intervals within each window. Engagement traces (one per annotator, five per gameplay video) undergo min-max normalisation, scaling values to [0, 1] for each trace, and the median trace is derived to mitigate inter-annotator disagreement (Grewe et al., 2007). The resulting engagement trace is segmented into 1-second windows, and the average engagement value (e) is calculated for each window.

Predicting user experience labels is challenging due to the subjective nature of such labels and the systematic reporting errors they might contain. Thus, following best practices from recent studies in the literature Makantasis et al. (2023); Pinitas et al. (2022b) we model engagement as categorical classification and we denote e as the engagement value within a time window. Assuming that the annotators exhibit consistent behaviour across games, we initially define game-agnostic classification labels by binarising the e values based on the median ground truth value of the entire set of affect annotation traces ($\bar{e}$) within a subcorpus (same participants within an annotation session). Consequently, the i_{th} time window falls under the *high engagement* and the *low engagement* class, respectively, when $e_i > \bar{e} + \epsilon$ and $e_i < \bar{e} - \epsilon$. Notably, the threshold $\epsilon = 0.1$ is used to eliminate ambiguous windows with annotation values close to $\bar{e}$ which may deteriorate the stability of the learning process. Since each game constitutes a domain of its own, we apply the relabelling method discussed in Section 7.2.1.1 to construct game-specific labels of engagement.

By utilising the unique game identifier $g_{ID} \in \{0, \ldots, 29\}$ (domain identifier n of Eq. 7.3), the set of (binary) labels derived in the previous step and the relabelling function of Eq. 7.3 we define low and high engagement class labels for each game. For the game with $g_{ID} = 0$ low and high engagement classes are represented by 0 and 1, respectively. For the game with $g_{ID} = 1$ the same classes are represented by 2 and 3. In general, for a game with $g_{ID} = n$, low and high engagement classes are represented by $2n$ and $2n + 1$, respectively. Since those classes are categories without a natural order, the game id value does not affect the output of our models. Finally, based on those labels we discard

games that yield less than 10 samples per class since they don't allow for sampling Q and S sets from both classes and can lead to overinflated performance within the few-shot learning setting. The resulting dataset—*GameVibe* Few-Shot—is the first dataset for few-shot experience modelling within video games. We refer to the four subcorpora of the dataset as $GVFS_1$, $GVFS_2$, $GVFS_3$, and $GVFS_4$, the key properties of which are summarised in Table 7.1.

7.3.2 | The RECOLA Few-Shot Dataset Preprocessing

The RECOLA Few-Shot dataset used in this study is derived from the RECOLA corpus, a publicly available multimodal dataset designed for affect recognition. RECOLA features recordings of 23 participants engaged in collaborative tasks via video-conference, annotated continuously for arousal and valence by six annotators. This dataset is particularly suited for few-shot learning due to its rich multimodal inputs (audio, visual, and physiological signals) and the inherent challenge of data scarcity in affective computing.

To prepare the dataset for few-shot learning tasks, the fine-grained features are standardised to a common scale using z-score normalisation, ensuring comparability across participants and modalities. For consistency and computational efficiency, the multimodal features are segmented into 1-second, non-overlapping time windows and the average feature vector is calculated per time window. Similarly, Visual representations are extracted using the pretrained InceptionResNet model, which generates 512-dimensional embeddings for each frame. Once again the resulting visual representation is the average representation within each time window.

The feature vectors of each modality and the corresponding frames are aligned to the annotation timestamps thus there is no need to account for the reaction time of the annotators. When it comes to the annotations of affect. The median trace is computed across annotators to mitigate inter-annotator variability while the average annotation value is calculated within each time window. Thus for each time window have access to three different fine-grained feature vectors (video, audio and physiology), a single visual representation vector extracted by InceptionResNet and one scalar value for arousal and valence.

Following the same strategy as in the GVFS dataset, we first extract binary domain-agnostic affect state labels as described in section 5.2.3.1. Given the participant-specific nature of RECOLA, each participant is treated as a distinct domain. To adapt the dataset for few-shot learning, we apply a domain-specific relabelling strategy, where binary affective state labels are encoded uniquely for each participant. For participant p the "low" and "high" classes are represented as $2p$ and $2p+1$, respectively. Participants with fewer

Table 7.1: High-level statistics of the GVFS dataset. Each subcorpus includes 5 unique annotators. The train / valid / test columns refer to the number of games in the train validation and test set, respectively. Values within parentheses correspond to the number of distinct classes (2 per game).

subcorpus	#samples	#games	#train / valid / test	Binary Majority
$GVFS_1$	1054	23	8 (16) / 8 (16) / 7 (14)	52.47%
$GVFS_2$	1026	22	8 (16) / 7 (14) / 7 (14)	52.14%
$GVFS_3$	797	19	7 (14) / 6 (12) / 6 (12)	54.20%
$GVFS_4$	1186	27	9 (18) / 9 (18) / 9 (18)	50.34%

Table 7.2: High-level statistics of the RECOLAFS dataset. Each affect dimension is annotated by 6 experts. The train / valid / test columns refer to the number of games in the train validation and test set, respectively. Values within parentheses correspond to the number of distinct classes (2 per game).

dimension	#samples	#participants	#train / valid / test	Binary Majority
Arousal	2754	18	6 (12) / 6 (12) / 6 (12)	60.02%
Valence	1706	16	6 (12) / 5 (10) / 5 (10)	59.60%

than 10 samples per class are excluded to ensure sufficient data for constructing support (S) and query (Q) sets in few-shot tasks. The properties of the RECOLAFS dataset, including the number of participants, and sample distributions, are summarised in Table 7.2.

7.4 | Results

In this section, we evaluate the performance of the proposed approach on affect classification from a small number of samples across contexts. We employ the GameVibe Few-Shot (GVFS) and RECOLA Few-Shot (RECOLAFS) datasets,that are specifically tailored to the task at hand, as our benchmarks. The subsequent subsections present the experimental protocol designed to assess our methods, and a comprehensive analysis of the results. First, we outline the experimental protocol (Section 7.4.1), including the metrics, evaluation setup, and training configurations. Finally, we present and discuss the results (Section 7.4.2), offering insights into the effectiveness of the model and its ability to generalise across different settings.

7.4.1 | Experiment Protocol

In this study, we evaluate the models optimised by the loss functions described in Section 3.4 under a few-shot learning (FSL) framework, using the N-way K-shot setting to assess their performance. This approach simulates realistic conditions where only a few labelled examples are available for each class. For each episode, we randomly sample N classes and K samples per class to form the support and query sets. The goal is for the model to generalise from these limited examples and correctly classify new instances during testing.

Following best practices established in prior studies on FSL Chen et al. (2022); Liu et al. (2021), we primarily report results for the widely-used 5-way (5w) classification task, where the model is trained to differentiate between five classes. This evaluation is performed in both 1-shot (1s) and 5-shot (5s) settings. In the 1-shot setting, the model is given only one labelled example per class, while in the 5-shot setting, it is provided with five examples. To further challenge the models and test their generalisation capabilities, we extend the evaluation to a more difficult 10-way (10w) setting, where the model must classify between 10 different classes, again in both 1-shot and 5-shot configurations. This progressive scaling of the number of classes and examples allows for a comprehensive understanding of the capacity of the the tested models across varying levels of difficulty.

7.4.1.1 | Training and Evaluation

The models are trained using early stopping, where training halts if there is no improvement in validation accuracy after 10 consecutive epochs (equivalent to 40-200 episodes, depending on the specific experiment). The best model is then selected based on the highest validation accuracy achieved. To optimise the models, we employ stochastic gradient descent (SGD) with momentum, coupled with a learning rate scheduler that reduces the learning rate α by half every 5 epochs (20-100 episodes). Through preliminary experiments, we determined that the optimal learning rate falls within the range of $\alpha \in (10^{-3}, 10^{-2})$, and hyperparameters are fine-tuned via a greedy search on the validation set. This ensures that the model is trained with parameters best suited for generalisation.

The evaluation protocol used in this study is designed to facilitate fair comparisons between the various few-shot learning models tested. Specifically, we measure model performance in terms of accuracy, following the standard evaluation protocol used in prototypical networks, one of the most well-known FSL methods. By adhering to this protocol, we ensure consistency and fairness in comparing the effectiveness of different models and approaches. To account for variability in model training and performance,

Table 7.3: **5-way few-shot experiments** (1-shot and 5-shot) across the GVFS subcorpora and on average. Mean accuracy of the Matching Network Loss (MN), Prototypical Network Loss (PN), Supervised Contrastive Loss (SC) and Silhouette Distance Loss (SD). Bold values indicate the highest accuracy obtained for each sub-corpus and backbone used. Underlined values denote methods whose accuracy is statistically equivalent to the accuracy obtained by SD as determined by the 95% confidence interval.

Backbone	Method	$GVFS_1$		$GVFS_2$		$GVFS_3$		$GVFS_4$		Average	
		5s	1s	5s	1s	5s	1s	5s	1s	5s	1s
I3D	MN	91.21	84.71	**91.34**	83.30	89.02	79.87	92.88	**85.42**	91.11	83.33
	PN	91.21	84.04	**91.34**	83.09	88.83	80.18	93.12	84.80	91.13	83.03
	SC	91.19	**84.80**	89.75	**83.76**	89.02	**82.53**	91.83	85.02	90.45	**84.03**
	SD	**91.71**	83.96	90.31	82.49	**89.38**	80.04	**93.24**	82.89	**91.16**	82.35
MVD	MN	81.73	74.62	84.95	75.00	80.06	74.36	84.53	75.16	82.82	74.79
	PN	81.73	74.13	84.95	74.69	80.07	74.36	84.52	74.31	82.82	74.37
	SC	77.77	75.96	82.36	75.71	78.25	72.89	81.36	77.60	79.94	75.54
	SD	**84.41**	**77.29**	**86.45**	**77.72**	**81.10**	**74.67**	**87.48**	**78.44**	**84.86**	**77.03**
VideoMAE	MN	78.86	66.31	75.83	64.14	73.54	61.69	79.96	64.0	77.05	64.04
	PN	77.95	64.8	75.83	62.96	73.35	61.51	79.96	63.64	76.77	63.23
	SC	81.73	74.27	79.77	69.00	76.44	62.93	83.1	74.67	80.26	70.22
	SD	**83.55**	**77.60**	**84.91**	**74.33**	**80.42**	**72.00**	**87.89**	**76.18**	**84.19**	**75.03**
VideoMAEv2	MN	91.20	83.82	90.99	81.78	89.40	78.98	92.88	82.62	91.12	81.80
	PN	91.89	83.20	90.93	82.11	89.40	**80.27**	92.95	83.73	91.29	82.33
	SC	**94.08**	83.33	**92.36**	**82.56**	**90.45**	79.11	**93.83**	**84.49**	**92.68**	82.37
	SD	91.07	**86.40**	91.26	82.45	90.02	79.02	92.25	81.78	91.15	**82.41**

we repeat each experiment five times. For each run, we sample 200 test episodes, which allows us to gather robust performance metrics. All significance tests are conducted at a 95% confidence interval (CI), with results considered significant if $p < 0.05$.

In addition to testing FSL models, we also establish a baseline using a conventional end-to-end learning approach. This baseline model shares the same architecture as the FSL models, as described in Section 7.2.2, including the use of pretrained backbones and identical layers and hyperparameters. The key difference is that the end-to-end model is trained on binary game-agnostic labels, predicting high or low engagement across unseen games. The model is optimised using the cross-entropy loss function and serves as a lower bound for performance, providing a benchmark against which the FSL models can be compared. By evaluating this baseline, we are able to assess how much improvement is gained by using few-shot learning techniques versus conventional methods.

7.4.2 | Results

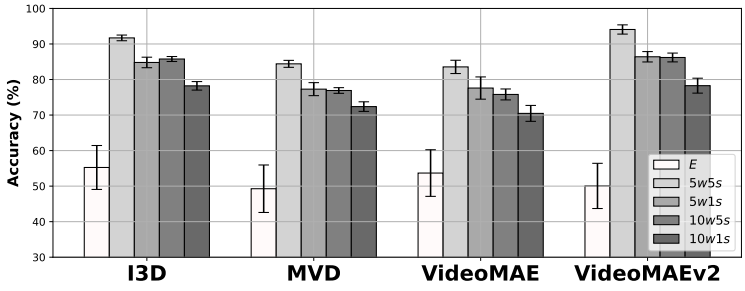
The proposed methodology is tested extensively across datasets and four few-shot settings. Table 7.3 showcases the average accuracy of the models for the 5-way few-shot experimental setting (e.g., 5 classes are compared per episode) and the backbone archi-

Table 7.4: **10-way few-shot experiments** (1-shot and 5-shot) across the GVFS subcorpora and on average. Mean accuracy of the Matching Network Loss (MN), Prototypical Network Loss (PN), Supervised Contrastive Loss (SC) and Silhouette Distance Loss (SD). Bold values indicate the highest accuracy obtained for each sub-corpus and backbone used. Underlined values denote methods whose accuracy is statistically equivalent to the accuracy obtained by SD as determined by the 95% confidence interval.

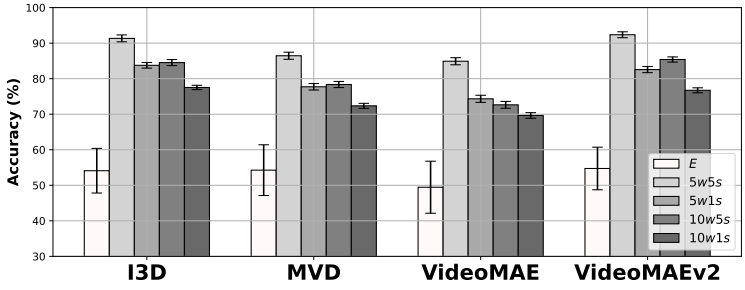
Backbone	Method	$GVFS_1$		$GVFS_2$		$GVFS_3$		$GVFS_4$		Average	
		5s	1s	5s	1s	5s	1s	5s	1s	5s	1s
I3D	MN	85.18	**78.22**	**84.53**	**77.53**	**82.65**	75.47	**87.16**	78.29	84.88	**77.38**
	PN	84.95	77.49	83.79	77.38	**82.65**	75.20	**87.16**	78.47	84.64	77.14
	SC	83.46	77.02	82.69	77.25	82.61	75.11	86.08	78.56	83.71	76.99
	SD	**85.77**	77.36	84.03	77.26	82.54	**75.82**	87.09	**78.60**	**84.86**	77.26
MVD	MN	73.37	67.2	74.89	67.29	70.23	65.93	75.31	68.11	73.45	67.13
	PN	73.13	68.09	74.88	67.53	70.23	65.93	75.31	68.18	73.39	67.43
	SC	72.51	66.44	74.34	66.89	69.37	66.0	73.92	68.04	72.54	66.84
	SD	**76.91**	**72.38**	**78.36**	**72.36**	**71.23**	**69.29**	**79.35**	**74.49**	**76.46**	**72.13**
VideoMAE	MN	69.08	58.31	66.36	57.17	63.92	54.62	70.25	58.02	67.40	57.03
	PN	68.36	59.73	66.36	56.95	63.92	54.93	70.25	57.11	67.22	57.18
	SC	70.23	58.87	69.73	57.19	**68.55**	55.76	74.56	58.29	70.77	57.53
	SD	**75.80**	**70.47**	**72.64**	**69.65**	68.44	**67.76**	**79.74**	**71.40**	**74.16**	**69.82**
VideoMAEv2	MN	84.56	75.80	84.08	74.88	82.38	72.71	86.34	76.40	84.34	74.95
	PN	84.27	76.67	83.72	74.88	82.16	72.13	86.33	76.22	84.12	74.98
	SC	**86.20**	**78.27**	**85.4**	76.15	**83.89**	73.22	**87.64**	**77.62**	**85.78**	**76.32**
	SD	85.25	77.33	84.58	**76.73**	82.44	**75.71**	87.11	77.56	84.85	76.83

Table 7.5: **5-way and 10-way few-shot experiments** (1-shot and 5-shot) across both affect dimensions of the RECOLAFS dataset. Mean accuracy of the Matching Network Loss (MN), Prototypical Network Loss (PN), Supervised Contrastive Loss (SC) and Silhouette Distance Loss (SD). Bold values indicate the highest accuracy obtained for each sub-corpus and backbone used. Underlined values denote methods whose accuracy is statistically equivalent to the accuracy obtained by SD as determined by the 95% confidence interval.

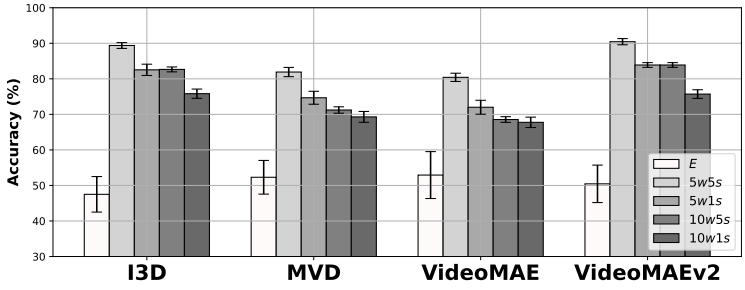
Representation	Method	Arousal				Valence			
		5w5s	5w1s	10w5s	10w1s	5w5s	5w1s	10w5s	10w1s
All Features	MN	74.39	58.40	66.30	51.86	63.65	56.40	57.99	49.67
	PN	73.84	58.93	66.73	52.40	63.33	56.93	59.15	49.93
	SC	**80.27**	60.13	**75.61**	53.13	**78.11**	53.47	**70.67**	49.87
	SD	74.11	**66.26**	69.83	**59.27**	67.31	**68.13**	58.75	**57.07**
Audiovisual Features	MN	58.48	46.53	55.63	40.73	49.95	47.47	44.37	42.93
	PN	61.68	47.73	54.72	42.47	51.65	46.40	46.45	40.53
	SC	**74.99**	54.67	**68.95**	47.73	**64.88**	47.87	**57.59**	44.47
	SD	68.21	**60.13**	58.40	**54.20**	57.04	**51.20**	44.95	**45.93**
InceptionResNet (VGGFace2)	MN	86.35	81.60	**79.44**	73.33	89.68	83.73	84.39	79.87
	PN	86.69	77.33	78.72	71.67	**91.81**	82.40	82.81	76.93
	SC	80.56	**82.13**	74.01	**74.20**	86.83	**86.13**	83.93	78.87
	SD	**86.69**	76.93	78.93	73.93	90.48	83.60	**84.83**	**79.90**



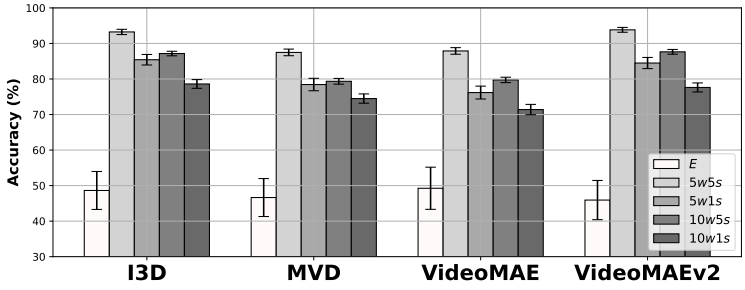
(a) $GVFS_1$



(b) $GVFS_2$

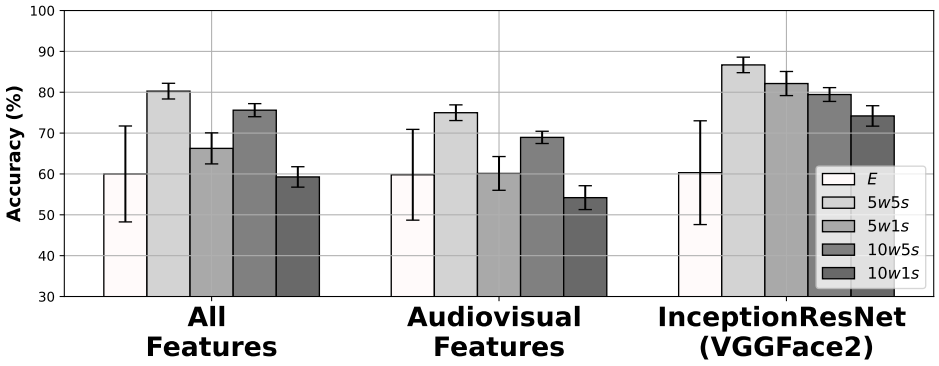


(c) $GVFS_3$

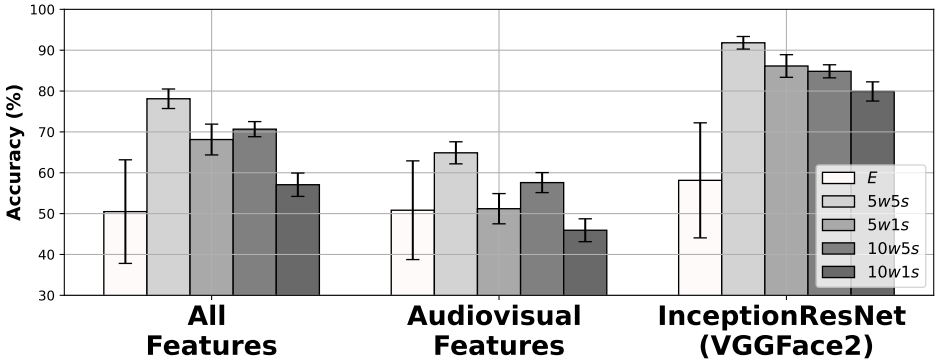


(d) $GVFS_4$

Figure 7.6: **GVFS Dataset** Average validation accuracy scores (%) for high-low few-shot engagement classification. Values are averaged across 1000 independent episodes; 95% confidence intervals are displayed as error bars.



(a) RECFS: Arousal



(b) RECFS: Valence

Figure 7.7: **RECOLAFS Dataset** Average validation accuracy scores (%) for high-low few-shot arousal (top) and valence (bottom) classification Dataset. Values are averaged across 1000 independent episodes; 95% confidence intervals are displayed as error bars.

tectures employed for the GVFS dataset. In particular, the SD method yields the highest accuracy in 20 out of 32 5-way experiments (outperforming significantly the rest of the models in 7 cases) showcasing the reliability of the contrastive silhouette-based optimisation objective. SC marks the highest accuracy in 9 out of 32 settings while performing statistically on par with SD in 7. Furthermore, MN and PN achieve the highest accuracy values in 2 out of 32 and 1 out of 32 cases respectively while in all of those cases, they perform statistically on par with SD. It is worth mentioning that all few-shot learners perform statistically on par in 17 out of 32 experiments indicating that one of the main sources of performance improvement is the training framework (see Section 7.2.1) itself.

Table 7.4 illustrates the average accuracy of the models for the 10-way few-shot experimental setting (e.g., 10 classes are compared per episode) and the backbone architectures employed on the GVFS dataset. The few-shot learners and the end-to-end model performed statistically on par in the remaining 3 experiments, where VideoMAE was

used as the backbone. Once again, the SD method marks the highest accuracy in 20 out of 32 10-way experiments while performing significantly better than the rest of the FSL models in 14 cases. It should be noted that SD was the only model that significantly outperformed the end-to-end model (E) across all 1-shot experiments and backbones. SC marked the highest accuracy in 7 experiments while performed statistically on par with SD in all of those cases. The same holds for MN and PN which achieved the highest accuracies in 5 and 2 experiments, respectively, without managing to significantly outperform SD in any of those. Moreover, all few-shot learners performed statistically on par in 14 out of 32 experiments further strengthening the efficacy of the proposed framework to generalise across domains with limited data regardless of the choice of the backbone and few-shot learning objective.

Importantly, the average performance of the models across the GVFS subcorpora showcases that accuracy improves significantly from 1-shot to 5-shot settings across all methods and backbones, reflecting the advantage of having more examples in few-shot learning. Furthermore, as the number of classes increases from 5-way to 10-way, a notable drop in performance is observed across all methods and backbones. This is due to the increased complexity and higher intra-class variability, making it harder for the models to distinguish between more classes with limited samples. Additional adversity comes from the subjective nature of experience annotation that inherently leads to noisy mappings between embeddings and labels.

Figure 7.6 compares the performance of the best few-shot learners (as indicated by the bold values of Tables 7.3, 7.4) and their corresponding end-to-end baseline across backbones for each subcorpus of the GVFS dataset. It is evident that FSL significantly outperforms the end-to-end baseline (E) across all few-shot learning experiments demonstrating the robustness of the proposed approach and the ability of the resulting FSL models to generalise across different domains within the same subcorpus. Furthermore, baseline models (E) struggle to learn engagement patterns likely due to the significant domain gaps existent between different video games, thereby yielding poor generalisation. Unlike our approach, which uses domain-specific information to address these challenges, the baseline models assume a single distribution across games and thus fail to capture their unique features.

In the case of RECOLAFS, table 7.5 summarises the results of 5-way and 10-way few-shot experiments (1-shot and 5-shot settings) conducted across two affect dimensions: arousal and valence. These experiments evaluate three feature representations—All Features, Audiovisual Features, and InceptionResNet (VGGFace2)—in conjunction with four few-shot learning methods: Matching Network Loss (MN), Prototypical Network Loss (PN), Supervised Contrastive Loss (SC), and Silhouette Distance Loss (SD). The

table reports mean accuracy values, with bold entries indicating the highest accuracy in each configuration, while underlined values denote results statistically equivalent to SD based on a 95% confidence interval. For arousal, the results reveal that the All Features representation performs well across most configurations. Specifically, the Supervised Contrastive Loss (SC) achieves the highest accuracy for the 5-way 5-shot and 10-way 5-shot settings, demonstrating its ability to effectively model arousal states when more training samples are available. Silhouette Distance Loss (SD) marks the highest accuracy in the 1-shot experiments while performing slightly worst than SC in the 5-shot settings. The Prototypical Network Loss (PN) yields moderate performance, particularly in 1-shot settings.

When restricted to Audiovisual Features, accuracy generally declines compared to All Features. Nevertheless, SC maintains its superiority in the 5-way 5-shot and 10-way 5-shot settings, showing its robustness even with reduced input dimensionality. SD exhibits comparable results in the 1-shot configurations. For the InceptionResNet (VGGFace2) representation, SD consistently outperforms other methods, achieving the highest accuracies in the 5way-5shot configuration. PN and MN also perform well in this setting, demonstrating its suitability for few-shot learning when supported by expressive feature representations. Furthermore SC marks the highest accuracy in both 5way-1shot and 10way-1shot configurations.

For valence, the results follow similar trends. Specifically, when All Features are used as input SC marks the highest accuracy in the 5-shot experiments while SD outperforms the rest of the models in the 1-shot experiments. The same behaviour is observed when using Audiovisual features as input. Howeve,r in this case, the models mark lower accuracies compared to the All Featurs. For Inception ResNet, all few-shot learners perform on par in the 5way-1shot and 10 way experiments. SD marks the highest accuracy in the 10 way experiments while SC is the best model in terms of absolute accuracy for the 5way-1shot setting. PN marks the highest accuracy in the 5 way-5 shot setting performing on par with MN. Overall, the results indicate that increasing the number of shots improves accuracy across all methods and configurations, as additional labelled examples provide more informative context for few-shot learning. Conversely, increasing the number of ways significantly decreases accuracy, reflecting the increased complexity of distinguishing between a larger number of classes. Similarly to GVFS, the performance improvement comes from the FSL framework and not the optimisation objective itself. However, it is evident that the backbone (input space) plays a key role in the performance of the FSL methods in RECOLAFS.

Figure 7.7 compares the performance of the best few-shot learners with their corresponding end-to-end baseline (E). The error bars represent 95% confidence intervals.

Table 7.6: **GVFS Dataset** Input Space Quality across Backbones and corpora. Sil, CH and DB refer to Silhouette, Calinski-Harabasz and Davies-Bouldin metrics, respectively. Bold values correspond to the best value across modalities.

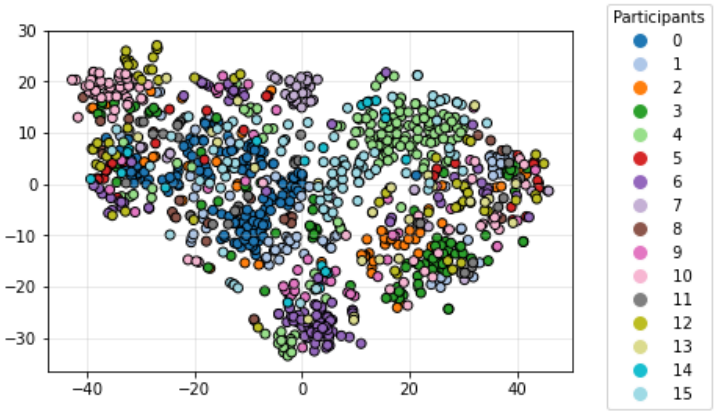
Subcorpus	Backbone	Metrics		
		Sil (↑)	CH (↑)	DB (↓)
$GVFS_1$	**I3D**	0.151	47.982	2.179
	MVD	0.131	88.784	2.434
	VideoMAE	0.042	32.345	2.826
	VideoMAEv2	0.133	38.838	2.007
$GVFS_2$	**I3D**	0.178	47.553	1.935
	MVD	0.139	74.929	1.974
	VideoMAE	0.037	32.216	2.738
	VideoMAEv2	0.153	40.978	1.966
$GVFS_3$	**I3D**	0.158	36.57	2.117
	MVD	0.134	62.786	2.275
	VideoMAE	0.046	22.836	2.860
	VideoMAEv2	0.088	30.27	2.194
$GVFS_4$	**I3D**	0.148	39.618	2.023
	MVD	0.095	72.425	2.234
	VideoMAE	0.029	27.093	2.770
	VideoMAEv2	0.123	34.015	2.044

Table 7.7: **RECOLAFS Dataset** Input Space Quality across Modalities. Sil, CH and DB refer to Silhouette, Calinski-Harabasz and Davies-Bouldin metrics, respectively. Bold values correspond to the best value across modalities.

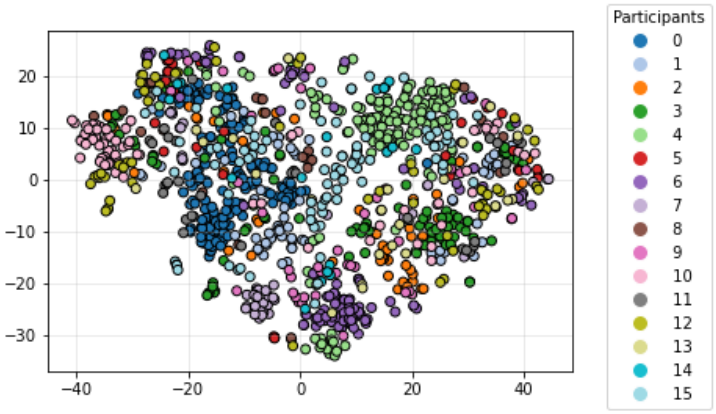
Modality	Metrics		
	Sil (↑)	CH (↑)	DB (↓)
All Features	-0.009	13.267	4.392
Audiovisual Features	-0.066	7.810	5.275
InceptionResNet (VGGFace2)	**0.184**	**99.028**	**1.893**

For both arousal and valence, when the pretrained Inception ResNet is used as input, the few-shot learners outperform the end-to-end baseline across all N way-K shot configurations. However, when All Features are used only the 5-shot few-shot learners manage to outperform the end-to-end baseline model (E). Finally the worst results are observed for Audiovisual Features since the baseline model B performs on par with almost every few-shot learner. The only exception is the 5way-5shot setting of arousal.

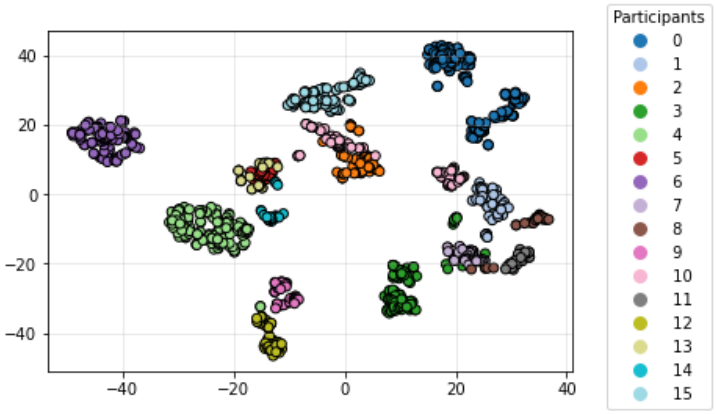
Figure 7.8 and Table 7.7 shed light into the behaviour of the proposed framework in RECOLAFS. In particular, the Silhouette Score (Sil) for both All Features and Au-



(a) All Features

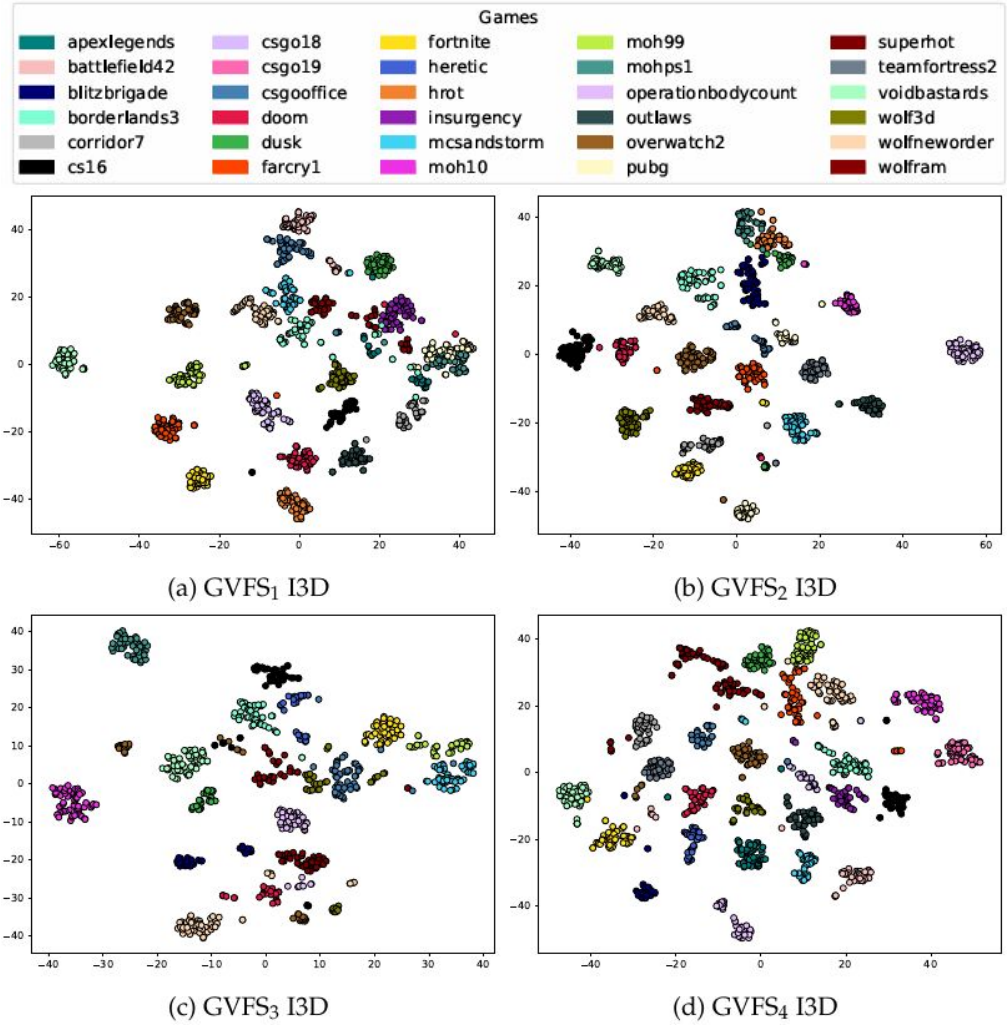


(b) Audiovisual Features



(c) InceptionResNet (VGGFace2)

Figure 7.8: **RECOLAFS Dataset** t-SNE plot illustration of the input space as shaped by All Features (a) Audiovisual Features (b) and InceptionResNet representations (c). The different colours correspond to the participant id.



(a) $GVFS_1$ I3D (b) $GVFS_2$ I3D

(c) $GVFS_3$ I3D (d) $GVFS_4$ I3D

Figure 7.9: **GVFS Dataset** t-SNE plot illustration of the input space as shaped by the I3D model. The different colours correspond to FPS games.

diovisual Features is negative, indicating overlapping or poorly separated domains, suggesting that the data points from different participants do not correspond to different domains in the feature space. This lack of multiple domains poses a significant challenge for the proposed FSL formulation, as it assumes that the data have distinct domain boundaries to effectively learn from a limited number of samples. In contrast, the InceptionResNet (VGGFace2) feature space achieves substantially better clustering metrics, including a positive Silhouette Score, a high Calinski-Harabasz Index, and a low Davies-Bouldin Index. These results highlight that the InceptionResNet input space enables the formation of distinct and compact participant clusters, thereby following the assumptions of the proposed FSL framework facilitating the success of the FSL ap-

proach. In addition, Figures 7.6, 7.9 and Table 7.6 demonstrate that even the slightest domain separation can be exploited by the proposed FSL frameworks since all backbones yield positive silhouette scores across all subcorpora of the GVFS corpus. It should be noted that the t-SNE plots for the remaining encoders of the GVFS dataset are presented in Appendix C

7.5 | Discussion

This work introduced a framework for learning from limited data across multiple domains, focusing on affect modelling in first-person shooter (FPS) games and dyadic interactions, where different games and participants represent distinct domains. The framework addresses the general problem of predicting affect states in new contexts by decomposing it into domain-specific tasks, incorporating domain-specific knowledge into the relabelling of affect labels. This approach was evaluated on the *GameVibe Few-Shot* and *RECOLA Few-Shot* datasets, designed specifically for few-shot experience modelling. Comparative analysis across the datasets' subcorpora and affect dimensions demonstrated that few-shot learning (FSL) models outperform conventional end-to-end classification methods, effectively distinguishing affect states in unseen domains, even with minimal labelled samples. While these findings highlight the promise of the proposed framework, a deeper analysis of the input feature spaces suggests that not all representations are equally suited for FSL tasks. The clustering quality analysis of the input space reveals that the proposed approach suffers from overlapping domains. This suggests that the success of the FSL framework depends not only on the methodology but also on the quality of the input space.

The study is limited by its focus on FPS games and does not account for individual viewer characteristics, such as personal interests, skill levels, or prior gaming experience, which could significantly influence engagement. Incorporating personalised features could yield more accurate engagement predictions, improving the framework's applicability to diverse gaming contexts. Additionally, extending the framework to other game genres—such as arcade, sports, puzzle, or strategy games—may reveal more nuanced engagement patterns and test the framework's generalisability across varied domains. Another key limitation is the sole reliance on accuracy as the evaluation metric. While accuracy remains the standard in the field, it does not capture the full scope of model performance, particularly in real-world applications where robustness to noisy data, computational efficiency, and scalability are critical. Future research should incorporate these additional performance metrics, along with analyses of trade-offs, such

as training time and computational complexity. Addressing these gaps will provide a more holistic understanding of the framework's effectiveness.

Additional avenues for future research include testing the framework across alternative modalities, such as audio features or graph-based representations, to explore their potential in modelling affect. Expanding the framework to handle more complex scenarios, such as domain-incremental learning, where the model must detect and adapt to new domains without explicit labels, is another promising direction. Moreover, evaluating the framework's applicability to tasks beyond affect state prediction, such as few-shot segmentation, preference learning, or even agent-based gameplay modelling, could further broaden its impact. Despite these challenges and open questions, the proposed framework remains a versatile and robust approach to few-shot learning in affective contexts. Its ability to generalise across domains positions it as a valuable tool not only for affective computing but also for broader applications in machine learning and AI.

7.6 | Summary

This chapter introduced a framework for affect modelling with limited data, leveraging few-shot learning (FSL) to enable generalisation across diverse domains with minimal labelled samples. The framework reformulates multidomain affect classification into domain-specific tasks, ensuring models effectively capture contextual patterns. Two datasets were adapted for this approach: the GameVibe Few-Shot (GVFS) dataset, which focuses on viewer engagement in first-person shooter games, and the RECOLA Few-Shot (RECOLAFS) dataset, which models arousal and valence in dyadic interactions. Domain-specific relabelling was applied to create unique class labels for each domain, facilitating FSL. Experiments demonstrated that FSL significantly outperformed end-to-end methods, particularly when paired with advanced loss functions like the Silhouette Distance (SD) loss. The framework proved effective in generalizing to unseen domains, particularly when domain boundaries were well-defined, and feature representations were robust. However, results highlighted the dependency on input feature space quality, as poorly clustered domains hindered performance. Overall, this work advances the field of affective computing by presenting a scalable and generalisable framework for affect modelling in data-scarce scenarios. The proposed methodology contributes not only to affective computing but also to broader applications in machine learning, offering a robust solution for learning from limited data across domains.

8

Discussion and Conclusions

This thesis investigated key challenges in affective computing and proposed innovative solutions through the lens of representation learning. The primary focus was on addressing the limitations of existing affective modelling techniques, particularly those related to data scarcity, multimodal integration, and domain generalisation. These challenges have long hindered the development of robust and generalisable models capable of performing reliably in real-world applications. By leveraging contrastive learning, few-shot learning (FSL), and the Learning Using Privileged Information (LUPI) framework, this research introduced novel methodologies that advanced the state-of-the-art in affective computing, enabling more scalable and adaptable systems.

The work began with an in-depth exploration of the theoretical and practical obstacles inherent to affect modelling, including the complexity of multimodal affective signals and the inherent subjectivity of emotion annotations. Emphasis was placed on the difficulties of working with noisy, incomplete, or scarce data, particularly in "in-the-wild" settings where emotional signals are often subtle and confounded by external factors. Addressing these issues required a rethinking of how affective representations are learned, leading to the development of data-efficient and generalisable approaches.

The thesis introduced new methods for representation learning that focused on extracting compact, meaningful, and multimodal embeddings of affective states. These methods integrated insights from contrastive learning to improve the discriminative power of learned representations by aligning semantically similar samples while contrasting them with dissimilar ones. Few-shot learning techniques were employed to enable models to generalise effectively with minimal labelled data, a crucial advancement for applications where large-scale annotation is impractical. The LUPI framework was leveraged to utilise auxiliary information during training, bridging the gap between controlled environments and real-world scenarios and enhancing the robustness of af-

fective models. These approaches were rigorously evaluated across diverse contexts, including video games. This thesis also contributed to the creation and collection of the GameVibe Dataset, a benchmark tailored for affective research..

At the beginning of Chapter 1, the scope of this investigation was outlined through a set of research questions designed to guide the exploration of models and methodologies in affective computing Chapter 2 provided a comprehensive review of the literature, situating this work within the broader landscape of affective computing, representation learning, and related methodologies. Chapter 3 detailed the methodological framework, outlining the approaches employed to address the challenges identified in the research questions. Chapter 4 introduced the affect corpora used in this study, discussing their characteristics, limitations, and relevance to the proposed methods. This chapter revisits the initial research objectives of Chapter 1, reflecting on them in the context of the findings and contributions presented throughout the thesis.

Objective 1: Contrastive Representations of Affect

Chapter 5 examined how contrastive learning principles can be applied to affective computing. Existing methods often struggle to learn discriminative multimodal representations that generalise to diverse subjects and contexts. To address this, a supervised contrastive learning framework was introduced in which affective labels are not merely treated as targets but integrated directly into the embedding space. This enabled the alignment of semantically similar affective states while pushing apart dissimilar ones. Evaluations on RECOLA and AGAIN demonstrated that SCL-derived embeddings capture finer-grained affective nuances, yielding improvements in downstream classification tasks. The analysis also showed that the choice of positive and negative sample definitions significantly influences the quality of the learned space, highlighting the need for task-aware construction of contrastive pairs. These findings provide evidence that SCL constitutes a powerful tool for enhancing the discriminability of affective representations.

Objective 2: Handling Missing Modalities

Chapter 6 addressed the challenge of missing modalities, a pervasive issue in affective computing caused by sensor failures, annotation gaps, or incomplete data collection protocols. This thesis proposed the use of the LUPI paradigm, whereby fine-grained features and multimodal fusions were made available as privileged information during training but not at test time. By constraining the model to rely on raw video frames under deployment conditions, the framework demonstrated improved robustness and

stability when certain modalities were unavailable. Empirical results on RECOLA and AGAIN showed that privileged supervision provides a scalable and practical strategy for mitigating the negative effects of incomplete data, thereby strengthening the reliability of affective models in realistic scenarios.

Objective 3: Knowledge Transfer from In-Vitro to In-Vivo Settings

Also in Chapter 6, this thesis explored how models trained under controlled, in-vitro settings could be transferred to more dynamic in-vivo conditions. Supervised contrastive learning was employed to pretrain teacher models with access to privileged information, which then guided student models under the LUPI framework. Experiments revealed that student models trained with privileged teachers consistently outperformed those relying solely on raw frames, in some cases approaching the performance of models with full multimodal access. This demonstrates that in-vitro models, when equipped with contrastive pretraining, can meaningfully enhance in-vivo learners, thus offering a pathway to bridge controlled experimentation and real-world deployment.

Objective 4: Few-Shot Learning for Data-Scarce Scenarios

Chapter 7 focused on the problem of data scarcity, which limits the practical scalability of affective systems. A novel framework was introduced that reformulated affective state prediction as a collection of few-shot learning tasks. By decomposing affect modelling into domain-specific, task-level problems, the framework enabled generalisation from only a handful of labelled samples. Extensive experiments on the GameVibe and RECOLA datasets showed that FSL-based methods, particularly metric-learning approaches, outperform conventional baselines in scenarios where annotated data is limited. These results underscore the value of few-shot methodologies in domains where continuous annotation is infeasible, providing a robust alternative to data-hungry deep learning models.

Objective 5: Feature Representations and Domain Generalisation

Finally, this thesis investigated the role of input feature representations in shaping the performance and generalisation of affective models. Through comparative experiments on the RECOLA and GameVibe datasets, it was shown that not all feature spaces are equally suitable for few-shot or contrastive frameworks. Certain deep embeddings and multimodal fusions provided more separable and robust representations, directly in-

fluencing classification accuracy and generalisability. Moreover, by formulating affect modelling as a set of domain-specific FSL tasks, the thesis demonstrated that models can adapt to novel domains, such as cross-game engagement prediction, with minimal labelled data. These findings highlight both the promise and the limitations of domain-transferable affect representations.

8.1 | Contributions

This section outlines the different contributions of this thesis to the field of affective computing. Although the research primarily focuses on representation learning, the insights gained can also be applied to other domains that suffer from the challenges of learning from limited and incomplete data. The key contributions of this thesis are presented below in a detailed yet concise way.

Contrastive Representations of Affect

1. **Development of Contrastive Learning Frameworks for Affective Representations:**
   This thesis advances the field of affective computing by introducing contrastive learning frameworks for affective representation modelling. It is the first to adapt the Supervised Contrastive Learning (SCL) framework for learning multimodal affect-infused representations, providing a novel approach to encoding affective information. Rigorous evaluation against end-to-end classification models on the RECOLA and AGAIN datasets highlighted the effectiveness of SCL in predicting arousal and valence across modalities. These contributions bridge gaps in the literature and lay a foundation for future research on using contrastive learning to develop robust, generalisable affective models.

2. **Design of Positive and Negative Sample Definitions for Emotional Data:**
   Although not a direct contribution of this thesis, this research direction explores the design of positive and negative sample definitions for emotional data, emphasising the critical role of affect labels as an integral part of representation learning. By introducing three supervised contrastive learning approaches, the research highlights how carefully defining positive and negative samples based on affect labels can enhance the discovery of robust, high-level representations. Evaluations on the RECOLA dataset for arousal prediction demonstrate that these representations not only outperform end-to-end models but also generalise effectively

subjects, showcasing their broader applicability to multimodal affective modelling tasks.

Learning from Missing Modalities

1. **Utilisation of Privileged Information for Multimodal Affect Modelling:**
   This thesis leverages the Learning Using Privileged Information (LUPI) framework to address the challenges of transferring affective models from controlled (in vitro) environments to real-world (in vivo) scenarios. The proposed approach incorporates auxiliary data—such as fine-grained features and fused multimodal inputs—that are available exclusively during training. By leveraging this additional information, the models learn richer and more robust representations, improving their generalisation to real-world conditions. Experimental results on the RECOLA and AGAIN datasets demonstrate the effectiveness of this approach, showing notable improvements in affect state predictions, particularly for arousal, across diverse settings.

2. **Optimisation of Teacher-Student Models in Affect Modelling:**
   The thesis further integrates Supervised Contrastive Learning (SCL) into the LUPI framework, refining the ability of models to capture transferable features. This research direction compares two teacher pretraining strategies: end-to-end modelling and supervised contrastive learning (SCL). Results demonstrate that SCL-trained teachers produce more robust representations, leading to student models with higher predictive power. These findings underscore the potential of the teacher-student framework to enhance affective computing in dynamic and challenging environments.

Affect Modelling with Limited Data

1. **Integration of Few-Shot Learning for Affective Representations:**
   This thesis explores the integration of few-shot representation learning for affect modelling addressing the challenges of data scarcity and domain generalisation. By decomposing multidomain classification into domain-specific few-shot tasks, the approach enables models to generalise across domains with limited data. Comparative analyses of few-shot techniques, including metric and contrastive learning, highlight their advantages over conventional methods, with few-shot learners consistently outperforming baselines. These contributions demon-

strate the potential of few-shot learning to advance affect modelling by enabling robust, domain-general representations in data-constrained scenarios.

2. **Introduction of a Novel Loss Function for Few-Shot Learning**
   This thesis introduced a novel loss function for few-shot learning. This loss function is inspired by the widely used silhouette score and accounts for both intra-class cohesion and inter-class separation. The proposed Silhouette Distance (SD) loss is thoroughly evaluated across diverse affective computing datasets, including scenarios with varying levels of complexity and data scarcity. Additionally, the behaviour of the SD loss is analysed on synthetic data to validate its properties and practical utility. These experiments demonstrate the effectiveness and versatility of the SD loss in enhancing few-shot learning performance, making it a valuable tool for advancing affect modelling.

3. **Evaluation of Input Feature Representations in Data-Limited Scenarios:**
   A comprehensive analysis is conducted to assess the impact of different input feature representations—such as multimodal data, audiovisual signals, and deep embeddings—on the performance and generalisation of affective models. These findings offer actionable guidelines for designing effective feature extraction pipelines. Additionally, this study sheds light into the inherent limitations of the proposed framework for few-shot domain generalisation.

4. **Application to Cross-Domain and Novel Contexts:**
   The proposed methods are validated across multiple domains, including cross-game engagement prediction, to evaluate their generalisability, scalability, and effectiveness in diverse contexts. By testing the framework on various tasks and settings, the study demonstrates its adaptability to different types of data and use cases. This contribution highlights the practical applicability of the developed methods, showing their potential to perform well in dynamic, real-world scenarios. The results underscore the framework's ability to scale across different domains while maintaining robust performance, making it suitable for a wide range of applications even out of the scope of affective computing.

5. **Development of the GameVibe Dataset:**
   A secondary contribution of this thesis is the design and collection of the GameVibe Dataset, a benchmark created for affect modelling and experience prediction in first-person shooter games. The dataset was collected in collaboration with researchers and academics at the Institute of Digital Games of the University of Malta, ensuring its relevance and rigour. In this work, a variation of the dataset,

the GVFS (GameVibe Few-Shot) dataset, is introduced as a benchmark specifically tailored for few-shot learning. This variant enables the study of viewer engagement prediction across diverse video game contexts, providing valuable insights into model performance in settings with limited data, and fostering further research in the area of affective computing in gaming environments.

8.1.1 | Towards a Unified Methodology for Affective Computing

Although this thesis investigated three methodological strands—supervised contrastive learning (SCL), learning using privileged information (LUPI), and few-shot learning (FSL)—these should not be seen as independent contributions, but as complementary elements of a unified framework for affect representation learning. The binding element is the recognition that affective information is not always directly evident in the raw input, but emerges from subtle, relational, and context-dependent cues. As such, methods that explicitly structure the representation space are required.

Supervised contrastive learning provides the theoretical foundation for this thesis by shaping embeddings to reflect affective similarity rather than superficial signal similarity. Unlike cross-entropy, which only enforces class separability, and unsupervised contrastive learning, which risks encoding irrelevant correlations, SCL leverages labels to capture the graded and relational nature of affect. This relational inductive bias aligns with the challenges of affective computing, where small differences in annotation may correspond to meaningful emotional distinctions.

The subsequent methodological contributions build directly on this foundation. The LUPI framework extends SCL representations to scenarios with missing modalities, allowing privileged teachers to guide student models under incomplete input conditions. Similarly, few-shot learning leverages the discriminative embeddings shaped through SCL to enable generalisation from very limited labelled data, reframing affect modelling as a collection of small, domain-specific tasks. Thus, both LUPI and FSL can be seen as natural extensions of the representational strength provided by supervised contrastive learning.

In this way, the thesis provides not three disjointed studies, but a coherent trajectory: beginning with the theoretical and empirical justification for supervised contrastive representations, extending this to teacher–student transfer under LUPI, and finally demonstrating their utility in few-shot and domain-generalisation settings. Together, these strands address the core challenges of affective computing—data scarcity, missing modalities, and domain variability—within a unified framework grounded in the principles of supervised contrastive learning.

8.2 | Limitations

This section describes the limitations of the work presented throughout the thesis, critically examining areas where the methodologies, approaches, or outcomes may fall short. In the context of affective computing, there is often no definitive or universally correct solution to a technical problem. Instead, the solutions devised are those deemed sufficiently effective or good enough to address the specific challenges at hand. This inherent ambiguity highlights the iterative and subjective nature of design, where trade-offs are inevitable, and choices are influenced by constraints such as available resources, data, or time. Consequently, while the proposed methods and frameworks demonstrate significant potential, they may not represent the ideal solution for all scenarios and should be interpreted within the scope of their application. Moreover, it is important to emphasise that the datasets employed in this thesis, RECOLA and AGAIN, are not directly comparable due to their fundamentally different annotation protocols. RECOLA relies on expert-annotated continuous traces bounded in $[-1, 1]$ (median of six annotators, $\varepsilon = 0.1$), whereas AGAIN is based on single self-reported scores normalised to $[0, 1]$ ($\varepsilon = 0.2$). These schemes capture distinct constructs with different levels of reliability, and therefore any cross-dataset equivalence is neither assumed nor tested in this work. All models were trained and evaluated strictly within each dataset, and results are reported separately.

Contrastive Representations of Affect

The findings presented in chapter 5 underscore the potential of supervised contrastive learning (SCL) for learning affect-infused representations, particularly in the context of binary classification of high versus low affect states across the RECOLA and AGAIN datasets. However, several methodological limitations must be acknowledged to contextualise the results. A significant limitation lies in the use of greyscale frames for visual input. While this approach was a deliberate choice to reduce computational complexity and memory requirements—thus enabling experimentation across multiple games, modalities, and modelling approaches—it inherently constrains the richness of the visual features available for learning. Colour information, especially in contexts such as the AGAIN dataset with its vibrant game environments, may carry arousal-related cues that are muted or entirely lost in grayscale representations.

Another methodological limitation concerns the absence of data augmentation techniques. While such techniques, including random cropping, flipping, brightness adjustments, and perturbations of numerical features, are known to improve model general-

isability to unseen data, they were excluded from this study to maintain focus on the primary objective of evaluating SCL's efficacy for affect representation learning. The datasets, particularly AGAIN, feature constrained diversity, and augmentation could introduce variability that enhances model robustness. Nevertheless, implementing and systematically evaluating augmentation strategies for both numerical and visual inputs was deemed outside the thesis's core scope, which prioritised foundational methodological contributions over dataset engineering.

The focus on binary classification (high vs. low affect states) rather than more nuanced regression or multi-class classification tasks also represents a deliberate simplification. This choice aligns with the thesis objectives of determining whether SCL can effectively differentiate between affective extremes. While regression or multi-class approaches might offer a richer representation of affective states, they would have necessitated substantial adjustments to the experimental setup and evaluation framework. Binary classification was chosen due to its alignment with prior literature (Makantasis et al., 2021b; Ng et al., 2015; Zhang et al., 2021) and the established focus of many representation learning approaches on classification tasks (Le-Khac et al., 2020; Saeed et al., 2021).

Additionally, the use of fixed time window lengths (1, 2, and 3 seconds) for temporal modelling may oversimplify the dynamic nature of affective changes. This static approach assumes uniform affective fluctuations across predefined intervals, which may not adequately capture the rapid affective shifts often observed in gameplay or other dynamic scenarios. Such fluctuations are inherently context-dependent and can vary significantly based on user interactions, game events, or environmental factors. Fixed windowing also risks missing subtle or transient affective states that occur at finer temporal resolutions, potentially leading to an incomplete representation of the underlying affective dynamics. While adaptive or event-based time windowing techniques could better reflect these variations, their implementation would require the integration of mechanisms to detect and adapt to changes in affective signals in real time. This would not only add considerable complexity to the temporal modelling pipeline but also demand significant computational and algorithmic advancements, making it a challenging endeavour given the resource and time constraints of this thesis.

Learning from Missing Modalities

The findings presented in chapter 6 highlight the potential of the LUPI framework and SCL in leveraging privileged information for affect modelling, particularly in binary classification tasks distinguishing high versus low affect states. Despite the promising

results across the RECOLA and AGAIN datasets, several limitations and considerations warrant discussion.

One key methodological limitation lies in the hyperparameter tuning protocol. The study employed a grid-search method to optimise the hyperparameter α, which governs the influence of the teacher model in the LUPI framework. While this approach ensured optimal performance for the datasets under investigation, it assumes that the same α value is effective across all contexts and applications, potentially overlooking dataset-specific nuances. Furthermore, hyperparameter tuning was conducted exclusively for 1-second time windows, with the assumption that the identified α value would generalise to other temporal settings, such as 2- and 3-second windows. This simplification may limit the robustness of the findings, as the optimal teacher influence might vary across different temporal contexts.

Another limitation is the manual selection of privileged features. While these features were chosen based on domain knowledge and task relevance, the manual approach risks overlooking complex, latent structures in the data that automated methods might identify and exploit. Automated feature selection techniques or deep learning-based joint feature extraction could offer richer and more nuanced representations of privileged information. For instance, attention mechanisms or feature-ranking algorithms could dynamically prioritise the most informative features, uncovering subtle patterns that manual curation may miss. However, the computational demands of such methods, coupled with the primary focus of this thesis on rigorously evaluating the benefits of the LUPI paradigm, constrained their inclusion in this work.

The study also focuses exclusively on binary classification tasks, limiting its exploration of broader affect modelling applications. Extending the LUPI paradigm to semi-supervised and self-supervised learning frameworks offers a promising avenue for future work. Semi-supervised approaches could leverage unlabelled data alongside privileged information, enhancing the ability to learn general-purpose representations for diverse affective tasks. Similarly, self-supervised methods could enable models to uncover intrinsic data structures without requiring explicit labels, a particularly advantageous approach for real-world scenarios where labelled data is often scarce or expensive to obtain. However, these extensions would require substantial methodological adjustments, including the development of new evaluation metrics and model architectures tailored to these paradigms.

Lastly, the reliance on fixed time window lengths for temporal modelling introduces another limitation. Static windows of 1, 2, and 3 seconds were used, assuming uniform affective changes across these intervals. As mentioned in the previous section, this approach may fail to capture the dynamic and context-dependent nature of affective

fluctuations, particularly in scenarios like gameplay, where rapid changes often occur. Adaptive or event-based time windowing techniques could better reflect the temporal dynamics of affect, providing a more accurate representation of arousal and valence changes. However, the added complexity and optimisation requirements of these techniques were deemed beyond the scope of this thesis. Future studies could explore such adaptive methodologies to enhance the temporal modelling of affect.

Affect Modelling with Limited Data

Chapter 7 introduced a framework for learning from limited data across multiple domains, focusing on affect modelling in first-person shooter (FPS) games and dyadic interactions, where different games and participants represent distinct domains. The framework addresses the general problem of predicting affect states in new contexts by decomposing it into domain-specific tasks and incorporating domain-specific knowledge into the relabelling of affect labels. This approach was evaluated on the *GameVibe Few-Shot* and *RECOLA Few-Shot* datasets, designed specifically for few-shot experience modelling. Comparative analysis across the datasets' subcorpora and affect dimensions demonstrated that few-shot learning (FSL) models outperform conventional end-to-end classification methods, effectively distinguishing affect states in unseen domains, even with minimal labelled samples. While these findings highlight the promise of the proposed framework, a deeper analysis of the input feature spaces suggests that not all representations are equally suited for FSL tasks. The clustering quality analysis of the input space reveals that the proposed approach suffers from poorly defined domains. This suggests that the success of the FSL framework depends not only on the methodology but also on the quality of the input space.

The study has several methodological limitations that warrant discussion. First, the framework's reliance on FPS games introduces a degree of domain specificity that may limit its applicability to other gaming genres or affective contexts. FPS games are characterised by high-paced, action-driven mechanics, which may elicit specific affective responses, such as heightened arousal or stress. However, these dynamics may not generalise to other genres, such as puzzle or strategy games, where affective states are influenced by slower-paced decision-making or problem-solving processes. This limitation raises questions about whether the framework's findings can be extended to contexts where the affective responses are qualitatively different. By focusing solely on FPS games, the study may overlook genre-specific affective dynamics, which could limit the broader applicability of the proposed framework to diverse gaming contexts.

Second, the framework does not account for individual viewer characteristics, such

as personal interests, skill levels, or prior gaming experience. These factors are widely acknowledged as critical determinants of engagement and affective states, as they shape how individuals perceive and interact with gaming content. For example, a highly skilled player may experience flow and immersion, while a novice may feel frustration in the same scenario. Similarly, personal preferences for certain game mechanics or aesthetics can significantly modulate affective responses. The omission of such individualised factors means that the framework assumes a "one-size-fits-all" approach, potentially leading to less accurate predictions. This limitation underscores the need for a more nuanced modelling approach that incorporates personalised features to capture the diverse and subjective nature of engagement and affect.

Another methodological limitation lies in the reliance on predefined domain labels for the few-shot learning tasks. The accuracy and validity of these domain labels are critical for effective domain-specific adaptation, as they influence how the model clusters and processes data. However, the process of defining domains is inherently subjective and may not adequately reflect the underlying structure of the data. For instance, domains defined based on game types or participants may not align with the true affective or behavioural patterns in the data, leading to poorly defined clusters. This issue is compounded in datasets with overlapping or ambiguous domain boundaries, where the lack of clear separation can hinder the model's performance. The dependence on such manually defined domain labels highlights a methodological constraint that could affect the scalability and robustness of the framework in more complex or less structured datasets.

8.3 | Future Work

This section explores the potential future research directions and extensibility of the methodologies introduced in this thesis, facilitating their application to a variety of domains and use cases. By examining how these methods can be extended and applied in different contexts, this discussion highlights opportunities for advancing both the methods themselves and the field as a whole.

Contrastive Representations of Affect

The findings presented in this chapter underscore the potential of SCL for developing representations that effectively retain affect-infused information. The results obtained through the classification of high versus low affective states across two distinct datasets, RECOLA and AGAIN, provide a solid foundation for extending this work into new di-

rections and applications. One promising avenue for future research is the incorporation of full-colour images in the visual input pipeline. Expanding beyond greyscale frames could enable models to capture richer visual features, particularly in scenarios where colour information plays a critical role in affect recognition. For instance, datasets with highly dynamic and colourful environments, such as AGAIN, may benefit from the additional cues provided by colour features.

Future work could also focus on integrating advanced data augmentation techniques to improve the generalisability and robustness of the learned representations. Techniques such as random cropping, flipping, brightness adjustments, and perturbations of numerical features could be systematically applied to both visual and numerical inputs. These augmentations would introduce variability during training, enhancing the model's ability to generalise to diverse and unseen data. Such strategies could be particularly valuable for datasets with constrained diversity, like AGAIN, and represent an important step towards creating more adaptable and reliable models. Another direction involves expanding the scope of affect modelling beyond binary classification. Exploring regression tasks or multi-class classification could allow for more nuanced affect recognition and a deeper understanding of affective states. This would involve developing new evaluation frameworks and adapting the experimental setup to handle finer granularity in affective data. Investigating these tasks would provide valuable insights into the broader applicability of SCL for affect representation learning and its potential for modelling complex emotional landscapes.

Temporal modelling presents another rich area for future exploration. While this thesis employed fixed time window lengths, adaptive or event-based time windowing could offer a more precise representation of the dynamic nature of affective states. Such an approach would account for the variability in the timing of affective changes, particularly in contexts like gameplay, where rapid fluctuations are common. Developing methodologies to implement adaptive temporal modelling would enhance the temporal resolution of affect analysis and contribute to a deeper understanding of affective dynamics. Finally, the integration of multimodal data beyond the current modalities used in this thesis could further enrich affect representation learning. Incorporating additional modalities, such as natural language, could provide complementary information and improve the robustness of affective models. This would involve designing multimodal fusion strategies that effectively combine diverse data streams, leveraging their unique contributions to enhance overall model performance.

Learning from Missing Modalities

The findings presented in this chapter highlight the potential of integrating the Learning Using Privileged Information (LUPI) framework and Supervised Contrastive Learning (SCL) for leveraging privileged information in affect modelling. The demonstrated improvements in robustness and performance across the RECOLA and AGAIN datasets provide a solid foundation for extending this work into new research directions and applications.

One promising avenue for future research is the integration of adaptive hyperparameter optimisation methods within the LUPI framework. While this study employed a grid-search protocol to tune the hyperparameter , more advanced approaches, such as Bayesian optimisation or population-based training, could enhance the efficiency and effectiveness of hyperparameter tuning. These methods could dynamically adapt values to dataset-specific characteristics or different time window lengths, thereby improving generalisability and performance across diverse datasets and applications. Another significant direction involves the automation of privileged feature selection. This study relied on manual selection based on domain knowledge, which, while interpretable, may have overlooked latent structures within the data. Future work could explore techniques such as automated feature selection or deep learning-based joint feature extraction to identify richer representations of privileged information. Methods like attention mechanisms or feature-ranking algorithms could dynamically prioritise the most relevant features, uncovering subtle patterns and enhancing model performance. These automated approaches would also improve the scalability and applicability of the LUPI framework in broader contexts.

Another compelling research direction involves extending the application of the LUPI paradigm to semi-supervised and self-supervised learning. These paradigms could enable the development of powerful general-purpose representations for affect modelling, particularly in scenarios with limited labelled data. The former could leverage unlabelled data alongside privileged information, while the latter could discover intrinsic data structures without explicit labels. Such methods would significantly enhance the scalability of affective computing systems, making them more applicable to real-world scenarios where data labelling is costly and time-intensive. Finally, incorporating LUPI into ranking and preference learning paradigms presents an exciting opportunity for advancing ordinal affect modelling tasks. Unlike binary classification, ranking models aim to capture the ordinal nature of affective states, such as levels of arousal or valence. By exploiting privileged information, LUPI-based models could infer subtle affective gradients, improving their ability to model subjective human experiences. De-

veloping methodologies to integrate LUPI into ranking and preference learning would contribute to a more nuanced understanding of affective states, addressing the complexities of ordinal affect modelling.

Affect Modelling with Limited Data

The framework introduced in this thesis offers a versatile approach for learning from limited data across multiple domains, focusing on affect modelling in FPS games and dyadic interactions. While the results demonstrate the promise of the proposed framework, they also pave the way for several exciting avenues of future research and potential extensions.

A key direction for future work is the incorporation of personalised features into the framework. Current predictions of affect states do not account for individual differences, such as personal interests, skill levels, or prior gaming experience, which can significantly influence engagement. Integrating personalised features could enhance the accuracy of engagement predictions, thereby improving the framework's applicability to diverse gaming contexts. This would involve collecting and incorporating user-specific data to model individual affective responses more effectively. Expanding the framework to other game genres, such as arcade, sports, puzzle, or strategy games, represents another important research direction. By applying the framework to these varied domains, researchers can explore more nuanced engagement patterns and assess its generalisability. Such extensions could reveal whether the domain-specific relabelling strategies and few-shot learning (FSL) approaches developed in this thesis are adaptable to a broader range of contexts.

Future research should also explore alternative evaluation metrics beyond accuracy. While accuracy is a widely used standard, it does not capture critical aspects of real-world performance, such as robustness to noisy data, computational efficiency, and scalability. Incorporating metrics that evaluate these dimensions, along with an analysis of trade-offs such as training time and resource demands, would provide a more comprehensive understanding of the framework's effectiveness and practical utility. Another promising avenue is testing the framework across alternative modalities. For example, incorporating textual features or graph-based representations could offer new perspectives on affect modelling. These modalities might capture complementary information that enhances the framework's performance and robustness.

The framework could also be extended to address more complex scenarios, such as domain-incremental learning, where the model must detect and adapt to new domains without explicit labels. This would involve developing methods for continuous learn-

ing and adaptation, allowing the model to remain effective in dynamically changing environments. Such advancements would significantly enhance the framework's applicability to real-world scenarios, where new domains and data types are introduced over time. Lastly, exploring applications beyond affect state prediction is another direction worth pursuing. The framework's versatility suggests it could be adapted for tasks such as few-shot segmentation, preference learning, or agent-based gameplay modelling. Extending the methodology to these areas would not only broaden its impact but also uncover novel insights into the potential of few-shot learning in diverse machine learning tasks.

8.4 | Ethical Impact and AI Act

The development and evaluation of emotion recognition systems raise material, ethical and regulatory considerations. Under the EU Artificial Intelligence (AI) Act, emotion recognition is regulated as a high-risk application, and certain uses are outright prohibited (e.g., emotion inference in workplaces and educational settings). Accordingly, any claim of practical applicability must be read in light of these requirements and prohibitions. This thesis should therefore be interpreted as a methodological contribution rather than a deployment blueprint. The reported systems were trained and tested on research datasets in controlled experimental settings; no systems are placed on the market, put into service, or deployed in real-world contexts. Any translation of these methods to practice would require prior demonstration of compliance with the AI Act's obligations for high-risk AI, and strict avoidance of prohibited use cases.

In particular, and without limitation, providers and deployers seeking to operationalise methods related to this work would need to: (i) establish and maintain a documented risk-management system (Article 9); (ii) satisfy data quality and data-governance requirements, including representativeness and relevance of data and documented data handling (Article 10); (iii) prepare and maintain technical documentation (Article 11); (iv) enable appropriate logging and record-keeping (Article 12); (v) provide transparency information and clear instructions to deployers and users (Article 13); (vi) ensure effective human oversight (Article 14); and (vii) meet accuracy, robustness, and cybersecurity requirements (Article 15). These obligations apply before placing a high-risk system on the market and throughout its lifecycle.

Furthermore, the Act prohibits certain emotion-recognition uses entirely—most notably inference of emotions in the workplace and in educational institutions—and these contexts are therefore out of scope for any future deployment derived from this thesis.

Any pilot or productisation must begin with a use-case screening to confirm that it is not a prohibited practice under Article 5, followed by a conformity-assessment pathway appropriate to the system's category.

Ethically, emotion recognition demands particular care regarding bias, explainability, and user agency. While the thesis investigates methods that can improve robustness and generalisation under research conditions, deployment-grade systems should additionally include: (a) bias and performance audits across relevant sub-populations; (b) clear communication of model purpose, limitations, and uncertainty; (c) meaningful human-in-the-loop controls (override, fallback, escalation); and (d) continuous post-market monitoring and incident reporting once in service. These safeguards complement, rather than replace, the legal obligations above.

In summary, this thesis provides evidence that representation-learning strategies (e.g., supervised contrastive learning and learning using privileged information) can strengthen methodological foundations for affect modelling in research settings. Any real-world deployment would require a separate compliance engineering effort to satisfy the AI Act, explicit exclusion of prohibited settings, and an end-to-end governance framework covering risk management, transparency, human oversight, and lifecycle monitoring.

8.5 | Summary

This chapter synthesised the core contributions, limitations, and future directions of the thesis, which sought to advance affective computing through innovative methodologies in representation learning. By addressing challenges such as data scarcity, domain generalisation, and multimodal integration, the research leveraged techniques like Supervised Contrastive Learning (SCL), Few-Shot Learning (FSL), and the Learning Using Privileged Information (LUPI) framework to propose scalable and robust solutions. The study introduced methods that enhance multimodal representation learning, enable generalisation across domains with limited labelled data, and utilise privileged information to bridge gaps between controlled and real-world scenarios. These approaches were rigorously evaluated, demonstrating their effectiveness in improving affect modelling across diverse datasets and tasks. The chapter also reflected on the broader implications of these contributions while acknowledging methodological limitations, such as the reliance on greyscale inputs, fixed time window lengths, and binary classification. Despite these constraints, the proposed methodologies offer significant potential for future research. Directions for further exploration include adaptive hyperparame-

ter optimisation, automated feature selection, and extending the frameworks to semi-supervised and self-supervised learning paradigms. Additionally, expanding the application scope to include domain-incremental learning and multimodal fusion presents opportunities to broaden the impact of these innovations. By laying a strong foundation, this thesis provides a pathway for advancing affective computing and related fields through robust, data-efficient, and adaptable methodologies.

A

Contrasting Representations of Affect

This chapter presents additional experiments regarding the appropriateness of SCL for affective computing, the influence of labels in the downstream task and the use of encoders based on ViT for modelling affect from frames.

A.1 | Supervised Contrastive Learning for Affect Modelling

Table A.1: Accuracy (%) of different representation learning approaches on the RECOLA dataset for arousal prediction (1-second windows). Results are reported for unimodal (audio, visual, physiology) and multimodal fusion settings. The baseline (B) refers to a supervised model trained with cross-entropy. The autoencoder reflects reconstruction-based unsupervised learning, while SimCLR and Barlow Twins represent unsupervised contrastive frameworks. SCL corresponds to supervised contrastive learning. Bold values correspond to the model with the highest accuracy on test set

RECOLA	**Arousal (1 sec)**			
Model	Audio	Visual	Physiology	Fusion
B	71.63	63.72	57.85	68.23
Autoencoder	70.15	61.37	56.58	65.92
SimCLR	69.52	63.22	55.18	66.00
BarlowTwins	68.58	62.48	54.90	65.38
SCL	**75.14**	**65.88**	**60.08**	**71.41**

The results presented in Table A.1 provide a comparative evaluation of reconstruction-based, unsupervised, and supervised contrastive approaches on the RECOLA dataset for arousal prediction. Several important patterns can be observed. First, the baseline

supervised model (B) trained with cross-entropy provides a strong point of comparison, particularly in the audio modality (71.63) and in the multimodal fusion setting (68.23). As expected, the autoencoder performs slightly below the baseline in all modalities. This outcome is consistent with the fact that reconstruction-based objectives primarily capture dominant statistical regularities in the input (e.g., spectral structure in audio, pixel intensities in visual signals, or low-level physiological dynamics) rather than affect-specific information. While such features retain some affect-relevant variance, the absence of explicit discriminative pressure results in weaker performance compared to a supervised baseline.

Second, unsupervised contrastive methods (SimCLR and Barlow Twins) perform at or below the baseline, with the gap being most evident in the physiology and fusion conditions. This finding is in line with theoretical expectations: affective information is not always directly observable in the raw input, and therefore unsupervised contrastive frameworks risk encoding invariances that are orthogonal to emotional content. SimCLR, for instance, constructs positive and negative pairs based on data augmentations, which may emphasise low-level similarities such as speaker identity, background, or visual appearance, without aligning the embedding space to affective dimensions. Barlow Twins applies a redundancy-reduction objective that enforces strong decorrelation between feature dimensions, but in doing so may suppress subtle affect-related cues, especially under partial feature gating. The feature gating (10% feature dropout) applied for constructing different views of the same sample further increases the difficulty of extracting affective information, compounding the performance gap with respect to supervised methods.

Finally, supervised contrastive learning (SCL) outperforms all alternatives for every modality. The consistent gains over the baseline confirm that SCL provides an advantage by explicitly using affective labels to structure the embedding space. Rather than relying on proxy similarities or reconstruction objectives, SCL enforces a relational inductive bias in which samples with similar labels are drawn together and those with divergent labels are pushed apart. From an information-theoretic perspective, SCL aims to maximise the mutual information between embeddings and affective annotations, thereby ensuring that the learned representations encode variance that is specifically relevant to emotional states. These results highlight both the limitations of unsupervised approaches for affect modelling and the importance of label-guided representation learning. While autoencoder, SimCLR, and Barlow Twins capture broad input structure or invariances, they do not consistently align the embedding space with affective information. In contrast, SCL leverages annotation signals to produce representations that are both discriminative and robust across participants.

A.2 | The Influence of Affective Labels in SCL

A.2.1 | Defining Contrastive Labels

Apart from *affect state* (g_a) that corresponds to the mean of the affect trace captured within a time window (Eq. A.1) Makantasis et al. (2021a); Melhart et al. (2022). This work explored two additional measures, the *affect change* score (c_a) which is the average of the absolute differences between consecutive annotation values (Eq. A.2), and the *affect trend* score (t_a) as the average of the differences between consecutive annotation values (Eq. A.3):

$$g_a = \frac{1}{w}\sum_{i=0}^{w} v_i \tag{A.1}$$

$$c_a = \frac{1}{w}\sum_{i=1}^{w} |v_i - v_{i-1}| \tag{A.2}$$

$$t_a = \frac{1}{w}\sum_{i=1}^{w} (v_i - v_{i-1}) \tag{A.3}$$

where w is the window size considered and v_i is the i-th annotation value of the time window.

Building on the measures described earlier, we investigate three distinct strategies for selecting positive and negative samples, which are outlined in detail below. It is important to emphasise that all the proposed contrastive labelling strategies utilize the same loss function for training the SCL models: the supervised contrastive loss.

A.2.1.1 | Contrasting Affect: High vs. Low

Contrastive labels can be intuitively constructed by treating windows with similar affect states as positive pairs and those with differing affect states as negative pairs. To quantify affect state similarity, we binarise the affect states g_a into "high" and "low" categories, assigning the same label to similar states and different labels to dissimilar states. The binarisation process relies on the median value of the ground truth affect annotations across the dataset ($\tilde{g}_a$) and incorporates a threshold ϵ. Specifically, a time window i is classified as "high" if $g_{a_i} > \tilde{g}_a + \epsilon$ and as "low" if $g_{a_i} < \tilde{g}_a - \epsilon$. The threshold ϵ, inspired by its use in works like Makantasis et al. (2019), helps exclude windows with ambiguous affect values near the median, as such ambiguities can compromise the stability of the SCL models and the quality of the learned representations. The filtered

dataset, which excludes these ambiguous annotations, serves as the foundation for all three contrastive labelling strategies.

A.2.1.2 | Contrasting Affect: Change vs. Unchanged

While the high-low pairing strategy relies on an absolute measure of affect, a similar binarisation approach can be applied based on *affect change* (see Eq. A.2), which is a relative measure. Specifically, a time window i is labelled as "change" if $c_{a_i} > \tilde{c}_a$ and as "unchanged" (i.e., no change) if $c_{a_i} \leq \tilde{c}_a$. Here, $\tilde{c}_a$ represents the median affect change value across the entire set of affect change traces. Using the median $\tilde{c}_a$ for binarisation ensures a balanced dataset. As in Section A.2.1.1, these labels are used to pair windows i and j as positive when both exhibit affect change or both exhibit no change, and as negative when one window exhibits affect change while the other does not.

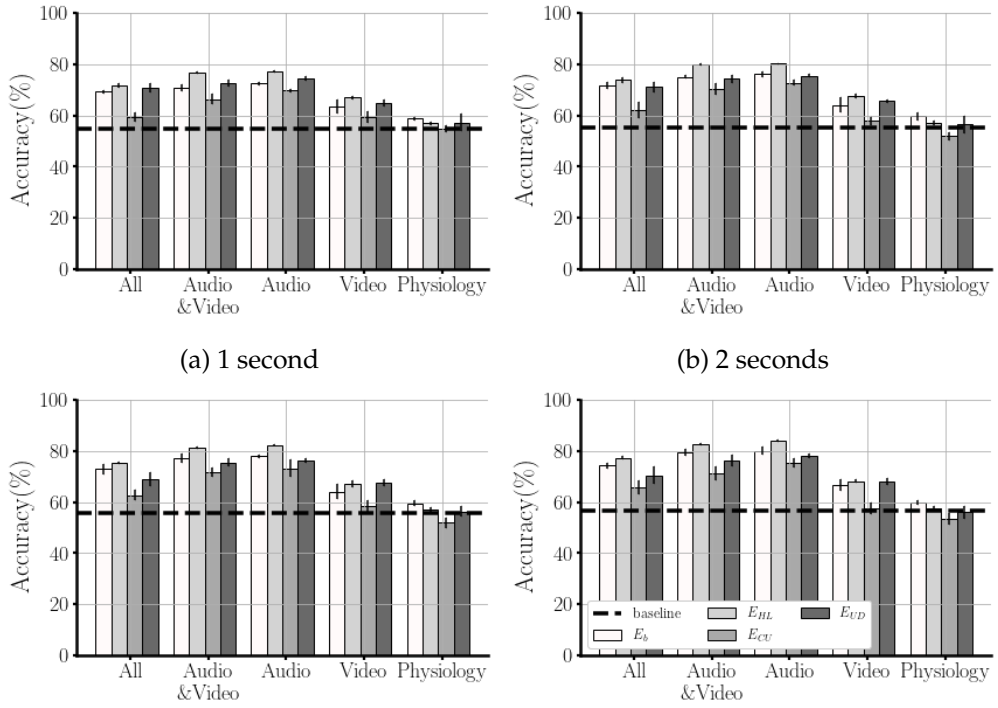
A.2.1.3 | Contrasting Affect: Uptrend vs. Downtrend

Inspired by Yannakakis et al. (2018) and following a similar approach to the pairing strategy described in Section A.2.1.2, this contrastive labelling strategy utilises a relative measure of affect, referred to as *affect trend*. Specifically, the i-th time window is assigned to the "uptrend" class if $t_{a_i} > \tilde{t}_a$; otherwise, it is assigned to the "downtrend" class. Here, $\tilde{t}_a$ denotes the median affect trend value across the entire set of affect trend traces. As before, these labels are used to define positive and negative pairs based on their class membership: samples from the same class form positive pairs, while samples from different classes form negative pairs.

The key distinction between the first contrastive labelling strategy and the latter two lies in their methodological approach. The first strategy is direct, as the "high" and "low" labels are derived directly from the actual magnitude of the affect annotation trace (see Eq. A.1). Conversely, the latter two strategies are indirect, as both "change" and "trend" are higher-order traces that represent the average absolute rate of change (Eq. A.2) and the curvature (Eq. A.3) of the annotation trace, respectively. Despite these differences, the binarisation criterion for all three strategies is based on the annotation traces of the entire affect corpus.

A.2.2 | Experiments

We wish to investigate how representations learned via contrastive learning perform for a downstream task classifying between high and low arousal states. We thus train three encoders via SCL as per our three contrastive labeling strategies: high-low arousal



(a) 1 second

(b) 2 seconds

(c) 3 seconds

(d) 4 seconds

Figure A.1: Average 5-fold validation accuracy scores (%) for high-low arousal classification as a downstream task. Values are averaged across 10 independent runs; 95% confidence intervals are displayed as error bars.

state (E_{HL}), arousal change-unchanged (E_{CU}), and uptrend-downtrend (E_{UD}). We train a probe model for each of the encoders as mentioned in Section 5.2. The baseline model, E_b, performs end-to-end arousal classification. An additional baseline always chooses the most frequent class in the training set (dotted line in Fig. A.1).

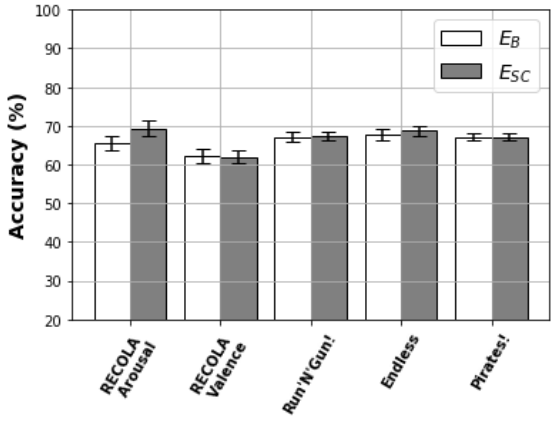
A.3 | The Influence of the Frame Encoder in Frame-based Affect Modelling

A.3.1 | Vision Transformer

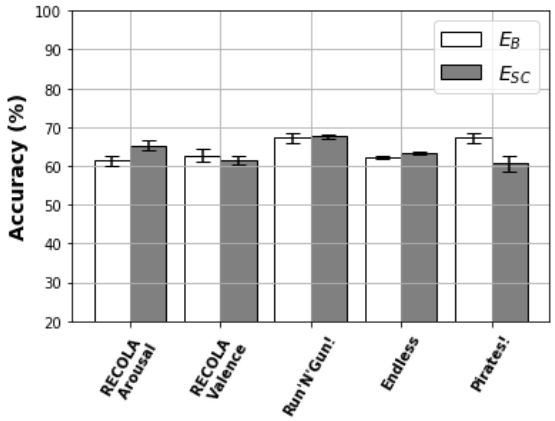
A Transformer is an architecture that leverages an attention mechanism to capture dependencies between input and output elements. While Transformers still employ an encoder-decoder structure, they eliminate recurrence, resulting in reduced training time and improved performance compared to other sequence transduction models. The Vision Transformer (ViT) is a Transformer-based architecture designed for image classification tasks. It takes a single image as input, maps it to a high-level vector representation, and passes this representation to a multilayer perceptron responsible for performing the classification task.

A.3.2 | Experiments

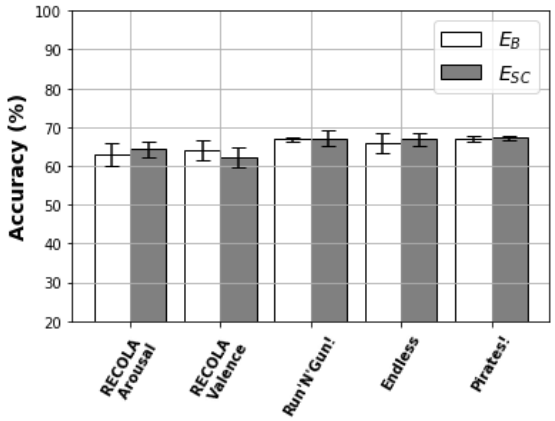
Drawing inspiration from previous studies Pinitas et al. (2022a, 2023), we employ a small ViT with two transformer blocks as the encoder to learn high-level vector representations of affect from sequences of frames. Since each time window contains 5 grayscale frames per second, the $5 \times 224 \times 224$ tensor of pixel values (normalised to the range $[0, 1]$) for a 1-second time window is transformed into a vector of 768 real values through this process.



(a) 1 second



(b) 2 seconds



(c) 3 seconds

Figure A.2: Frame-based affect modelling for RECOLA and AGAIN using a ViT encoder. The bars represent the average 5-fold validation accuracy scores (%) for high-low arousal classification as a downstream task. 95% confidence intervals are displayed as error bars.

B

Learning from Missing Modalities

This chapter presents additional experiments regarding the influence of the hyperparameter α in the LUPI process.

B.1 | Influence of Teacher Importance Hyperparameter

Table B.1: The effect of α parameter on students' average binary classification accuracy (%) on the RECOLA dataset when the Feature (top) and Fusion (bottom) models are used as teachers. Bold values indicate the highest classification accuracy achieved across all different values of α.

Feature Teacher	**Arousal**		**Valence**	
Student ($\alpha = 0$)	59.41	59.41	60.84	60.84
Student ($\alpha = 0.25$)	**63.90**	**64.61**	58.02	58.14
Student ($\alpha = 0.5$)	59.88	59.94	**60.98**	**61.58**
Student ($\alpha = 0.75$)	61.78	62.38	59.98	61.45
Student ($\alpha = 1$)	51.25	50.95	56.71	55.68
Fusion Teacher	**Arousal**		**Valence**	
Student ($\alpha = 0$)	59.41	59.41	60.84	60.84
Student ($\alpha = 0.25$)	60.14	60.68	60.80	**62.16**
Student ($\alpha = 0.5$)	**61.06**	61.71	61.91	61.40
Student ($\alpha = 0.75$)	59.82	**62.03**	**62.22**	62.38
Student ($\alpha = 1$)	57.18	56.78	59.80	57.61

In this section, we analyze the influence of the teacher model on the performance of the student model, modulated by the parameter α in Eq. (3.2). Specifically, we evaluate two teacher models: the Feature Teacher and the Fusion Teacher. These models are initially trained to model affect. In particular we employed two training methods.

Table B.2: The effect of α parameter on students' average binary classification accuracy (%) on the AGAIN dataset when the Feature (top) and the Fusion (bottom) models are used as teachers. Bold values indicate the highest classification accuracy achieved across all different values of α.

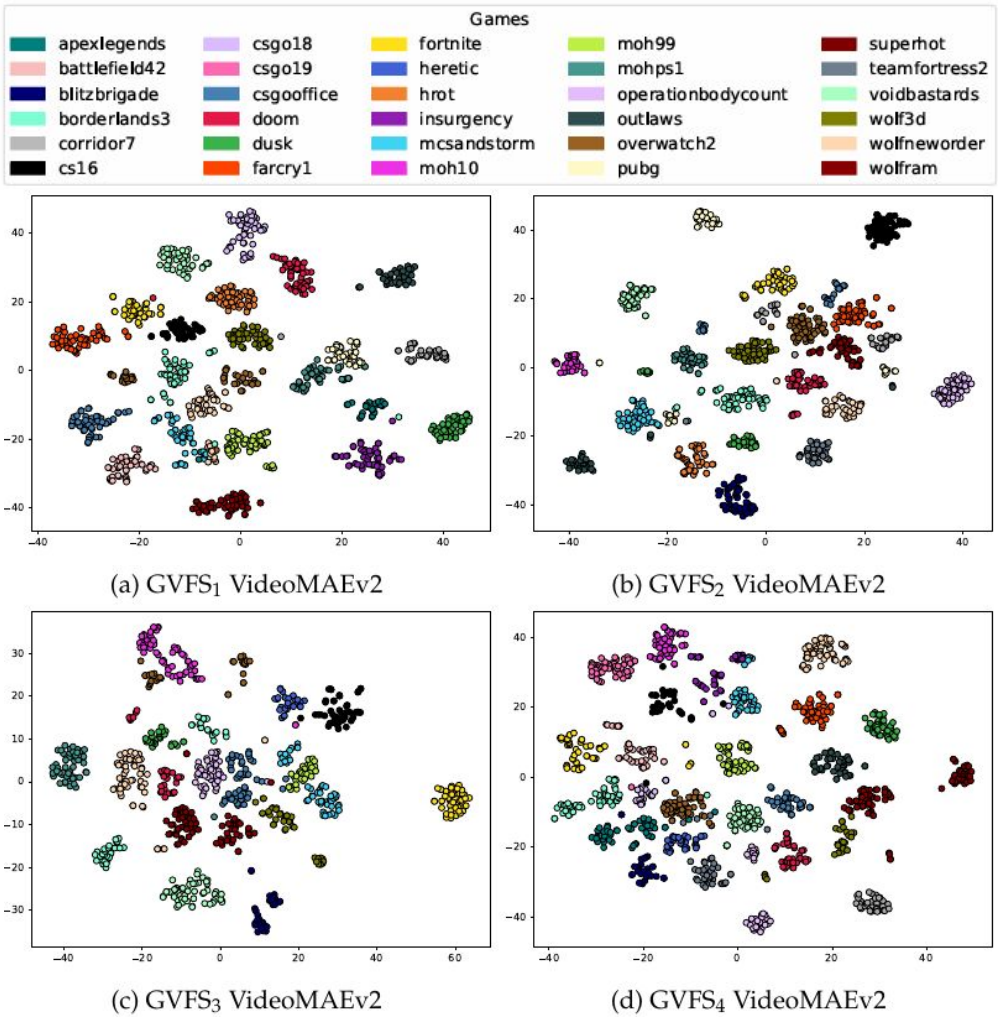
Feature Teacher	**Run'N'Gun!**		**Pirates!**		**Endless**	
Student ($\alpha = 0$)	67.40	67.40	67.63	67.63	65.96	65.96
Student ($\alpha = 0.25$)	67.32	68.25	**67.47**	**67.90**	65.99	66.37
Student ($\alpha = 0.5$)	68.92	68.78	67.31	67.22	65.58	65.27
Student ($\alpha = 0.75$)	**70.63**	**71.15**	66,87	66.11	**66.67**	**67.36**
Student ($\alpha = 1$)	59.26	61.58	52.41	53.09	56.13	57.09
Fusion Teacher	**Run'N'Gun!**		**Pirates!**		**Endless**	
Student ($\alpha = 0$)	67.40	67.40	67.63	67.63	65.96	65.96
Student ($\alpha = 0.25$)	68.05	69.30	**67.26**	**67.87**	65.42	65.51
Student ($\alpha = 0.5$)	68.60	68.59	63.51	64.78	67.17	66.84
Student ($\alpha = 0.75$)	**68.87**	**69.69**	65.66	67.16	**67.73**	**67.35**
Student ($\alpha = 1$)	60.10	60.53	55.38	56.02	58.00	59.38

The conventional apporach of end-to-end classification and teacher prtraining via SCL. Subsequently, student models are trained using five distinct values for the parameter α: $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. When $\alpha = 0$, the student model relies solely on the ground truth labels, without utilizing any privileged information. Conversely, with $\alpha = 1$, the student exclusively follows the teacher's guidance, completely disregarding the ground truth labels.

The results of this investigation are summarized in Tables B.1 and B.2. It is important to note that the abbreviation SC refers to students trained under the LUPI paradigm, where the teacher model is pretrained using Supervised Contrastive Learning (SCL). In contrast, B denotes students trained with teachers that were pretrained as end-to-end classifiers. For this investigation, a time window of 1 second was employed.

C

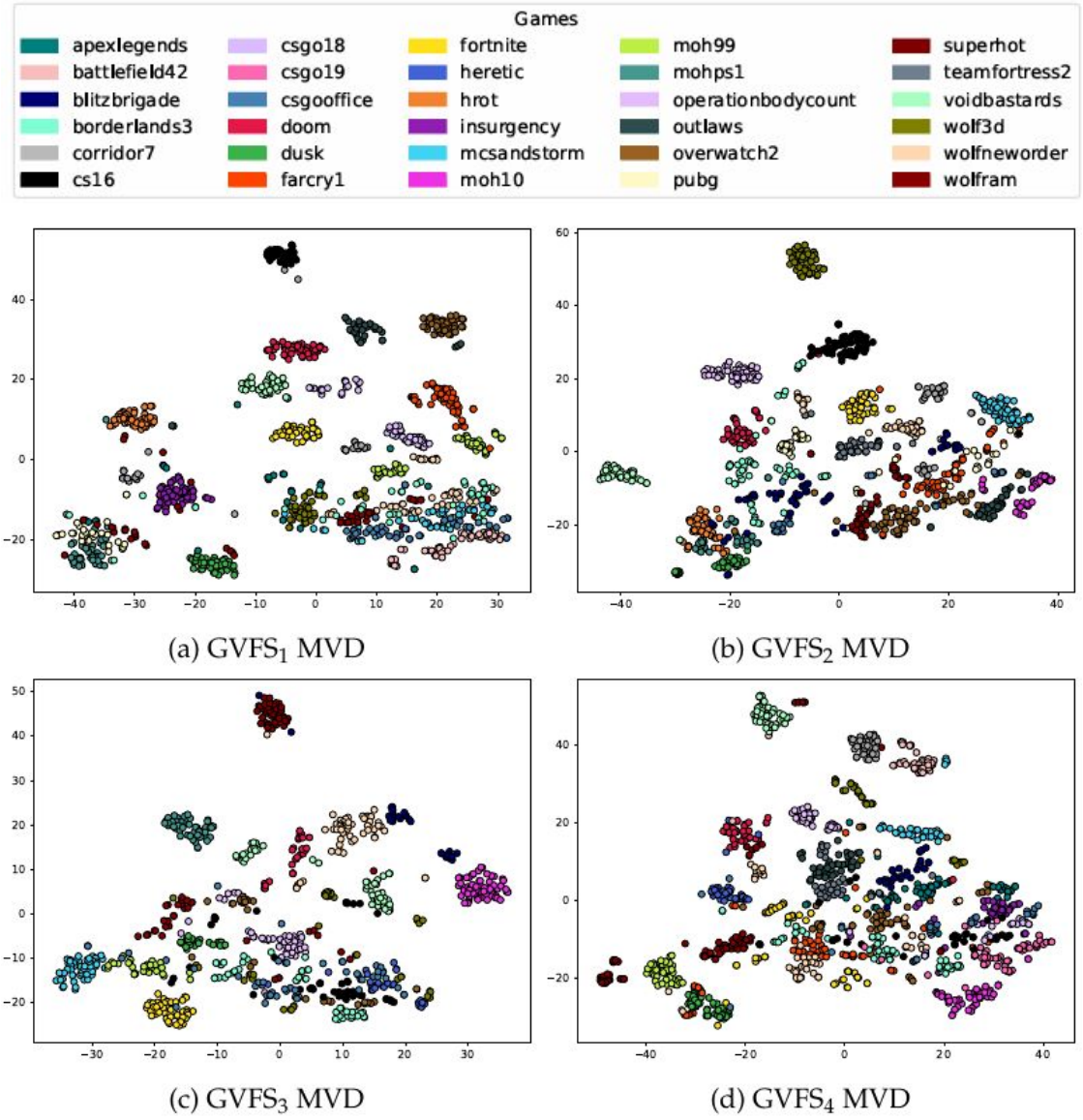
Learning with Limited Data



(a) GVFS$_1$ VideoMAEv2 (b) GVFS$_2$ VideoMAEv2

(c) GVFS$_3$ VideoMAEv2 (d) GVFS$_4$ VideoMAEv2

Figure C.1: **GVFS Dataset** t-SNE plot illustration of the input space as shaped by the VideoMAEv2 model. The different colours correspond to FPS games.



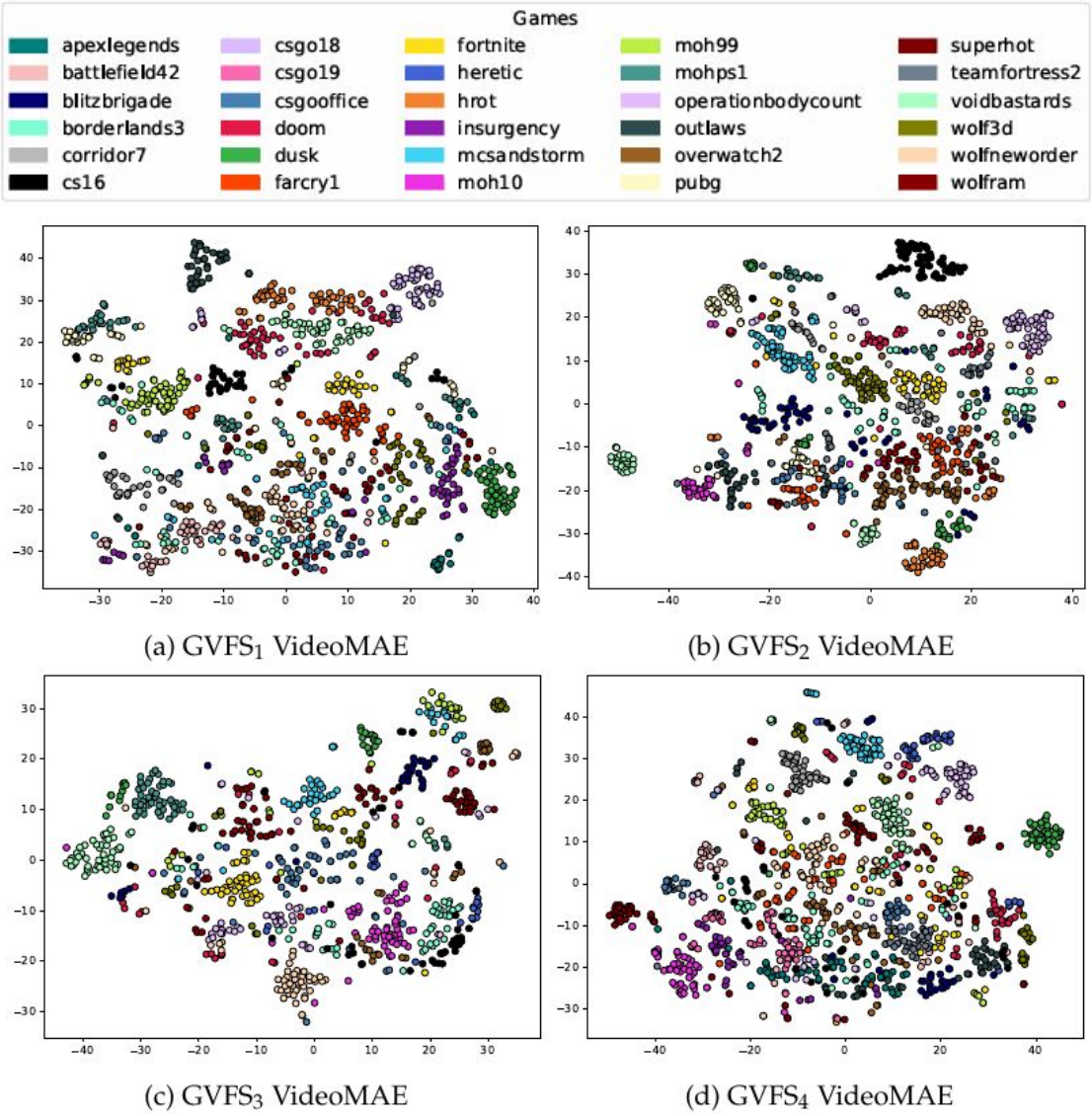
(a) $GVFS_1$ MVD

(b) $GVFS_2$ MVD

(c) $GVFS_3$ MVD

(d) $GVFS_4$ MVD

Figure C.2: **GVFS Dataset** t-SNE plot illustration of the input space as shaped by the MVD model. The different colours correspond to FPS games.



(a) $GVFS_1$ VideoMAE (b) $GVFS_2$ VideoMAE

(c) $GVFS_3$ VideoMAE (d) $GVFS_4$ VideoMAE

Figure C.3: **GVFS Dataset** t-SNE plot illustration of the input space as shaped by the VideoMAE model. The different colours correspond to FPS games.

References

Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58, 2021.

Jacob Aday, Will Rizer, and Joshua M. Carlson. Chapter 2 - neural mechanisms of emotions and affect. In Myounghoon Jeon, editor, *Emotions and Affect in Human Factors and Human-Computer Interaction*, pages 27–87. Academic Press, San Diego, 2017. ISBN 978-0-12-801851-4. doi: https://doi.org/10.1016/B978-0-12-801851-4.00002-1. URL `https://www.sciencedirect.com/science/article/pii/B9780128018514000021`.

Ralph Adolphs and David J Anderson. The neuroscience of emotion: A new synthesis. 2018.

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

James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of school psychology*, 44(5):427–445, 2006.

Matthew Barthet, Chintan Trivedi, Kosmas Pinitas, Emmanouil Xylakis, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Knowing your annotator: Rapidly testing the reliability of affect annotation. In *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos*, 2023.

Matthew Barthet, Maria Kaselimi, Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Gamevibe: A multimodal affective game corpus. *arXiv preprint arXiv:2407.12787*, 2024.

Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Proc. of the IEEE Int. Conf. on affective computing and intelligent interaction*, pages 77–83, 2015.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*, 2021.

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Uta K Bindl and Sharon K Parker. 32 feeling good and performing well? psychological engagement and positive behaviors at work. *Handbook of employee engagement: Perspectives, issues, research and practice*, 385, 2010.

Anthony F. Botelho, Ryan S. Baker, and Neil T. Heffernan. Improving Sensor-Free Affect Detection Using Deep Learning. In Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 40–51, Cham, 2017. Springer International Publishing. ISBN 978-3-319-61425-0. doi: 10.1007/978-3-319-61425-0_4.

Margaret M Bradley and Peter J Lang. Measuring emotion: Behavior, feeling, and physiology. In *Cognitive neuroscience of emotion*, page 242–276. Oxford University Press, 2000.

Aihua Cai, Wenxin Hu, and Jun Zheng. Few-shot learning for medical image classification. In *International Conference on Artificial Neural Networks*, pages 441–452. Springer, 2020.

Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

Elizabeth Camilleri, Georgios N Yannakakis, and Antonios Liapis. Towards general models of player affect. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2017.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10492–10500, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xuanchi Chen, Xiangwei Zheng, Kai Sun, Weilong Liu, and Yuang Zhang. Self-supervised vision transformer-based few-shot learning for facial expression recognition. *Information Sciences*, 634:206–226, 2023.

Anca-Nicoleta Ciubotaru, Arnout Devos, Behzad Bozorgtabar, Jean-Philippe Thiran, and Maria Gabrani. Revisiting few-shot learning for facial expression recognition. *arXiv preprint arXiv:1912.02751*, 2019.

Mohamed Dahmane and Jean Meunier. Emotion recognition using dynamic grid-based hog features. In *Proc. of the IEEE Int. Conf. on Automatic Face & Gesture Recognition*, pages 884–888, 2011.

Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. In *Proceedings of the International Conference on Multimodal Interaction*, pages 440–445, 2019.

Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the International Conference on Computer Vision*, pages 1502–1512. IEEE, 2021.

Amitava Dutta. Integrating ai and optimization for decision support: A survey. *Decision Support Systems*, 18(3-4):217–226, 1996.

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

Jing Fan, Dayi Bian, Zhi Zheng, Linda Beuscher, Paul A Newhouse, Lorraine C Mion, and Nilanjan Sarkar. A robotic coach architecture for elder care (rocare) based on multi-user engagement models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(8):1153–1163, 2016.

Kexin Feng and Theodora Chaspari. Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing*, 14(2):1627–1633, 2021.

Jan Feyereisl and Uwe Aickelin. Privileged information for data clustering. *Information Sciences*, 194:4–23, 2012.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Julian Frommel, Claudia Schrader, and Michael Weber. Towards emotion-based adaptive games: Emotion recognition via input and performance features. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 173–185, 2018.

Giorgos Giannakakis, Eleftherios Trivizakis, Manolis Tsiknakis, and Kostas Marias. A novel multi-kernel 1d convolutional neural network for stress recognition from ecg. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. IEEE, 2019.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

Ian Goodfellow. Deep learning, 2016.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, 7(4): 774, 2007.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3071–3074, 2019.

Jad Haddad, Olivier Lézoray, and Philippe Hamel. 3d-cnn for facial emotion recognition in videos. In *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15*, pages 298–309. Springer, 2020.

Ross Harper and Joshua Southern. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Transactions on Affective Computing*, 13(2):985–991, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Stat*, 1050, 2015.

Christoffer Holmgård, Georgios N Yannakakis, Hector P Martinez, and Karen-Inge Karstoft. To rank or to classify? annotating stress for reliable ptsd profiling. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2015.

Yan Huang, Stefanus Jasin, and Puneet Manchanda. "level up": Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research*, 30(3):927–947, 2019.

Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the International Conference on Multimedia*, pages 801–894. Association for Computing Machinery, 2014.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multitask emotion recognition model with knowledge distillation and task discriminator. *arXiv preprint arXiv:2203.13072*, 2022.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Contrastive learning for prompt-based few-shot language learners. *arXiv preprint arXiv:2205.01308*, 2022.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille, 2015.

Soonil Kwon et al. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167:114–177, 2021.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 53:95–108, 2014.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017a.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020a.

Lin Li, Tiong-Thye Goh, and Dawei Jin. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Computing and Applications*, 32(9):4387–4415, May 2020b. doi: 10.1007/s00521-018-3865-7.

Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6329–6333. IEEE, 2021.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017b.

Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8635–8643, 2021.

Huan Liu, Ke Li, Jianping Fan, Caixia Yan, Tao Qin, and Qinghua Zheng. Social image-text sentiment classification with cross-modal consistency and knowledge distillation. *IEEE Transactions on Affective Computing*, 2022.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. Ranktrace: Relative and unbounded affect annotation. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2017.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. 2016.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *arXiv preprint arXiv:2109.01797*, 2021.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2022.

Konstantinos Makantasis. Affranknet+: ranking affect using privileged information. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2021.

Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. From pixels to affect: a study on games and player experience. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2019.

Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. The pixels and sounds of emotion: General-purpose representations of arousal in games. *IEEE Trans. on Affective Computing*, 2021a.

Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. Privileged information for modeling affect in the wild. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, 2021b.

Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, and Georgios N Yannakakis. From the lab to the wild: Affect modeling via privileged information. *IEEE Transactions on Affective Computing*, 2023.

Regan L Mandryk and M Stella Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, 65(4): 329–347, 2007.

Héctor P Martínez and Georgios N Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*, pages 34–41, 2014.

Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. Learning deep physiological models of affect. *IEEE Computational intelligence magazine*, 8(2):20–33, 2013.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. PAGAN: Video affect annotation made easy. In *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction*, pages 130–136, 2019.

David Melhart, Daniele Gravina, and Georgios N Yannakakis. Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, pages 1–10, 2020.

David Melhart, Antonios Liapis, and Georgios N. Yannakakis. The Affect Game AnnotatIoN (AGAIN) dataset. *arXiv preprint arXiv:2104.02643*, 2021a.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. Towards general models of player experience: A study within genres. In *2021 IEEE Conference on Games (CoG)*, pages 01–08. IEEE, 2021b.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. The arousal video game annotation (again) dataset. *IEEE Transactions on Affective Computing*, 13(4):2171–2184, 2022.

David Melhart, Julian Togelius, Benedikte Mikkelsen, Christoffer Holmgård, and Georgios N Yannakakis. The ethics of ai in games. *IEEE Transactions on Affective Computing*, 2023.

Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proc. of the Int. Conf. on multimodal interaction*, pages 443–449, 2015.

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

Run Ning, CL Philip Chen, and Tong Zhang. Cross-subject eeg emotion recognition using domain adaptive few-shot learning networks. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1468–1472. IEEE, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sicheng Pan, Gary J.W. Xu, Kun Guo, Seop Hyeong Park, and Hongliang Ding. Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach. *IEEE Transactions on Games*, pages 1–12, 2023. doi: 10.1109/TG.2023.3348230.

Rosalind W Picard. *Affective computing*. MIT press, 2000.

Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Rankneat: outperforming stochastic gradient search in preference learning tasks. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1084–1092, 2022a.

Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Supervised contrastive learning for affect modelling. In *Proceedings of the International Conference on Multimodal Interaction*, pages 531–539, 2022b.

Kosmas Pinitas, David Renaudie, Mike Thomsen, Matthew Barthet, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Predicting player engagement in tom clancy's the division 2: A multimodal approach via pixels and gamepad actions. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 488–497, 2023.

Kosmas Pinitas, Konstantinos Makantasis, and Georgios N Yannakakis. Across-game engagement modelling via few-shot learning. *arXiv preprint arXiv:2409.13002*, 2024a.

Kosmas Pinitas, Nemanja Rasajski, Konstantinos Makantasis, and Georgios N Yannakakis. Silhouette distance loss for learning few-shot contrastive representations. *Proceedings of Machine Learning Research*, 1: 18, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proc. of the IEEE Int. conf. and workshops on automatic face and gesture recognition*, 2013.

Shuvendu Roy and Ali Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the International Conference. on Multimodal Interaction*, pages 253–257. Association for Computing Machinery, 2021.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *Proceedings of the International Conference on Multimedia and Expo*. IEEE, 2003.

Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 2014.

Nicu Sebe, Ira Cohen, and Thomas S Huang. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*, pages 387–409. World Scientific, 2005.

Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *Proc. of the IEEE Int. Conf. on computer vision*, pages 825–832, 2013.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 2022.

Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3461–3470, 2017.

Debaditya Shome and Tejaswini Kar. Fedaffect: Few-shot federated learning for facial expression recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4168–4175, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Alexis D. Souchet, Stéphanie Philippe, Domitile Lourdeaux, and Laure Leroy. Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review. *International Journal of Human–Computer Interaction*, 38(9):801–824, 2022.

Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing*, 13(2):1037–1043, 2020.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Choo-Yee Ting, Wei-Nam Cheah, and Chiung Ching Ho. Student engagement modeling using bayesian networks. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 2939–2944. IEEE, 2013.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, January 2021. doi: 10.1038/s42256-020-00280-0.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093, 2022.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, March 2016. doi: 10.1109/ICASSP.2016.7472669.

Chintan Trivedi, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Revisiting lp-constrained softmax loss: A comprehensive study. *arXiv preprint arXiv:2206.09616*, 2022.

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

Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, April 2021. doi: 10.1016/j.inffus.2020.10.011.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(1):2023–2049, 2015.

Vladimir Vapnik and Rauf Izmailov. Knowledge transfer in svm and neural networks. *Annals of Mathematics and Artificial Intelligence*, 81(1):3–19, 2017.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023a.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6312–6322, 2023b.

Shangfei Wang, Menghua He, Yachen Zhu, Shan He, Yue Liu, and Qiang Ji. Learning with privileged information using bayesian networks. *Frontiers of Computer Science*, 9:185–199, 2015.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Yu Wang, Mark Cartwright, and Juan Pablo Bello. Active few-shot learning for sound event detection. In *Interspeech*, pages 1551–1555, 2022.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8): 1–38, 2022.

Su Xue, Meng Wu, John Kolen, Navid Aghdaie, and Kazi A Zaman. Dynamic difficulty adjustment for maximized engagement in digital games. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 465–471, 2017.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 2023.

Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017.

Georgios N Yannakakis and David Melhart. Affective game computing: A survey. *Proceedings of the IEEE*, 2023.

Georgios N Yannakakis and Julian Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions. In *Proc of the Int. Conf. on Affective Computing and Intelligent Interaction*, pages 248–255. IEEE, 2017.

Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 12(1):16–35, 2018.

Yufeng Yin, Liupei Lu, Yao Xiao, Zhi Xu, Kaijie Cai, Haonan Jiang, Jonathan Gratch, and Mohammad Soleymani. Contrastive learning for domain transfer in cross-corpus emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Guangyi Zhang and Ali Etemad. Distilling eeg representations via capsules for affective computing. *arXiv preprint arXiv:2105.00104*, 2021.

Na Zhang, Mindi Ruan, Shuo Wang, Lynn Paul, and Xin Li. Discriminative few shot learning of facial dynamics in interview videos for autism trait classification. *IEEE Transactions on Affective Computing*, 14 (2):1110–1124, 2022a.

Tianyi Zhang, Abdallah El Ali, Alan Hanjalic, and Pablo Cesar. Few-shot learning for fine-grained emotion recognition using physiological signals. *IEEE Transactions on Multimedia*, 25:3773–3787, 2022b.

Yuhao Zhang, Md Zakir Hossain, and Shafin Rahman. Deepvanet: A deep end-to-end network for multi-modal emotion recognition. In *Procceedings of the 18th International Conference on Human-Computer Interaction (INTERACT)*, page 227–237, 2021.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *European Conference on Computer Vision*, pages 297–313. Springer, 2022.

Wenming Zheng, Hao Tang, Zhouchen Lin, and Thomas S Huang. Emotion recognition from arbitrary view facial images. In *Proc. of the European Conf. on Computer Vision*, pages 490–503. Springer, 2010.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

Xinyi Zou, Yan Yan, Jing-Hao Xue, Si Chen, and Hanzi Wang. When facial expression recognition meets few-shot learning: A joint and alternate learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5367–5375, 2022.