School of Computer Science
Faculty of Science and Engineering
University of Nottingham
Malaysia

UG FINAL YEAR DISSERTATION REPORT

Optimising Multi-Modal Frameworks for Multi-Label Classification: An Empirical Study on the EMOTIC Dataset

Student's Name	: Tinaabishegan Baladewan
Student Number	: 20408241
Supervisor Name	: Dr. Tissa Chandesa
Year	: 2025

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE (HONS)
THE UNIVERSITY OF NOTTINGHAM

Optimising Multi-Modal Frameworks for Multi-Label Classification: An Empirical Study on the EMOTIC Dataset

Submitted in May 2025, in partial fulfillment of the conditions of the award of the degrees B.Sc.

Tinaabishegan Baladewan

School of Computer Science

Faculty of Science and Engineering

University of Nottingham

Malaysia

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature _____

Date 04 / 05 / 2025

Acknowledgement

I would like to begin by expressing my sincere gratitude to my Final Year Project (FYP) supervisor, Dr Tissa Chandesa, for his invaluable guidance, support, and encouragement throughout my project in these two semesters. I am deeply grateful for his assigning of this topic, which has helped me to gain a better grasp of the computer vision field at hand, and I really appreciate for his tireless efforts in reviewing my project proposal, ethics form, interim report, and final report. His constructive feedback and suggestions have been extremely helpful in improving my research project, and I am grateful for his unwavering commitment to my success.

I would also like to extend my gratitude to all the lecturers in the School of Computer Science who conducted FYP lectures and provided me with the necessary steps to carry out the project, document my work appropriately, and demonstrate my work. Their dedication to teaching and support has been critical to my academic success.

In addition, I would like to thank my classmates and friends for their support and assistance during the research and project development process. Their willingness to help and patience in solving problems have been a great source of motivation for me.

Finally, I am deeply grateful to my family members for their encouragement and understanding throughout my final year. Their constant love and companionship have been my anchor during my ups and downs.

Once again, I express my heartfelt appreciation to everyone who has played a part in my academic journey, and I am forever grateful for their support.

Abstract

Recognising human emotions within their natural context is a complex yet crucial task for advancing human-computer interaction and affective computing. This dissertation addresses the challenge of multi-label emotion classification using the EMOTIC dataset, which features 'in-the-wild' images annotated with context, body, and face regions, alongside 26 discrete emotion categories. This dissertation proposes and systematically evaluates a multi-modal deep learning framework designed to integrate these diverse visual cues while tackling inherent dataset challenges like class imbalance and potential label noise. Through a staged experimental process, this dissertation investigated various deep learning backbones, feature fusion strategies, multi-label loss functions, and data augmentation techniques. The optimal configuration identified employs Swin Transformer backbones for context and body streams, EfficientNet-B7 for the face stream, integrated via a Weighted Fusion mechanism. Training utilised a custom Discrete Loss function with dynamic weighting, basic image-level augmentation (random flips, colour jitter), and ADASYN feature-space resampling to mitigate class imbalance. This optimised framework was evaluated on the held-out EMOTIC test set, achieving a mean Average Precision (mAP) of 32.00%. This result demonstrates the effectiveness of the systematic multi-modal approach, yielding performance comparable to several established benchmarks and highlighting the importance of architectural choices, adaptive fusion, and targeted augmentation for robust emotion recognition in context.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem and Motivation

The ability to perceive, interpret, and respond appropriately to human emotions constitutes a cornerstone of social intelligence, underpinning the nuanced interactions and empathy that characterise effective human relationships. As artificial intelligence (AI) systems become increasingly interwoven into the fabric of daily life, manifesting in applications ranging from conversational agents and personalised healthcare monitoring systems to interactive educational software and sophisticated driver-assistance technologies—imbuing these systems with affective capabilities is recognised as a critical step towards achieving truly natural, intuitive, and supportive human-computer interaction (HCI) (Mobbs et al., 2025; Feng and Chaspari, 2020). The field of Affective Computing, as described by Mobbs et al. (2025), is dedicated precisely to this challenge, developing computational systems capable of recognising, interpreting, processing, and even simulating human affects. Success in this endeavour promises not only to enhance user experience but also to unlock new possibilities in areas requiring social awareness and emotional understanding from machines.

Historically, the computational pursuit of emotion recognition was heavily influenced by early psychological models focusing on basic, universal emotions, often leading to research centred on analysing isolated facial expressions (Kopalidis et al., 2024). Methodologies frequently involved datasets featuring posed, prototypical expressions captured under controlled laboratory conditions. While such studies provided valuable foundational insights into the facial muscle configurations associated with certain emotional states, this paradigm inherently lacks ecological validity. In authentic human experience, emotions are rarely expressed or perceived through the face alone; rather, as Turkstra et al. (2023) note, they are embedded within, and inextricably linked to, a rich tapestry of contextual information encompassing the physical environment, prevailing social dynamics, accompanying body language, and other multimodal signals. Extensive research in psychology and cognitive science, including studies by Turkstra et al. (2023) and Aviezer et al. (2011), has unequivocally demonstrated that context does not merely supplement facial information but actively interacts with it, significantly modulating, and in some cases fundamentally altering, the interpretation of an observed expression. An identical smile, for instance, might be interpreted as joy, amusement, embarrassment, or even contempt depending entirely on the surrounding situational and bodily cues. Consequently, models relying solely on facial analysis are inherently limited in their ability to achieve robust, human-like emotional understanding in complex, real-world settings.

Acknowledging these limitations, the focus within Affective Computing has progressively shifted towards Context-Aware Emotion Recognition (CAER). This paradigm, discussed by Kosti et al. (2019), seeks to develop systems capable of inferring apparent emotional states by adopting a more holistic perspective, considering the individual within their broader environmental and situational context. The development of large-scale datasets specifically designed to facilitate this research has been crucial. Among these, the EMOTIC (Emotions in Context) dataset, introduced by Kosti et al. (2017),

stands out as a significant resource. It comprises thousands of 'in-the-wild' images featuring annotated individuals, captured in diverse, uncontrolled settings (Kosti et al., 2017). EMOTIC's value lies not only in its scale and ecological validity but also in its rich annotation schema, providing bounding boxes for context, body, and face regions, alongside a comprehensive set of 26 discrete emotion categories applied in a multi-label fashion (Kosti et al., 2017).

However, the very realism that makes EMOTIC valuable also introduces substantial challenges that motivate the work presented in this dissertation. Firstly, as highlighted by Costa et al. (2023), the subjective nature of emotion perception, compounded by the complexity of annotating 26 nuanced categories via crowdsourcing, results in considerable label noise and inter-annotator disagreement. Models must therefore be robust to potentially inconsistent or inaccurate ground truth labels. Secondly, the dataset exhibits severe class imbalance, with common emotions like 'Happiness' appearing far more frequently than rarer states like 'Annoyance' or 'Doubt/Confusion' (Audibert et al., 2024). This imbalance can heavily bias standard machine learning models towards majority classes, hindering their ability to recognise less frequent but equally important emotions. Thirdly, the 'in-the-wild' nature means images often contain significant variations in viewpoint, illumination, image quality, subject scale, and, critically, occlusion of facial or body features (Costa et al., 2023). These factors necessitate models that do not rely solely on any single cue but can effectively integrate partial information from multiple sources. Addressing these combined challenges requires sophisticated and robust deep learning methodologies.

1.2 Aim

This dissertation confronts these challenges directly by developing and systematically evaluating a multi-modal deep learning framework explicitly designed for multi-label emotion classification using the EMOTIC dataset. The primary aim of this project is therefore to design, implement, and comprehensively assess a robust system capable of effectively integrating visual information derived from the surrounding context, the subject's body language, and their facial expression to predict the apparent discrete emotions depicted in static images, while addressing the inherent difficulties posed by the dataset.

1.3 Objectives

To realise this aim, the following objectives were systematically addressed through the design and execution of the codebase associated with this project, as detailed in Sections 4 and 5, respectively:

i) **Investigate Feature Extraction Backbones:** Evaluate the representational power of diverse pre-trained deep learning architectures (including CNNs like EfficientNet, MobileNetV3, and Transformers like Swin, DeiT) for capturing emotion-

relevant visual cues from context, body, and face streams independently (Elharrouss et al., 2024).

ii) **Explore Feature Fusion Strategies:** Investigate and compare various techniques (e.g., simple concatenation, weighted fusion, attention mechanisms, cross-modal transformers, bilinear pooling) to determine the most effective method for combining the features extracted from the three separate visual streams (Li and Tang, 2024; Guo et al., 2023).

iii) **Evaluate Multi-Label Loss Functions:** Assess the suitability of different loss functions (e.g., Binary Cross-Entropy, weighted Discrete Loss, Focal Loss, Asymmetric Loss, Dice Loss) for handling the multi-label output and addressing the severe class imbalance present in the EMOTIC dataset (Audibert et al., 2024; Yasuda et al., 2024; Ridnik et al., 2021; Yeung et al., 2023).

iv) **Assess Data Augmentation Techniques:** Investigate the impact of image-level augmentation strategies (geometric and photometric transformations, mixing techniques) and dataset-level resampling methods (specifically ADASYN) on model generalisation, robustness, and performance in the context of class imbalance (Shantharam and Schwenker, 2024; Halim et al., 2023).

v) **Evaluate Overall Performance:** Determine the effectiveness of the optimally configured framework by measuring its performance on the held-out EMOTIC test set using the standard mean Average Precision (mAP) metric (Etesam et al., 2024).

Through a staged experimental process detailed in Section 5, this work identified an optimal configuration combining Swin Transformer backbones for context and body, EfficientNet-B7 for the face, a weighted fusion mechanism, a custom discrete loss function, basic image augmentation, and ADASYN resampling. This optimised framework achieved a mean Average Precision (mAP) of 0.32 on the EMOTIC test set, demonstrating the efficacy of the proposed systematic multi-modal approach.

1.4 Dissertation Overview

The subsequent sections of this dissertation are organised as follows: Section 2 presents a comprehensive review of the literature relevant to emotion recognition, multi-modal learning, and the EMOTIC dataset. Section 3 provides a concise overview of the work undertaken. Section 4 details the methodology employed, describing the dataset, the proposed multi-modal architecture, and the specific algorithms and techniques investigated. Section 5 reports the experimental setup, presents the evaluation results including comparative analyses derived from the staged experiments, and discusses the findings. Finally, Section 6 concludes the dissertation by summarising the key outcomes, reflecting on the contributions and limitations, and outlining potential avenues for future research.

2 Related Work

2.1 Introduction to Emotion Recognition in Context

Understanding human emotion is a cornerstone of social intelligence, enabling nuanced interaction and empathy. As artificial intelligence (AI) systems become increasingly integrated into human environments, equipping them with the capacity to perceive and interpret human emotions, a field known as Affective Computing (AC), is paramount (Mobbs et al., 2025). This capability holds transformative potential across diverse domains, including human-computer interaction (HCI), healthcare, education, robotics, and entertainment (K et al., 2024). However, computationally modelling and recognising emotion presents significant challenges due to its inherent complexity, subjectivity, and context-dependency (Shu et al., 2018). This review synthesises the literature pertinent to multi-modal emotion classification, with a specific focus on approaches utilising the EMOTIC dataset, providing a foundation for advanced research in context-aware emotion recognition.

2.1.1 Psychological Foundations

Effective computational models of emotion must be grounded in psychological theory to ensure they capture the phenomenon meaningfully (Calvo and Mac Kim, 2013). As Marsella et al. (2010) explain, the process of translating abstract psychological concepts into explicit, formal computational structures not only enables implementation but also serves to concretise these theories, potentially revealing ambiguities or underspecified elements and thereby fostering theoretical refinement. This interplay highlights a bidirectional influence: psychological frameworks guide computational design, while the rigour of implementation can feed back into and extend psychological theorising (Fathalla, 2020). Furthermore, computational models allow for the exploration of theoretical predictions, such as the temporal dynamics of emotion, in ways that complement traditional laboratory methods (Marsella et al., 2010).

Two dominant psychological frameworks underpin most computational emotion representation: categorical and dimensional models.

Categorical Models: These models posit the existence of a limited set of discrete, fundamental emotions, often considered universal and innate (Korsmit et al., 2023). The most influential is Ekman's model of six basic emotions: happiness, sadness, anger, fear, disgust, and surprise (Matsuda et al., 2013). This approach is prevalent in AC, where systems often aim to classify inputs into predefined emotional labels (Calvo and Mac Kim, 2013). While intuitive and widely adopted, categorical models face limitations in capturing the full spectrum of human affective experience, particularly subtle, mixed, or complex emotional states (Korsmit et al., 2023). Their strict boundaries may oversimplify the fluid nature of emotion.

Dimensional Models: In contrast, dimensional models represent emotions within a

continuous multi-dimensional space, rather than as distinct categories (Korsmit et al., 2023). The most common dimensions are Valence (representing the hedonic quality, from pleasant to unpleasant) and Arousal (representing the level of activation or energy, from calm to excited) (Calvo and Mac Kim, 2013). The Valence-Arousal-Dominance (VAD) model incorporates a third dimension, Dominance (representing the sense of control over the situation or oneself), which is particularly relevant in social contexts (Calvo and Mac Kim, 2013). Dimensional models offer advantages in representing emotional nuances, intensity variations, and potentially greater generalisability across languages and cultures (Calvo and Mac Kim, 2013). However, they can struggle to differentiate emotions that occupy similar positions in the dimensional space, such as anger and fear, which might both be characterised by negative valence and high arousal (Korsmit et al., 2023).

2.1.2 The Imperative of Context

Traditional research in automatic Facial Expression Recognition (FER) often relied on analysing isolated, static images of faces, frequently featuring posed, prototypical expressions under controlled laboratory conditions (Kopalidis et al., 2024). While valuable for understanding basic facial muscle configurations associated with certain emotions, this approach lacks ecological validity (Turkstra et al., 2023). In real-world interactions, faces are rarely perceived in isolation; they are embedded within a rich tapestry of contextual information (Turkstra et al., 2023).

This context is multifaceted, encompassing:

- Scene/Environment: The physical surroundings, objects present, and overall situation (Turkstra et al., 2023).
- Body Language: Posture, gestures, and whole-body movements (K et al., 2024).
- Social Situation: The presence and actions of other individuals, social dynamics, and interactions (Turkstra et al., 2023).
- Multimodal Cues: Information from other channels, such as auditory cues (voice tone, speech content) (Mobbs et al., 2025).
- Observer Factors: The perceiver's own mood, prior knowledge, cultural background, and biases (Turkstra et al., 2023).

Crucially, extensive research demonstrates that context does not merely supplement facial information but actively interacts with it, significantly influencing and sometimes fundamentally altering the perceived emotion (Turkstra et al., 2023). Identical facial expressions can be interpreted differently depending on the accompanying body posture or situational cues (Turkstra et al., 2023). In some cases, contextual information can even override strong facial signals (Turkstra et al., 2023). This highlights the limitations of relying solely on isolated facial analysis for understanding emotions as they occur naturally (Turkstra et al., 2023).

The integration of facial cues and context appears to be a fundamental aspect of human perception, operating rapidly and relatively automatically (Aviezer et al., 2011). Studies such as those by Aviezer et al. (2011) have shown that individuals cannot easily disregard contextual information even when instructed to focus solely on the face, and this integration process persists even under cognitive load, suggesting it requires minimal attentional resources. This automaticity implies that context is not simply processed separately and then deliberately combined with facial information; rather, the two streams of information are likely integrated early and interactively during perception. The strength of this integration often correlates with the congruence between the observed facial expression and the expression typically expected in that specific context (Aviezer et al., 2011).

2.1.3 Key Challenges

Despite significant progress, building robust and accurate computational emotion recognition systems remains a formidable challenge, encompassing psychological, perceptual, and technical hurdles.

Inherent Nature of Emotion:

- Subjectivity and Ambiguity: Emotion is an internal state, and its outward expression can be ambiguous or perceived differently by observers. Defining ground truth is inherently difficult (Zhang et al., 2022). Perceived emotion (annotated in datasets like EMOTIC) may not perfectly match the actual felt emotion (Shu et al., 2018).
- Cultural Variation: While some basic expressions might have universal elements (Turkstra et al., 2023), the display rules and interpretation of emotions can vary significantly across cultures (K et al., 2024).
- Intensity and Nuance: Emotions vary in intensity, and real-world experiences often involve mixed or compound emotions, which are harder to capture with simple categorical labels (Wang et al., 2024).
- Dynamics: Emotions are dynamic processes that unfold over time. Static images, as used in EMOTIC, capture only a snapshot, missing crucial temporal information (Wang et al., 2024; Richoz et al., 2018).

Technical Challenges in Vision-Based Systems:

- Environmental Variations: Performance can be severely affected by changes in illumination, background clutter, and image resolution (K et al., 2024).
- Pose and Occlusion: Variations in head pose and occlusion of facial features (e.g., by hands, objects, masks) are common in 'in-the-wild' scenarios and significantly hinder recognition, particularly for face-centric models (Kopalidis et al., 2024). Facial occlusion is noted as a specific challenge within the EMOTIC dataset (Costa et al., 2023).

- Identity Bias: Models might inadvertently learn identity-specific features rather than emotion-specific expressions, leading to poor generalisation across individuals (Kopalidis et al., 2024).
- Data Scarcity and Imbalance: Acquiring large-scale, accurately annotated emotion datasets is difficult and expensive (Feng and Chaspari, 2020). Most datasets suffer from significant class imbalance, with some emotions being much rarer than others (Audibert et al., 2024; Limami et al., 2024).

Challenges Specific to Contextual Emotion Recognition:

- Context Definition and Feature Extraction: Defining what constitutes relevant context and extracting meaningful features from complex scenes remains challenging (Yang et al., 2024).
- Modelling Interactions: Capturing the complex interplay between the person, the environment, and social dynamics (if applicable) requires sophisticated modelling techniques (Mittal et al., 2020).
- Context Bias: Models may learn spurious correlations between irrelevant background elements and specific emotion labels present in the training data, leading to poor generalisation when these correlations do not hold in new environments (Yang et al., 2024). This is a recognised issue when working with datasets like EMOTIC.
- Annotation Complexity: Annotating emotions in context, especially with a rich label set like EMOTIC's 26 categories, is highly complex and prone to inter-annotator disagreement (Costa et al., 2023).

Addressing these multifaceted challenges requires robust algorithms, representative datasets, appropriate evaluation methodologies, and a continued dialogue between computer science and psychology.

2.2 The EMOTIC Dataset: A Resource for Contextual Analysis

The EMOTIC (EMOTions In Context) dataset was introduced by Kosti et al. (2017) to specifically address the need for large-scale data enabling research into context-aware emotion recognition. Its primary goal is to facilitate the development of systems capable of inferring apparent emotional states by considering not just the individual, but also the surrounding situation depicted in static images (Kosti et al., 2017).

2.2.1 Dataset Genesis, Structure, and Annotation Schema

The dataset comprises 18,316 images containing 23,788 annotated people, sourced from real-world, non-controlled environments ('in the wild') to maximise ecological va-

lidity (Kosti et al., 2017). Some of the source images originate from existing large-scale datasets, namely MSCOCO and ADE20k (Kosti et al., 2017).

Annotation was performed using the Amazon Mechanical Turk (AMT) platform (Kosti et al., 2017). To ensure annotation quality, the creators employed a two-stage process: a qualification task to filter workers based on their ability to correctly annotate control images, and the insertion of control images during the main annotation task to monitor consistency (Kosti et al., 2017). Despite these measures, the subjective nature of emotion and the complexity of the annotation task mean that label noise remains a potential issue, as discussed further in Section 2.2.2. Based on the guidance given by Kosti et al. (2017), the training split (70% of data) was annotated by a single annotator per image, while the validation (10%) and test (20%) sets received annotations from 5 and 3 annotators respectively, allowing for some measure of agreement analysis but also potentially introducing systematic differences between splits. Examples from the dataset are shown in Figure 1.

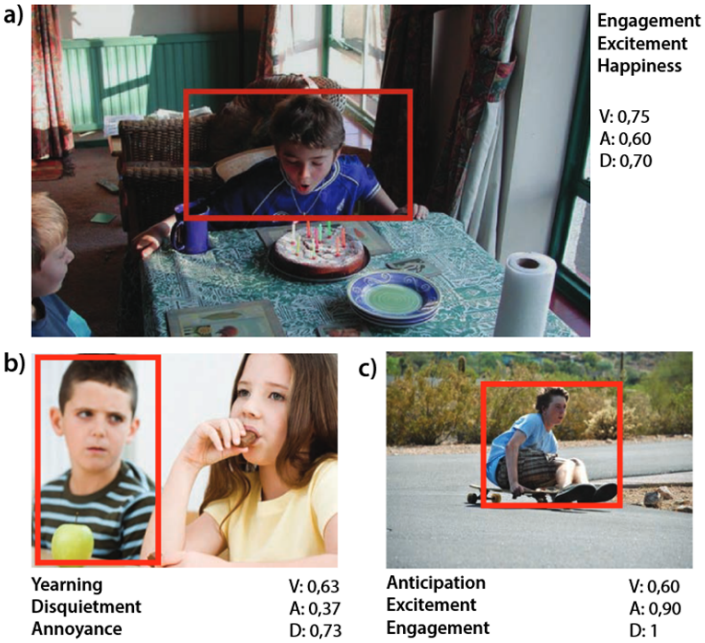

Figure 1: Illustrative examples from the EMOTIC dataset (Kosti et al., 2017). Bounding boxes localise individuals within contextual scenes (a, b, c), accompanied by their corresponding multi-label discrete emotion categories and continuous VAD annotations.

2.2.2 Critical Appraisal: Strengths and Limitations

The EMOTIC dataset represents a significant contribution to affective computing, but like any large-scale, real-world dataset, it possesses both strengths and limitations that researchers must consider.

Strengths:

- Contextual Focus: It was one of the first large datasets specifically designed for

emotion recognition in context, driving research in this crucial area (Kosti et al., 2017).

- Ecological Validity: The use of 'in-the-wild' images depicting natural situations provides higher ecological validity compared to datasets using posed expressions in controlled settings (Kosti et al., 2017). It presents challenges closer to real-world application scenarios.
- Rich Annotations: The dual representation system (26 discrete categories + VAD dimensions) offers flexibility and allows for the investigation of fine-grained emotional states beyond basic emotions (Kosti et al., 2017).
- Person Localisation: Providing bounding boxes facilitates methods that explicitly process person-specific features alongside global context (Kosti et al., 2019).
- Context Diversity: Compared to some other contextual datasets (e.g., CAER), EMOTIC has been noted for its more diverse range of contextual representations (Costa et al., 2023).

Limitations:

- Annotation Noise and Inconsistency: Review of relevant and current literature (Costa et al., 2023; Etesam et al., 2024) suggests this being the most significant limitation. The subjective nature of emotion perception, combined with the complexity of 26 categories and the use of crowdsourcing (AMT), inevitably leads to noise and disagreement among annotators (Costa et al., 2023). Analysis has revealed conflicting annotations for the same person (e.g., simultaneous annotation of 'Happiness' and 'Pain', or 'Fear' and 'Confidence') (Costa et al., 2023). The difference in the number of annotators between training (1) and validation/test sets (5/3) could also introduce biases or affect the reliability of evaluation (Etesam et al., 2024). The richness of the annotation scheme, while a strength, simultaneously contributes to this challenge; annotating 26 subtle categories reliably is difficult, especially for non-expert annotators. This necessitates models robust to label noise and careful interpretation of evaluation metrics.
- Class Imbalance: The dataset suffers from severe class imbalance across the 26 discrete categories, as illustrated in Figure 2, with some emotions appearing far more frequently than others (Audibert et al., 2024). This poses a significant challenge for training models that perform well on minority classes. The exclusion of a 'Neutral' category might further complicate learning, potentially forcing annotators (and thus models) to assign emotional labels even to neutral expressions, possibly based on perceived traits (Costa et al., 2023).
- Occlusion and Scale: Being 'in-the-wild', the images in EMOTIC often contain faces that are small, partially occluded, or not clearly visible, making facial expression analysis difficult or impossible in many instances (Costa et al., 2023). This reinforces the need for context and body cues but is a limitation for face-dependent methods.

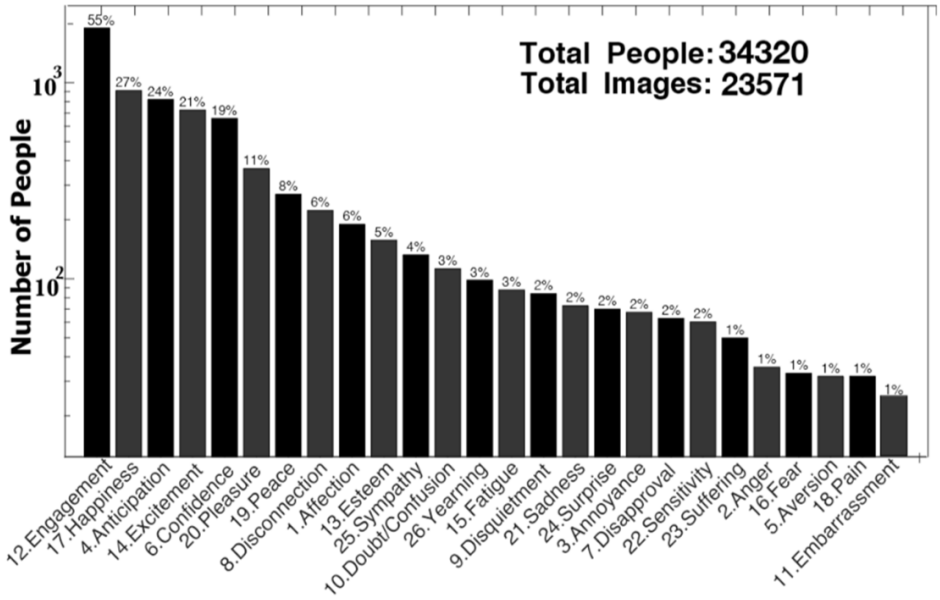

Figure 2: Class distribution for the 26 discrete emotion categories in the EMOTIC dataset, illustrating the significant imbalance between frequent and rare emotions.

- Static Nature: The dataset contains only static images, lacking the temporal information inherent in real-world emotional expression and perception (Wang et al., 2024). Dynamic cues can significantly aid recognition (Richoz et al., 2018).
- Potential Biases: As images sourced from MSCOCO and ADE20k (Kosti et al., 2017) and the web, they may contain inherent demographic biases (e.g., ethnicity, gender, age distribution) or cultural biases in expression (Costa et al., 2023). Furthermore, context bias, where models learn spurious correlations between specific scene types and emotions (e.g., beaches always correlating with happiness), is a documented problem (Yang et al., 2024).
- Reproducibility: Some influential works using EMOTIC have not made their code publicly available (Kosti et al., 2017), hindering direct comparison and reproduction of results (Etesam et al., 2024).

2.2.3 Benchmarking Emotion Recognition on EMOTIC

The task benchmarked on EMOTIC within this dissertation is multi-label classification over the 26 discrete emotion categories. Performance is typically measured using mean Average Precision (mAP), which accounts for the ranking of predicted labels and is suitable for multi-label problems (Etesam et al., 2024). Some studies also report performance on predicting the continuous VAD dimensions (Kosti et al., 2017).

The original work by Kosti et al. (2017) established baseline performance using a two-branch CNN architecture (Kosti et al., 2019). One branch processed the cropped person region (body features), and the other processed the entire image (context features), which were then fused. Depending on the specific fusion strategy and loss function, their reported mAP on the test set was approximately 27% (Kosti et al., 2017).

Subsequent research has explored various architectural and methodological improvements, leading to gradual increases in benchmark performance. Key developments, summarised in Table 1, include:

- **High-Level Context**: de Lima Costa et al. (2023) focused on improving context representation, achieving an mAP of 30.02%.
- **Knowledge-Enhanced Methods (DRM/TEKG)**: Chen et al. (2023) incorporated structured emotion commonsense knowledge and interpersonal relations using their framework, reporting mAPs of 26.48% (DRM configuration) and 31.36% (TEKG configuration). This approach explicitly models relationships and leverages external knowledge graphs to enhance context understanding.
- **EmotiCon (Depth)**: Mittal et al. (2020) proposed a model integrating multiple context types (modalities, semantics, social interactions modelled using depth maps), significantly improving performance to 35.5%.
- **Vision-Language Models (VLMs)**: Approaches leveraging large pre-trained VLMs have shown promise. Etesam et al. (2024) demonstrated that linear probing on CLIP features (EmotionCLIP) achieved an mAP of 32.9%.

Table 1: Summary of Key EMOTIC Benchmark Results (Discrete Categories, mAP)

Method Name	Key Technique(s)/Streams	Backbone(s)	Reported mAP (%)
EMOTIC Baseline (Kosti et al., 2017)	Two-branch CNN (Body, Context)	CNN (Low-Rank)	27.38
DRM (Chen et al., 2023)	Structured Knowledge & Relations (Config 1)	ResNet-50	26.48
High-Level Context (de Lima Costa et al., 2023)	Enhanced Context Representation	ResNet-50	30.02
TEKG (Chen et al., 2023)	Structured Knowledge & Relations (Config 2)	ResNet-50	31.36
EmotionCLIP (Etesam et al., 2024)	CLIP Linear Probe	CLIP (ViT variant)	32.91
EmotiCon (Depth) (Mittal et al., 2020)	Multi-Context, Depth Maps	CNN	35.48

Analysis of the benchmark results in Table 1 indicates a clear progression. Initial methods like the EMOTIC baseline focused on fusing separate body and context streams using standard CNNs. Subsequent work explored enhanced context representations (de Lima Costa et al., 2023). Incorporating external knowledge, such as structured emotion commonsense and interpersonal relations (DRM/TEKG), provided further gains (Chen et al., 2023). A significant jump in performance was achieved by Mittal et al. (2020) with EmotiCon, particularly the depth-based variant, which explicitly modelled multiple context types including social dynamics. More recently, the trend has shifted towards leveraging large pre-trained models like CLIP (Etesam et al., 2024), suggesting that robust contextual emotion recognition benefits significantly from higher-level semantic understanding and reasoning capabilities beyond simple feature aggregation.

2.3 Deep Learning Paradigms for Visual Emotion Recognition

Deep learning (DL) has become the predominant approach for tackling visual emotion recognition tasks, largely supplanting traditional machine learning methods that relied on handcrafted features (K et al., 2024). The ability of deep neural networks to

automatically learn hierarchical feature representations directly from data has proven highly effective for complex perceptual tasks like interpreting facial expressions and contextual scenes (Shen, 2024; Pereira et al., 2024).

2.3.1 Foundational Deep Learning Approaches: CNNs and Transformers

Two main families of deep learning architectures dominate the field of computer vision, including emotion recognition: Convolutional Neural Networks (CNNs) and Transformers.

Convolutional Neural Networks (CNNs): CNNs have long been the standard for image analysis tasks (Aleissaee et al., 2023). Their architecture, built around convolutional layers, pooling layers, and activation functions (like ReLU), is inherently suited to processing grid-like data such as images (Shen, 2024). Convolutional filters act as local feature detectors, learning hierarchical representations from simple edges and textures in early layers to more complex object parts and shapes in deeper layers (Elharrouss et al., 2024). Key properties like parameter sharing and translation equivariance make CNNs relatively data-efficient and effective at capturing spatial hierarchies (Aleissaee et al., 2023). They have been widely applied to FER and CAER, often forming the backbone for feature extraction (K et al., 2024; Chen et al., 2021).

Transformers: Originally developed for natural language processing (NLP) tasks (Islam et al., 2024), Transformer architectures were adapted for computer vision with the introduction of the Vision Transformer (ViT) (Shen, 2024). Instead of convolutions, Transformers rely on the self-attention mechanism (Islam et al., 2024). Self-attention allows the model to weigh the importance of different image patches relative to each other, enabling it to capture long-range dependencies and model global context effectively (Shen, 2024). This global receptive field contrasts with the inherently local receptive field of convolutional layers. Transformers have shown strong performance, particularly on large datasets, and as demonstrated by Min et al. (2024), are increasingly applied to emotion recognition, offering advantages in modelling global scene context or holistic body posture (Tarekegn et al., 2021).

Both CNNs and Transformers have distinct advantages and disadvantages for visual emotion recognition (Shen, 2024; Takahashi et al., 2024). CNNs leverage strong spatial inductive biases (locality, translation equivariance), making them generally more sample-efficient and adept at learning local patterns like facial features (Aleissaee et al., 2023). However, their limited receptive field in early layers can make capturing global context or long-range spatial relationships less direct. Transformers excel at modelling global relationships via self-attention but typically lack these built-in spatial biases (Khan et al., 2023). This makes them more powerful when sufficient data is available but often requires large-scale pre-training to learn basic visual properties that CNNs learn more readily (Khan et al., 2023).

2.3.2 Comparative Review of Feature Extraction Backbones

In many deep learning systems for emotion recognition, a pre-trained network, referred to as the 'backbone', is used to extract initial visual features from the input image or specific regions (like the person bounding box) (Elharrouss et al., 2024). These features are then fed into subsequent layers or modules for task-specific processing, such as fusion and classification. The choice of backbone significantly impacts performance and efficiency.

Convolutional Architectures

- **VGG (e.g., VGG-16, VGG-19)**: Developed by the Visual Geometry Group at Oxford, VGGNets are characterised by their simplicity, using stacked 3x3 convolutional layers followed by max-pooling (Chen et al., 2021). Increasing depth (16 or 19 layers) led to improved performance on ImageNet (Cheng and Zhou, 2020). VGG has served as a backbone in early deep learning approaches for emotion recognition (Terven et al., 2023) and remains a common baseline (Akhand et al., 2021). Its main drawback is the large number of parameters and computational cost (Elharrouss et al., 2024).
- **ResNet (e.g., ResNet-50, ResNet-101)**: Residual Networks introduced skip connections (residual blocks) that allow gradients to flow more easily through very deep networks, overcoming the vanishing gradient problem encountered by earlier deep models like VGG (Elharrouss et al., 2024). This enabled the training of networks with 50, 101, 152, or even more layers (Elharrouss et al., 2024). ResNets are highly influential and widely used as strong backbones for various vision tasks (Terven et al., 2023), including emotion recognition, often outperforming shallower CNNs and VGG (Terven et al., 2023). Variants like ResNeXt introduce grouped convolutions (cardinality) for further improvements (Elharrouss et al., 2024).
- **EfficientNet (e.g., B0-B7)**: EfficientNets were designed using neural architecture search and a novel compound scaling method (Terven et al., 2023). As described by Liu et al. (2022), this method systematically scales network depth, width, and input resolution together to achieve a better balance between accuracy and computational efficiency (measured in FLOPs) compared to previous models. EfficientNets offer a family of models (B0 to B7) with increasing capacity and performance, providing good options when computational resources are a constraint (Alayón et al., 2023). Studies comparing backbones have shown EfficientNet to be competitive (Alayón et al., 2023).
- **MobileNet (e.g., MobileNetV2, MobileNetV3)**: MobileNets are specifically designed for efficiency on mobile and embedded devices (Terven et al., 2023). They achieve this primarily using depthwise separable convolutions, which significantly reduce the number of parameters and computations compared to standard convolutions (Terven et al., 2023). MobileNetV2 introduced inverted residuals and

linear bottlenecks, while MobileNetV3 incorporated architecture search and optimised activation functions (h-swish) (Terven et al., 2023). While highly efficient, they typically trade some accuracy for this reduction in complexity (Urnisha et al., 2024). Their suitability for emotion recognition depends on the required accuracy versus computational budget (Urnisha et al., 2024; Franzoni et al., 2020).

Transformer Architectures

- **ViT (Vision Transformer)**: The original ViT demonstrated that a pure Transformer architecture could achieve state-of-the-art results on image classification, challenging CNN dominance (Tarekegn et al., 2021). It processes images by dividing them into fixed-size patches, linearly embedding them, adding position embeddings, and feeding the sequence to a standard Transformer encoder (Khan et al., 2023). Its main limitations are the lack of strong image-specific inductive biases, leading to a heavy reliance on large-scale pre-training (e.g., JFT-300M) for optimal performance, and quadratic complexity with respect to the number of patches (Khan et al., 2023).
- **Swin Transformer**: Swin Transformers introduced a hierarchical structure and shifted window self-attention (Liu et al., 2022). By computing self-attention within local windows and shifting these windows across layers to allow cross-window connections, Swin achieves linear complexity relative to image size and incorporates a degree of locality bias, making it more suitable as a general-purpose vision backbone (Khan et al., 2023). As shown by Kim et al. (2023), Swin often matches or surpasses ViT and ResNet performance on various tasks, including detection and segmentation, often with better efficiency or less reliance on massive pre-training datasets (Khan et al., 2023).
- **DeiT (Data-efficient image Transformers)**: DeiT models specifically addressed the data-hungriness of ViTs (Alayón et al., 2023). Using knowledge distillation from a pre-trained CNN teacher and other training refinements, DeiT achieves competitive performance on ImageNet using only ImageNet training data, making Transformers more accessible when large proprietary datasets like JFT-300M are unavailable (Alayón et al., 2023).

2.3.3 The Role of Transfer Learning (ImageNet Pre-training)

Transfer learning (TL) is a cornerstone technique in modern computer vision, particularly when dealing with tasks where labelled data is limited, such as often is the case for specific emotion datasets (Feng and Chaspari, 2020). The standard practice involves taking a backbone network (CNN or Transformer) pre-trained on a large-scale dataset, typically ImageNet (containing over a million images across 1000 object categories, as described by Akhand et al. (2021)), removing its original classification head, and fine-tuning the network (either fully or partially) on the target dataset (e.g., an emotion dataset like EMOTIC or FER2013) (Urnisha et al., 2024).

The effectiveness of this approach stems from the assumption that features learned on the large source dataset (e.g., edge detectors, texture analysers, part detectors learned from ImageNet) are general enough to be useful for the target task (Akhand et al., 2021). Pre-training provides a powerful initialisation, allowing the model to converge faster and achieve better performance on the target task compared to training from scratch, especially with limited target data (Feng and Chaspari, 2020). Numerous studies, including those by Urnisha et al. (2024) and Akhand et al. (2021), have demonstrated significant accuracy improvements in emotion recognition tasks by leveraging ImageNet pre-trained models. For instance, models pre-trained on ImageNet have shown strong performance even when fine-tuned on relatively small or specialised emotion datasets (Akhand et al., 2021).

2.4 Strategies for Multi-Modal Feature Fusion

Recognising emotions accurately in real-world scenarios, as depicted in datasets like EMOTIC, necessitates integrating information from multiple sources or modalities (K et al., 2024). Relying solely on facial expressions is often insufficient due to ambiguity, occlusion, or the subtle nature of the emotion (Turkstra et al., 2023). Body posture, gestures, and the surrounding visual context provide crucial complementary information that can significantly enhance understanding (K et al., 2024).

2.4.1 Rationale for Fusing Context, Body, and Face Cues

The core motivation for multi-modal fusion in this domain, as outlined by Li and Tang (2024), is that different cues offer distinct but related perspectives on the emotional state. The face provides fine-grained expressive details (when visible), the body conveys overall posture and energy level, and the context offers situational information that helps disambiguate expressions or infer likely affective states (K et al., 2024). For example, a smiling face might indicate happiness in a social gathering but schadenfreude or nervousness in a different context.

The goal of fusion is to combine these disparate information streams into a unified representation that is richer and more discriminative than any single modality alone (Jiao et al., 2024). However, effective fusion faces several challenges (Guo et al., 2023):

- Heterogeneity: Different modalities have different data structures and statistical properties (e.g., structured features from a face detector vs. global scene features) (Jiao et al., 2024).
- Alignment: Features from different modalities may not be naturally aligned spatially or temporally (though less critical for static images like EMOTIC) (Jiao et al., 2024).
- Noise and Reliability: Different modalities can have varying levels of noise or reliability depending on the situation (e.g., face occluded, context ambiguous)

(Jiao et al., 2024). A robust fusion mechanism should ideally adapt to these variations.

- Information Complementarity vs. Redundancy: Fusion should leverage complementary information while handling redundancy effectively (Jiao et al., 2024).

2.4.2 A Taxonomy and Critical Analysis of Fusion Techniques

Fusion techniques vary by stage and mechanism, as surveyed by Gao et al. (2020) and Pawłowski et al. (2023):

- Early vs. Late Fusion: Combining at the feature level allows learning of low-level interactions but is sensitive to missing data, whereas decision-level fusion is simpler and more robust to missing data but misses low-level interactions (Pawłowski et al., 2023). Hybrid or intermediate fusion approaches are common (Pawłowski et al., 2023).
- Simple Methods: Concatenation simply joins feature vectors, increasing dimensionality (Gao et al., 2020). Element-wise operations like sum, average, or product require aligned dimensions and risk information loss (Gao et al., 2020; Guo et al., 2023).
- Weighted/Gated Methods: These methods use learnable weights (Pawłowski et al., 2023) or gates, such as the Gated Multimodal Units (GMU) proposed by Arevalo et al. (2017), to adaptively control modality contributions. This offers more flexibility than static methods but increases complexity.
- Attention Mechanisms: Self-attention (within modality) and cross-modal attention (between modalities) dynamically focus on relevant information, explicitly modelling inter-modal relationships (Lu et al., 2023; Guo et al., 2023; Praveen and Alam, 2024).
- Transformer-Based Fusion: Utilising Transformer layers (incorporating self- and cross-attention) allows modelling of complex dependencies between modality tokens (Gao et al., 2020; Wang et al., 2023). These methods are powerful but computationally intensive (Nagrani et al., 2021).
- Bilinear Pooling: Techniques like bilinear pooling aim to capture second-order interactions between features (e.g., through an outer product) (Winterbottom et al., 2022). Standard bilinear pooling suffers from high dimensionality; compact variants like Multimodal Compact Bilinear pooling (MCB) or Multimodal Low-rank Bilinear pooling (MLB) offer efficient approximations (Winterbottom et al., 2022), with low-rank approaches explored by Chu et al. (2021).
- Advanced Strategies: Other approaches include hierarchical fusion (integrating information at multiple levels) (Zeng et al., 2023), ensemble methods (combining predictions from diverse models) (Peng et al., 2015), applying layer normalisation before fusion to stabilise training (Tarekegn et al., 2021), and bottleneck fusion (compressing features before combination) (Nagrani et al., 2021).

The trend in multimodal fusion research is towards dynamic, adaptive methods like attention, gating, and transformer-based approaches, which can better handle varying cue reliability and model complex inter-modal dependencies (Chu et al., 2021). This justifies their exploration for challenging tasks like emotion recognition on the EMOTIC dataset.

2.5 Mitigating Data Imbalance in Emotion Recognition

As previously noted, class imbalance is a pervasive issue in emotion recognition datasets, including EMOTIC (Limami et al., 2024). This imbalance, where certain emotions (majority classes) are far more prevalent than others (minority classes), can significantly skew model training, leading to models that perform well on common emotions but poorly on rarer ones, thus limiting their real-world utility (Tarekegn et al., 2021). Several strategies can be employed to mitigate this problem, broadly categorised as data-level, algorithm-level (discussed previously under loss functions in Section 4.5), and feature-level approaches.

2.5.1 The Challenge of Imbalanced Emotion Datasets

The reasons for imbalance in emotion datasets are manifold: some emotions are intrinsically less frequent in daily life, data collection methods might introduce sampling biases, and annotation can be more difficult or ambiguous for certain subtle or complex emotions. Regardless of the cause, the consequence is that standard machine learning algorithms, which often implicitly assume balanced class distributions, tend to develop a bias towards the majority classes, as these contribute more significantly to the overall loss or error signal during training (Tarekegn et al., 2021). This results in poor recall and F1-scores for minority classes, even if overall accuracy appears high.

2.5.2 Image-Level Augmentation Techniques

Data augmentation artificially increases the size and diversity of the training dataset by applying various transformations to existing images (Shantharam and Schwenker, 2024). This helps improve model robustness, reduce overfitting, and can indirectly help with imbalance by creating more varied examples of minority classes. Common techniques involve geometric transformations (e.g., random flips, rotations, scaling, cropping, translation) and photometric transformations (e.g., adjusting brightness, contrast, saturation, adding noise) (Shantharam and Schwenker, 2024). These techniques are further detailed in Section 4.6.1.

2.6 Chapter Summary

This section provided a review of the literature relevant to context-aware emotion recognition. It began by establishing the psychological foundations, contrasting categorical and dimensional models of emotion, and highlighting the critical role of context in interpreting facial expressions, moving beyond traditional FER approaches. Key challenges inherent in emotion recognition, particularly in contextual settings and using vision-based systems, were discussed, including subjectivity, ambiguity, environmental variations, data imbalance, and annotation complexity. The EMOTIC dataset was introduced as a primary resource for this task, detailing its structure, annotation schema, strengths (ecological validity, rich annotations), and significant limitations (annotation noise, class imbalance, static nature). A summary of benchmark results on EMOTIC, presented in Table 1, illustrated the evolution of approaches, from early CNN fusion methods to more recent knowledge-enhanced and VLM-based techniques achieving higher performance. The review then covered dominant deep learning paradigms (CNNs and Transformers), comparing representative backbone architectures (VGG, ResNet, EfficientNet, MobileNet, ViT, Swin, DeiT) and the importance of transfer learning. Finally, strategies for multi-modal feature fusion were taxonomised and critically analysed, alongside a discussion of the pervasive challenge of data imbalance in emotion datasets and the role of image-level augmentation as a mitigation strategy. This review establishes the background and motivates the methodological choices explored in subsequent sections.

3 Description of Work

Following the problem identification and literature review presented in Sections 1 and 2, respectively, this dissertation details the development and evaluation of a multi-modal deep learning framework for context-aware, multi-label emotion classification using the EMOTIC dataset (Kosti et al., 2017). The core technical work, described in detail in Section 4 (Methodology) and Section 5 (Experiments, Results and Discussion), focuses on constructing, systematically testing, and refining this framework to address the specific challenges posed by the dataset and the task.

The methodology leverages the 'in-the-wild' nature of the EMOTIC dataset, employing a multi-modal approach that processes three distinct visual streams corresponding to the context (overall scene), the body (person bounding box), and the face (cropped facial region). Feature extraction for each stream utilises various pre-trained deep learning backbones, including both Convolutional Neural Networks (EfficientNet-B7, MobileNetV3) and Vision Transformers (Swin Transformer, DeiT), capitalising on transfer learning from ImageNet.

A significant aspect of the work involves a systematic exploration and comparison of diverse feature fusion techniques, ranging from simple concatenation to more complex methods like weighted fusion, attention mechanisms, cross-modal transformers, and bilinear pooling, to effectively integrate information from the three modalities.

To address the dataset's multi-label nature (26 discrete emotion categories) and severe class imbalance, specialised multi-label loss functions (such as Binary Cross-Entropy, a custom weighted Discrete Loss, Focal Loss, Asymmetric Loss, and Dice Loss) were implemented and evaluated. Furthermore, the impact of various data augmentation strategies was investigated. This included image-level augmentations (e.g., flips, colour jitter, affine transformations, noise, mixing techniques) applied during training, and dataset-level resampling using ADASYN applied in the feature space to synthetically balance class distributions.

A systematic, multi-stage experimental procedure, detailed in Section 5, was employed to compare these different components (backbones, fusion methods, loss functions, augmentation levels). Each stage built upon the best performing configuration from the previous one, allowing for methodical optimisation and analysis of each component's contribution. This process culminated in identifying an optimal framework configuration.

The implementation relied on Python and the PyTorch deep learning library, utilising supporting libraries like TIMM, NumPy, Scikit-learn, and Matplotlib. Experiments were conducted adhering to the standard EMOTIC data splits and evaluation protocols, primarily using mean Average Precision (mAP) as the key performance metric. The final performance of the optimised framework was rigorously evaluated on the held-out test set, with results and analysis presented in Section 5. Finally, Section 6 provides a concluding discussion, summarising findings, contributions, limitations, and potential directions for future research.

4 Methodology

This section details the methodology employed to develop and evaluate the multi-modal deep learning framework for context-aware, multi-label emotion classification on the EMOTIC dataset. It describes data preprocessing steps, the proposed system architecture, the specific techniques investigated for feature extraction, feature fusion, loss computation, data augmentation, and the evaluation metrics used.

4.1 Data Preprocessing and Input Streams

The raw data for this project consists of pre-extracted image arrays corresponding to the context, body, and face regions for each annotated person in the EMOTIC dataset, alongside their categorical labels. The preprocessing pipeline involves several steps:

1. **Loading Data:** The pre-computed NumPy arrays are loaded. The face array, originally single-channel grayscale, is stacked to create a 3-channel image compatible with standard pre-trained models expecting RGB input.
2. **Input Stream Generation:** For each annotated person instance, three distinct input visual streams are prepared, as illustrated in Figure 3:
    - Context Stream: The entire image containing the person is used as the context input.
    - Body Stream: An image cropped around the annotated person bounding box serves as the body input.
    - Face Stream: A cropped region focusing specifically on the face area is used as the face input.
3. **Image Transformations:** Before being fed into the feature extraction backbones, each image stream undergoes a series of transformations defined using ‘torchvision.transforms‘:
    - Conversion to PIL Image format.
    - Data Augmentation (Training Only): Various image augmentation techniques are applied during training, as detailed in Section 4.6.1.
    - Conversion to PyTorch Tensors.
    - Normalisation: Pixel values are normalised using mean and standard deviation statistics appropriate for the specific backbone model being used (typically ImageNet statistics, unless specified otherwise). Specific normalisation parameters are applied per stream based on pre-calculated dataset statistics or standard ImageNet values.

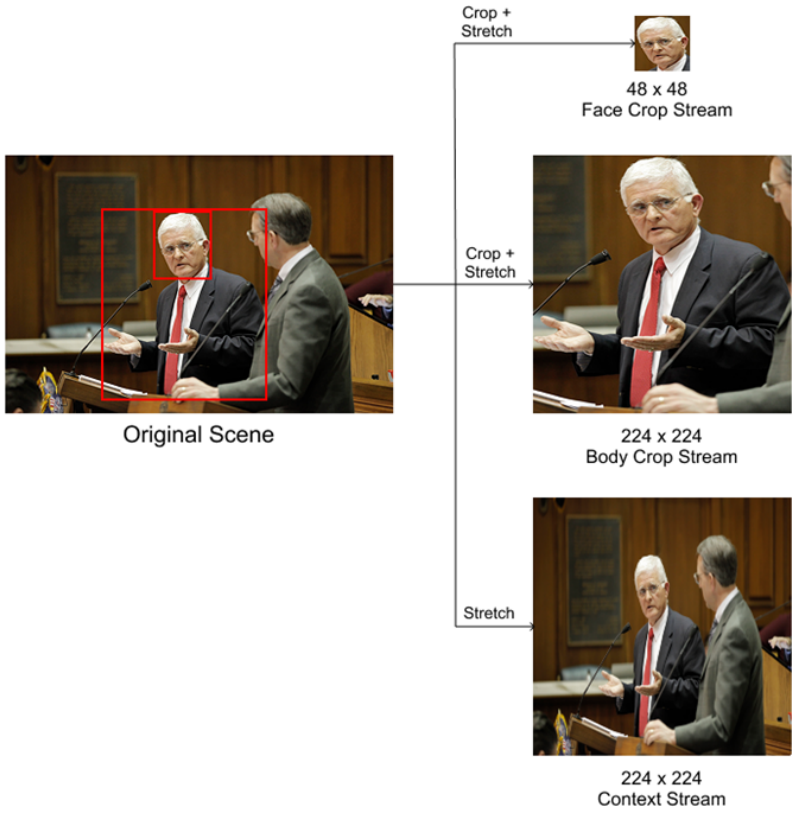

Figure 3: Input Stream Generation pipeline, showing the extraction of Context, Body, and Face streams from the original scene.

4.2 Proposed Multi-Modal Framework

The core of the proposed methodology is a multi-modal deep learning framework, specifically designed to address the challenges of context-aware emotion recognition by integrating information from multiple visual streams derived from the EMOTIC dataset: the overall scene context, the subject's body language, and their facial expression. This multi-modal approach is motivated by the limitations of relying solely on facial expressions, which can be ambiguous or occluded in real-world 'in-the-wild' scenarios, and the established importance of context and body cues in human emotion perception (Turkstra et al., 2023; Aviezer et al., 2011). By processing these streams in parallel, the framework aims to capture a more holistic representation of the affective state.

The overall architecture follows a structured, parallel processing pipeline, illustrated conceptually in Figure 4:

- **Input:** The process begins with the pre-processed context, body, and face images corresponding to a single annotated person instance within an EMOTIC image (as described in Section 4.1).

- **Feature Extraction:** Each of these three input streams is independently fed into a dedicated deep learning backbone network (detailed in Section 4.3). These networks, typically pre-trained on large-scale datasets like ImageNet (Akhand et al., 2021), are responsible for learning hierarchical visual features and extracting high-level semantic representations relevant to each specific stream (scene understanding for context, posture/action for body, expression details for face).
- **Feature Fusion:** The high-level feature vectors extracted from the three parallel backbones are then combined using a dedicated feature fusion module (detailed in Section 4.4). Recognising that simple concatenation may not optimally capture inter-modal dependencies, various fusion strategies are investigated to determine the most effective method for integrating the complementary and potentially redundant information.
- **Classification:** The resulting fused feature vector, which encapsulates information from all three modalities, is passed through one or more fully connected layers that constitute the classification head. This head maps the integrated features to the final prediction space.
- **Output:** The framework outputs a vector of scores or probabilities (via a Sigmoid activation) for each of the 26 discrete emotion categories defined in the EMOTIC dataset, enabling multi-label prediction (detailed in Section 4.5).

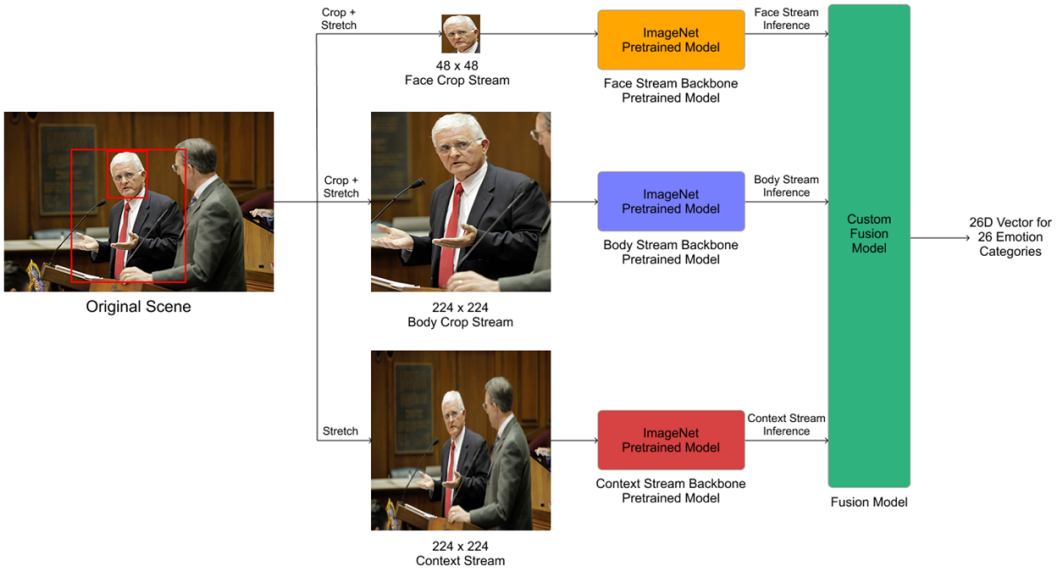

Figure 4: Architecture of the proposed Multi-Modal Framework for emotion prediction, illustrating the parallel processing of context, body, and face streams, followed by feature fusion and classification.

4.3 Feature Extraction Backbones

A crucial component influencing the framework's performance is the choice of backbone networks used for feature extraction within each of the three streams. Leveraging the power of transfer learning (Feng and Chaspari, 2020), this research primarily considered architectures pre-trained on the large-scale ImageNet dataset. This approach allows the models to benefit from general visual features learned on millions of images, significantly reducing training time and data requirements compared to training from scratch, which is particularly advantageous given the moderate size and specific nature of emotion datasets. The selection of backbones for experimental evaluation aimed to cover diverse and representative architectures from both the CNN and Vision Transformer families, as reviewed in Section 2:

Convolutional Neural Networks (CNNs): These architectures excel at learning spatial hierarchies through convolutional filters.

- EfficientNet-B7: Chosen as a representative high-performance CNN, EfficientNet models achieve a strong balance between accuracy and computational efficiency by systematically scaling network depth, width, and resolution using a compound scaling method (Liu et al., 2022). The B7 variant represents the higher end of this family, selected to explore the potential for high accuracy.
- MobileNetV3: Included to evaluate a highly efficient architecture designed for resource-constrained environments (Terven et al., 2023). Its use of depthwise separable convolutions significantly reduces computational cost, making it relevant for potential real-time applications or as a comparison point regarding the

trade-off between efficiency and performance in the context of feature extraction for emotion.

Vision Transformers (ViTs): These architectures utilise self-attention mechanisms to capture global dependencies within the input.

- Swin Transformer: Selected for its hierarchical structure and use of shifted window attention, which provides linear computational complexity with respect to image size and incorporates some spatial locality bias, making it a strong general-purpose vision backbone often outperforming earlier ViTs (Liu et al., 2022). Its effectiveness across various vision tasks motivated its inclusion.
- DeiT (Data-efficient image Transformer): Included as a representative ViT variant specifically designed to achieve competitive performance using only ImageNet pre-training, without requiring massive proprietary datasets (Alayón et al., 2023). This makes it a practical choice for leveraging the Transformer architecture and assessing its capabilities relative to CNNs under standard pre-training conditions.

For each stream (context, body, face), different combinations of these backbones were tested experimentally (detailed in Section 5) to identify the most effective setup for capturing relevant features for that specific modality. To prepare the features for fusion, the final classification layer (head) of the pre-trained backbone is typically removed or replaced with an identity layer, allowing the network to output a high-dimensional feature vector representing the learned visual information.

4.4 Feature Fusion Strategies

Combining the features extracted from the context, body, and face streams is critical for leveraging the complementary information each provides. This research systematically investigated a wide array of feature fusion techniques, ranging from simple baselines to more complex attention-based methods. The evaluated strategies, illustrated in Figures 5 through 14, include:

1. **Simple Concatenation:** As shown in Figure 5, this baseline approach flattens the input feature tensors from each stream (Context C, Body B, Face F). These flattened vectors are then directly concatenated along the feature dimension. The resulting single, high-dimensional vector undergoes further processing through a sequence comprising a Linear layer, BatchNorm1d, ReLU activation, and Dropout, before feeding into the final classification head. This method relies entirely on these subsequent layers to learn inter-modal interactions.

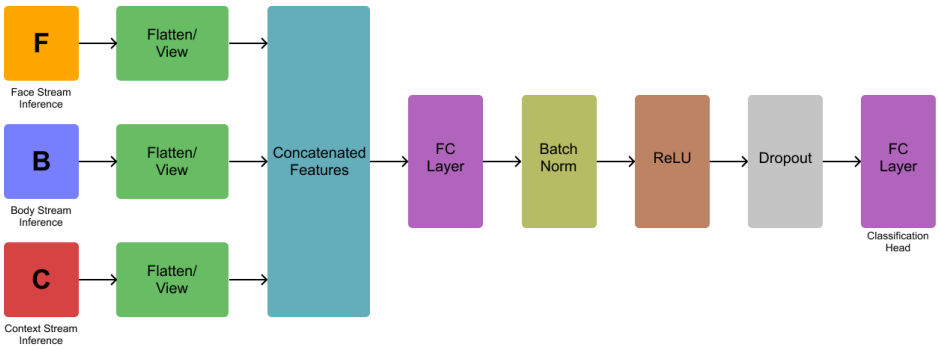

Figure 5: Simple Concatenation Fusion Model Architecture.

2. **Weighted Fusion:** Illustrated in Figure 6, this method introduces learnable parameters to prioritise modalities (Pawłowski et al., 2023). Flattened input features (C, B, F) are first projected into a common 256-dimension space using separate Linear layers followed by ReLU activation. Three distinct learnable scalar weights (W_C, W_B, W_F) are then element-wise multiplied with their respective projected feature vectors. The final fused representation is obtained by summing these weighted vectors and explicitly normalising the result by the sum of the three weights (plus a small epsilon for numerical stability). This adaptively weighted vector then passes through BatchNorm1d, Dropout, and the classification head.

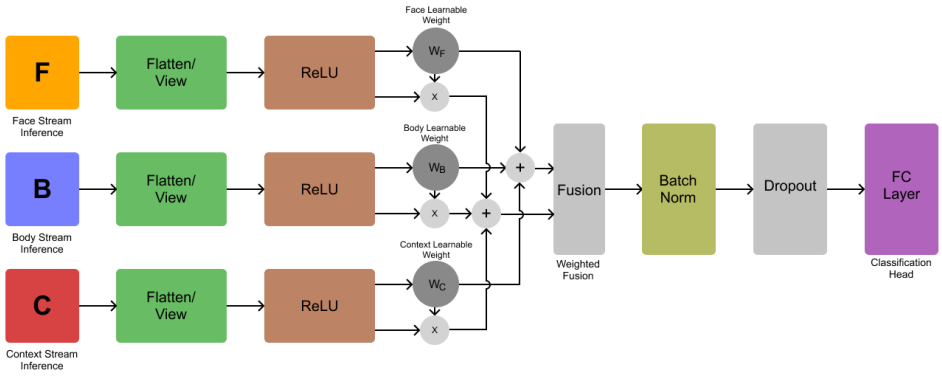

Figure 6: Weighted Fusion Model Architecture.

3. **Attention-Based Weighted Fusion:** Depicted in Figure 7, this approach computes data-dependent weights using an attention mechanism. Features (C, B, F) are flattened and concatenated. This concatenated vector is then fed into a separate small neural network (Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ Softmax), which outputs a vector of attention weights having the same dimension as the concatenated features. These attention weights are applied to the original concatenated features via element-wise multiplication. The resulting attention-weighted feature vector is then processed through subsequent shared layers (Linear, BatchNorm1d, ReLU, Dropout) before classification.

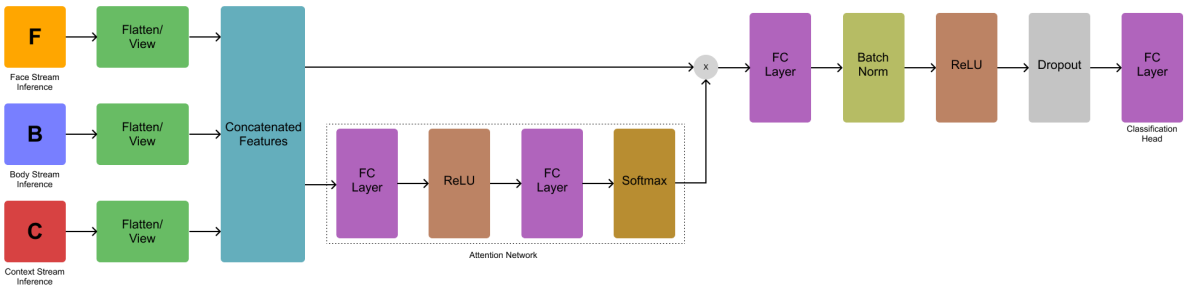

Figure 7: Attention-Based Weighted Fusion Model Architecture.

4. **Transformation Fusion:** Shown in Figure 8, unlike simple concatenation, this method applies separate transformations to each modality before combining them. Flattened features from C, B, and F are each passed through their own transformation module (Context Transform, Body Transform, Face Transform), consisting of a Linear layer projecting to 128 dimensions followed by a ReLU activation. The three resulting transformed vectors are then concatenated side-by-side. This combined vector is subsequently processed by shared layers (Linear, BatchNorm1d, ReLU, Dropout) leading to the classification head.

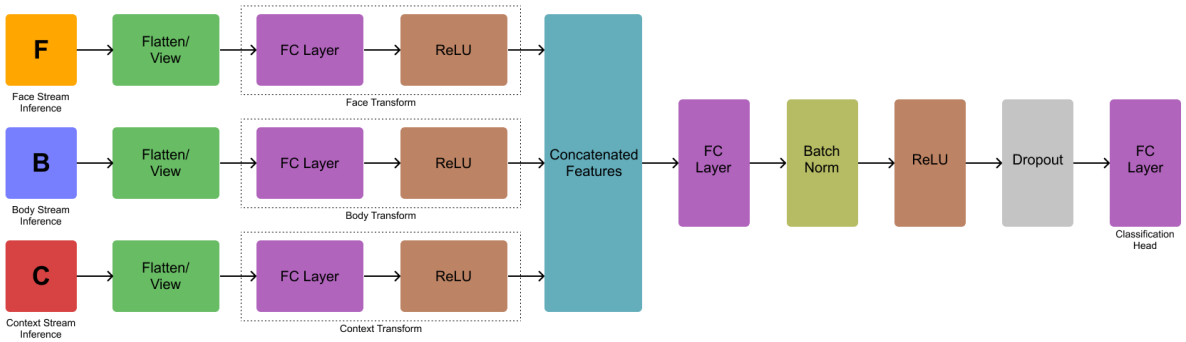

Figure 8: Transformation Fusion Model Architecture.

5. **Cross-Modal Transformer Fusion:** Figure 9 illustrates this technique, which explicitly models interactions between modalities using Transformer mechanisms (Wang et al., 2023). Input tensors are projected into a common embedding dimension using 1x1 Conv2d layers. Features are spatially pooled and reshaped into token sequences. Body and Face tokens are concatenated. A MultiheadAttention layer uses Context tokens as the query and concatenated Body + Face tokens as key and value. The architecture incorporates standard Transformer block components: the attention output is added to the original Context tokens (residual connection), followed by LayerNorm. This is then passed through a Feed-Forward Network (FFN: Linear $\rightarrow$ ReLU $\rightarrow$ Linear), with another residual connection adding the input of the FFN to its output. The final fused representation is obtained by averaging the resulting tokens before the classification layers.

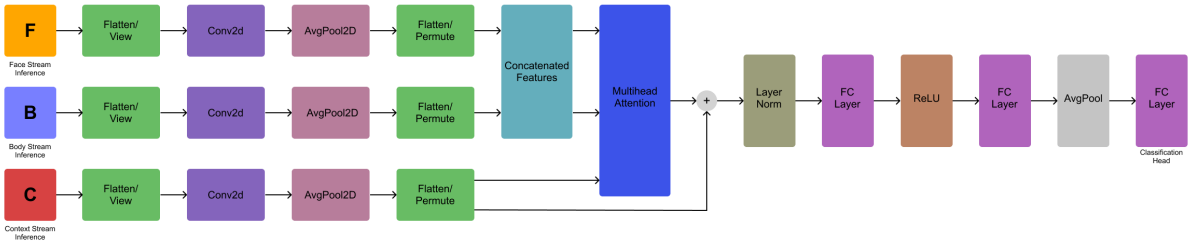

Figure 9: Cross-Modal Transformer Fusion Model Architecture.

6. **Low-Rank Bilinear Fusion:** This method, shown in Figure 10, captures second-order feature interactions efficiently (Chu et al., 2021). Flattened C, B, F features are first projected to an intermediate dimension (256) with Linear layers and ReLU. For each projected vector, two separate Linear layers compute low-rank factors (U and V, projecting to dimension rank). The element-wise product of these U and V factors forms the bilinear representation for that modality. These three low-rank representations (Bilinear C, Bilinear B, Bilinear F) are concatenated and passed through a final processing block (Linear $\rightarrow$ ReLU $\rightarrow$ BatchNorm1d $\rightarrow$ Dropout) before classification.

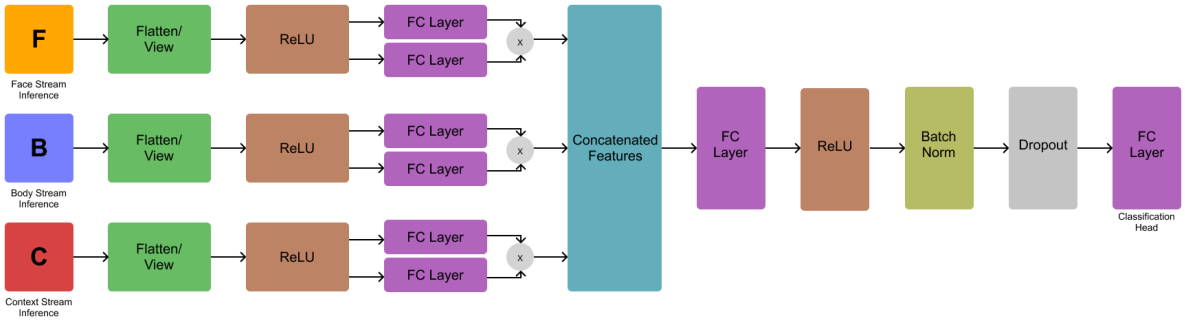

Figure 10: Low-Rank Bilinear Fusion Model Architecture.

7. **Gated Residual Fusion:** Figure 11 shows this method, which employs dynamic gating to control information flow (Arevalo et al., 2017). Flattened C, B, F features are projected (Linear + ReLU). An initial fused representation is formed by averaging these projections. This initial fusion then feeds into three separate gate networks (each 'Linear -¿ Sigmoid') to compute scalar gate values for C, B, and F. Each gate value multiplicatively modulates its corresponding original projected feature. These gated features are then sequentially added back to the initial fused representation via residual connections. The result passes through an integration block (Linear $\rightarrow$ ReLU $\rightarrow$ BatchNorm1d $\rightarrow$ Dropout) before the classification head.

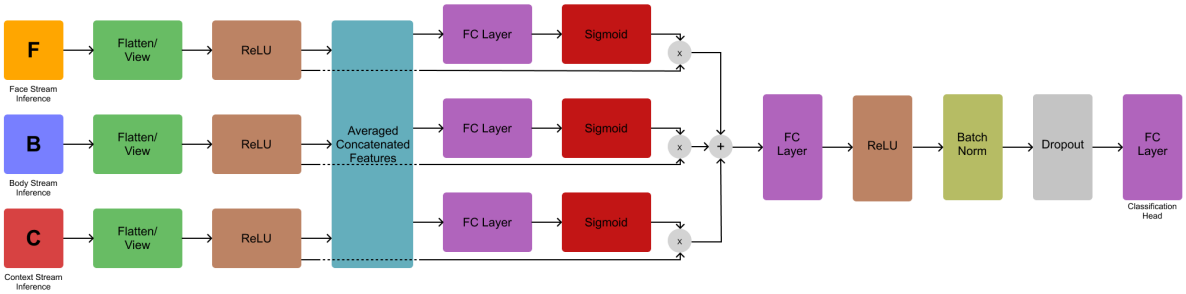

Figure 11: Gated Residual Fusion Model Architecture.

8. **Hierarchical Fusion:** This approach, illustrated in Figure 12, fuses modalities in stages, assuming a semantic hierarchy (Zeng et al., 2023). Flattened Body and Face features are first concatenated and passed through a dedicated fusion block. The resulting fused Body-Face representation is then concatenated with the flattened Context features. This final concatenated vector is processed by subsequent layers (Linear $\rightarrow$ ReLU $\rightarrow$ BatchNorm1d $\rightarrow$ Dropout) before classification.

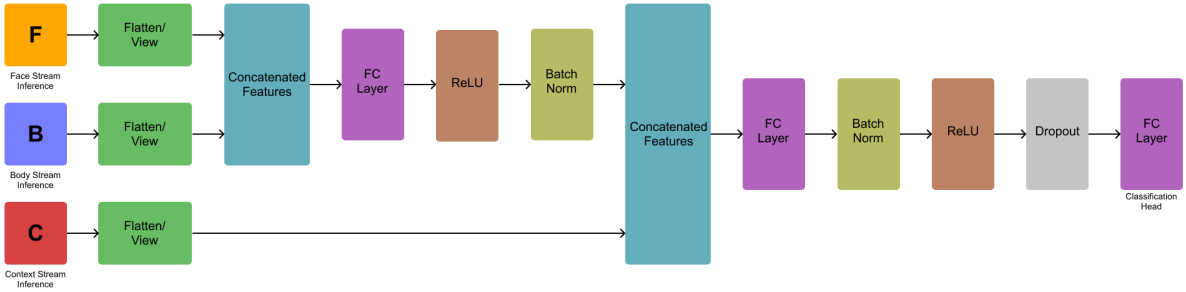

Figure 12: Hierarchical Fusion Model Architecture.

9. **Layer Normalisation Fusion:** Figure 13 depicts this method, which focuses on pre-normalisation. LayerNorm is applied independently to each flattened feature vector (C, B, F) before they are combined (Tarekegn et al., 2021). The aim is to stabilise training and balance the influence of different streams prior to their concatenation and processing by shared subsequent layers (Linear $\rightarrow$ BatchNorm1d $\rightarrow$ ReLU $\rightarrow$ Dropout).

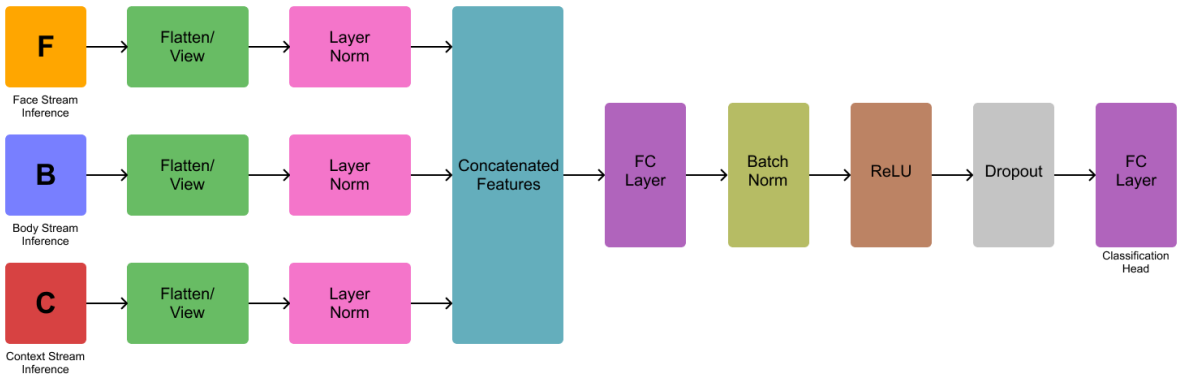

Figure 13: Layer Normalisation Fusion Model Architecture.

10. **Bottleneck Fusion:** Illustrated in Figure 14, this strategy emphasises information compression before fusion (Nagrani et al., 2021). Flattened features from C, B, and F are each projected through separate Linear layers followed by ReLU down to a lower, shared 'bottleneck' dimension (e.g., 128). These compressed, lower-dimensional representations are then concatenated and fed into the subsequent shared processing layers (Linear $\rightarrow$ BatchNorm1d $\rightarrow$ ReLU $\rightarrow$ Dropout) for final classification.

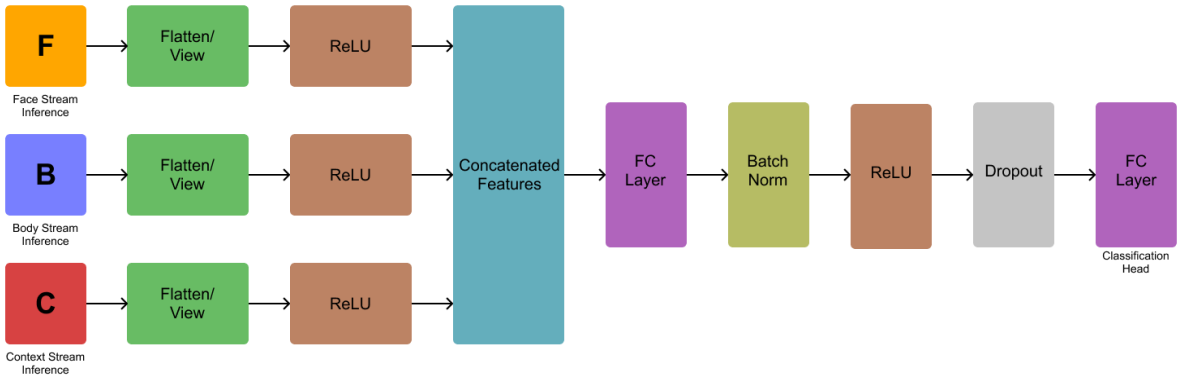

Figure 14: Bottleneck Fusion Model Architecture.

The goal of evaluating these diverse strategies, detailed in Section 5, was to identify fusion mechanisms that effectively balance model complexity, computational cost, and the ability to capture the nuanced interactions between context, body, and face cues for emotion recognition.

4.5 Multi-Label Classification and Loss Functions

The output layer of the framework consists of a fully connected layer that maps the fused feature vector to 26 output units, corresponding to the 26 discrete emotion categories in EMOTIC. A Sigmoid activation function is applied to each output unit to

produce independent probabilities for each emotion label, suitable for the multi-label classification task. The choice of loss function is critical for training effective multi-label classification (MLC) models, especially given the severe class imbalance in EMOTIC (Audibert et al., 2024). This research evaluated several loss functions:

Binary Cross-Entropy (BCE) Loss: This is the standard baseline loss for multi-label problems, treating each label classification independently (Audibert et al., 2024). It measures the difference between the predicted probability distribution $\mathbf{p}_i$ and the true binary distribution $\mathbf{y}_i$ for each sample i. The PyTorch implementation 'BCEWithLogitsLoss' combines Sigmoid activation and BCE calculation for better numerical stability. However, it is highly sensitive to class imbalance. The loss for a batch of N samples and L labels is given by Equation (1):

$$L_{BCE} = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}[y_{il}\log(\sigma(x_{il})) + (1-y_{il})\log(1-\sigma(x_{il}))] \tag{1}$$

where x_{il} is the logit output for sample i, label l, y_{il} is the true label, and $\sigma(\cdot)$ is the sigmoid function.

Discrete Loss: To counteract imbalance, BCE can be weighted (Yasuda et al., 2024). The DiscreteLoss implemented in this work uses a weighted squared error formulation. It calculates the squared difference between the predicted probability $p_{il} = \sigma(x_{il})$ and the true label y_{il}, multiplied by a weight w_l. Weights are calculated dynamically based on the inverse log frequency of positive labels within the current batch, as shown in Equation (2):

$$L_{Discrete} = \sum_{i=1}^{N}\sum_{l=1}^{L} w_l(p_{il} - y_{il})^2 \tag{2}$$

(Note: The summation is used directly as the loss value in the implementation, rather than the mean).

Focal Loss: Designed by Ridnik et al. (2021) to address class imbalance by reducing the loss contribution from easy, well-classified examples. It adds a modulating factor $(1-p_t)^\gamma$ to the standard cross-entropy loss, where p_t is the probability of the correct class and $\gamma \geq 0$ is a tunable focusing parameter. Higher γ values increase the focus on hard-to-classify examples. The WeightedFocalLoss variant further incorporates per-class weights α_l, as defined in Equation (3):

$$L_{FL} = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}\alpha_{t,l}(1-p_{t,il})^\gamma\log(p_{t,il}) \tag{3}$$

where $p_{t,il} = p_{il}$ if $y_{il} = 1$, else $1 - p_{il}$. $\alpha_{t,l}$ represents optional class weighting.

Asymmetric Loss (ASL): Specifically engineered for MLC imbalance by Ridnik et al. (2021), ASL decouples the focusing parameters for positive (γ_+) and negative (γ_-) samples. Typically, γ_- is set higher than γ_+ to aggressively down-weight easy negatives while preserving the loss contribution from rare positives. It can also incorporate probability margin clipping. A simplified form is shown in Equation (4):

$$L_{ASL} \approx -\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}[y_{il}(1-p_{il})^{\gamma_+}\log(p_{il}) + (1-y_{il})(p_{il})^{\gamma_-}\log(1-p_{il})] \tag{4}$$

(Actual implementation may include clipping).

Dice Loss: Originating from image segmentation (Yeung et al., 2023), this loss directly maximises the Dice Similarity Coefficient (DSC), a measure of overlap. It is inherently robust to class imbalance. The loss is typically defined as $1 - DSC$, calculated using Equation (5):

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}\sum_{l=1}^{L} p_{il}y_{il} + \epsilon}{\sum_{i=1}^{N}\sum_{l=1}^{L}(p_{il} + y_{il}) + \epsilon} \tag{5}$$

where ϵ is a small smoothing constant.

Multi-Label Loss with Self Correction (MLLSC): Inspired by methods designed to handle noisy or missing labels (Ghiassi et al., 2023), this loss uses the model's own prediction confidence to correct the loss calculation. It employs thresholds (τ for positive confidence, τ' for negative confidence) to estimate whether a predicted positive or negative label is likely correct or likely a result of noise/error, adjusting the standard log-loss calculation accordingly for positive (L_{pos}) and negative (L_{neg}) components. The loss is computed as shown in Equation (6):

$$L_{MLLSC} \approx -\frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}[y_{il}L_{pos,il} + (1 - y_{il})L_{neg,il}] \tag{6}$$

where $L_{pos,il}$ and $L_{neg,il}$ are modified log-losses based on confidence thresholds τ, τ' applied to the predicted probability p_{il}. Specifically, $L_{pos,il} = \mathbb{I}(p_{il} > \tau)\log(p_{il}) + \mathbb{I}(p_{il} \leq \tau)\log(1 - p_{il})$ and $L_{neg,il} = \mathbb{I}(p_{il} < \tau')\log(1 - p_{il}) + \mathbb{I}(p_{il} \geq \tau')\log(p_{il})$, where $\mathbb{I}(\cdot)$ is the indicator function.

The comparative performance of these loss functions, governed by Equations (1) through (6), in addressing the specific challenges of the EMOTIC dataset is presented and discussed in Section 5.

4.6 Data Augmentation

To enhance model generalisation, improve robustness to variations in input data, and mitigate the detrimental effects of class imbalance inherent in the EMOTIC dataset, two primary categories of data augmentation techniques were systematically employed and evaluated during the training phase: image-level augmentation and dataset-level augmentation (resampling) applied in the feature space.

4.6.1 Image-Level Augmentation

Image-level augmentation involves applying transformations directly to the input images (context, body, and face streams) on-the-fly during training batch preparation. This process artificially expands the diversity of the training set without requiring additional labelled data, thereby reducing the likelihood of overfitting and encouraging the model to learn more invariant feature representations (Shantharam and Schwenker, 2024). The specific transformations used include:

Geometric Transformations These transformations alter the spatial orientation or geometry of the image, simulating variations in viewpoint, camera angle, or subject position.

- **Random Horizontal Flip:** This technique mirrors the image along its vertical axis with a 50% probability. This value is a common default in image augmentation pipelines and libraries (e.g., Shorten and Khoshgoftaar, 2019) as it provides a simple yet effective way to potentially double the useful dataset size for tasks where horizontal orientation is not semantically critical. While other probabilities could be chosen, 50% ensures an equal chance of seeing the original or flipped image during training. Optimal probabilities are often task-dependent and require empirical tuning based on validation set performance, which was considered beyond the scope of this initial parameter setting phase. An example is shown in Figure 15.

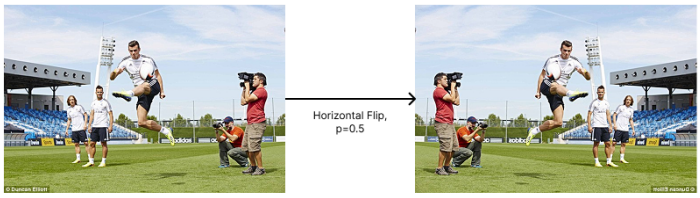

Figure 15: Example of Random Horizontal Flip.

- **Random Affine Transformation:** This applies a combination of geometric transformations randomly within specified ranges. The transformations included rotation (e.g., ±15 degrees), translation (shifting the image horizontally and vertically by a fraction of its size, e.g., ±15%), and scaling (zooming in or out, e.g., between 80% and 120%). These ranges are standard defaults often used to simulate moderate viewpoint changes (Shorten and Khoshgoftaar, 2019). This simulates a wider range of viewpoint changes and variations in subject distance or framing. An example is shown in Figure 16.

Figure 16: Example of Random Affine Transformation.

Photometric Transformations These transformations modify the pixel values related to colour and brightness, simulating variations in lighting conditions, camera sensors, or post-processing effects.

- **Colour Jitter:** This randomly adjusts the brightness, contrast, saturation, and hue of the image within defined limits (e.g., varying brightness and contrast by up to 40%, saturation by up to 40%, and hue by up to 10%). These specific ranges are common defaults in libraries like PyTorch's 'torchvision.transforms' and represent a moderate level of colour variation often encountered in real-world images (Shorten and Khoshgoftaar, 2019). This helps the model become less sensitive to variations in colour and illumination. An example is shown in Figure 17.

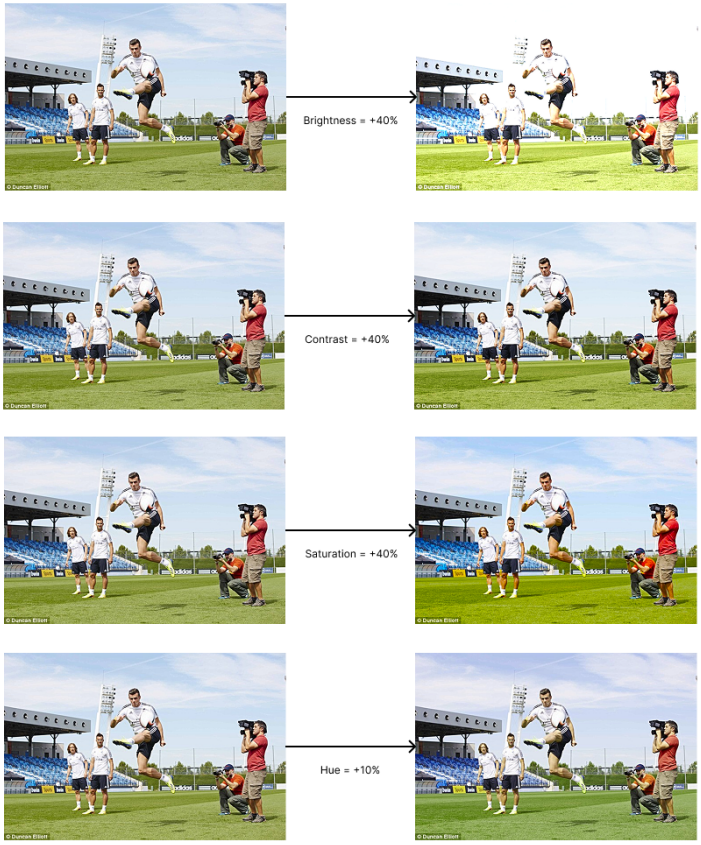

Figure 17: Example of Colour Jitter.

Noise and Blur These techniques simulate degradations in image quality.

- **Random Gaussian Blur:** Applies a Gaussian blur filter to the image with a randomly chosen standard deviation (sigma) within a specified range (e.g., 0.1 to 2.0), applied with a 50% probability. The sigma range allows for varying degrees of blur, while the 50% probability, as noted by Shorten and Khoshgoftaar

(2019), is a standard choice balancing the introduction of blur with retaining original sharpness. This simulates effects like slight defocus or atmospheric conditions and encourages the model to learn more robust features. An example is shown in Figure 18.

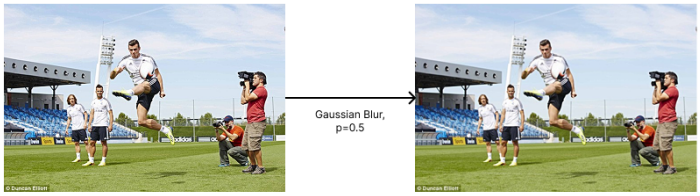

Figure 18: Example of Random Gaussian Blur.

- **Random Gaussian Noise:** Adds noise sampled from a Gaussian distribution (with a specified mean, typically 0, and standard deviation, e.g., 0.05 relative to pixel intensity range) to the image pixels, applied with a 50% probability. The standard deviation of 0.05 represents a moderate level of noise, and the 50% probability is again a common default (Shorten and Khoshgoftaar, 2019). This simulates sensor noise or transmission artefacts, potentially improving model robustness. An example is shown in Figure 19.

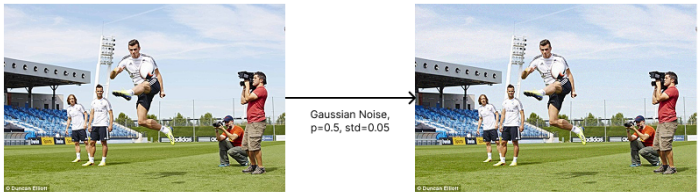

Figure 19: Example of Random Gaussian Noise.

Mixing Augmentations These advanced techniques create new training samples by combining information from multiple images, acting as a strong form of regularisation.

- **Mixup:** Generates a new sample by performing a convex combination (weighted average) of two randomly selected images and their corresponding labels (Shantharam and Schwenker, 2024). The mixing coefficient (lambda) is typically sampled from a Beta distribution. This encourages the model to learn smoother decision boundaries and exhibit more linear behaviour between training samples, potentially improving generalisation and robustness (Shantharam and Schwenker, 2024).
- **CutMix:** Involves cutting a random rectangular patch from one training image and pasting it onto another randomly selected training image (Yu et al., 2024). The ground truth label is mixed proportionally to the area of the combined patches. This forces the model to recognise objects/patterns based on partial views and utilise information from the entire image rather than focusing only on the most

discriminative parts, improving robustness and localisation capabilities (Mensink and Mettes, 2023).

These image-level augmentations were applied probabilistically during the data loading phase for training samples only. Different combinations and intensities of these techniques (often grouped into levels denoted as 'none', 'basic', 'full' corresponding to increasing complexity) were systematically applied to the context, body, and face streams, and their effects were evaluated experimentally, as detailed in Section 5. The choice of specific probability values (e.g., 50%) and parameter ranges follows common practices and defaults widely used in the field and established libraries (Shorten and Khoshgoftaar, 2019), though it is acknowledged that optimal values ideally should be tuned based on validation performance for the specific task and dataset.

4.6.2 Dataset-Level Augmentation

Distinct from image-level augmentation, dataset-level techniques were explored to directly address the severe class imbalance present in the EMOTIC dataset (Audibert et al., 2024). These methods typically operate in the feature space after initial feature extraction by the backbone networks, aiming to create a more balanced distribution before the fusion and classification stages. The primary techniques investigated were oversampling methods that generate synthetic data for minority classes:

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE is a widely adopted algorithm that generates synthetic samples for minority classes (Chawla et al., 2002). For each instance belonging to a minority class (considering each emotion label independently in the multi-label context), SMOTE identifies its k nearest neighbours within the same class in the feature space. It then creates synthetic samples by linearly interpolating between the original instance and one or more of these randomly selected neighbours. Specifically, for a minority sample x_i, it selects a random neighbour x_{zi} from its k nearest minority neighbours. A new synthetic sample x_{new} is generated as $x_{new} = x_i + \delta \times (x_{zi} - x_i)$, where δ is a random number between 0 and 1. This process effectively creates new samples along the line segments joining minority class instances and their neighbours, expanding the decision region for the minority class. This approach avoids simple duplication of minority samples, potentially leading to better model generalisation by creating more diverse, yet plausible, minority examples (Halim et al., 2023).

ADASYN (Adaptive Synthetic Sampling): ADASYN builds upon SMOTE by adaptively generating more synthetic samples for minority class instances that are considered "harder to learn" (He et al., 2008). The difficulty is typically estimated by examining the k-neighbourhood of a minority instance: if the neighbourhood contains a higher proportion of majority class instances, that minority instance is deemed harder to classify correctly. ADASYN assigns a density distribution based on this difficulty measure and generates proportionally more synthetic samples around the harder-to-learn minority instances, effectively focusing the oversampling effort near the decision boundaries where misclassifications are more likely (He et al., 2008).

In this work, these resampling techniques were typically applied in a label-wise manner due to the multi-label nature of the problem. That is, for each under-represented emotion category, synthetic feature vectors were generated based on the feature vectors of instances labelled with that emotion, aiming to balance the representation of each category before the final classification stage. The effectiveness of these methods is evaluated in Section 5.

4.7 Evaluation Metrics

Evaluating Multi-Label Classification (MLC) performance requires metrics that go beyond simple accuracy, as standard accuracy does not adequately capture performance when multiple labels can be assigned or when classes are imbalanced (Zhang and Zhou, 2013). Common metrics used in this study include:

Precision: This metric measures the accuracy of the positive predictions made by the model for a specific label l. It answers the question: "Of all instances predicted as label l, what fraction were actually label l?" High precision indicates that the model makes few False Positive errors for that label. It is particularly important when the cost of falsely assigning a label is high. Precision is calculated using Equation (7):

$$Precision_l = \frac{TP_l}{TP_l + FP_l} \tag{7}$$

where TP_l is the number of true positives and FP_l is the number of false positives for label l.

Recall (Sensitivity): This metric measures the model's ability to identify all relevant instances for a specific label l. It answers the question: "Of all instances that truly belong to label l, what fraction did the model correctly identify?" High recall indicates that the model makes few False Negative errors, meaning it successfully finds most instances of that label. It is crucial when failing to detect a relevant label is costly. Recall is calculated using Equation (8):

$$Recall_l = \frac{TP_l}{TP_l + FN_l} \tag{8}$$

where FN_l is the number of false negatives for label l.

F1 Score: This metric provides a single measure that balances both Precision and Recall by calculating their harmonic mean. The harmonic mean is used because it penalises extreme values more heavily than the arithmetic mean, meaning that a high F1 score requires both high precision and high recall. It is often considered a more informative measure than accuracy, especially on imbalanced datasets where high accuracy can be achieved by simply predicting the majority class. F1 Score is calculated using Equation (9):

$$F1_l = \frac{2 \times Precision_l \times Recall_l}{Precision_l + Recall_l} = \frac{2TP_l}{2TP_l + FP_l + FN_l} \tag{9}$$

Mean Average Precision (mAP): This is a ranking-based metric widely used for evaluating MLC systems, particularly in information retrieval and object detection contexts, and is the standard metric reported for EMOTIC benchmarks (Etesam et al., 2024). Average Precision (AP) for a single class l summarises the shape of the precision-recall curve for that class, effectively representing the model's ability to rank positive instances higher than negative ones across various decision thresholds. It is calculated as the weighted sum of precisions achieved at each threshold where recall increases, effectively approximating the area under the precision-recall curve, as shown in Equation (10):

$$AP_l = \sum_{k=1}^{N} (Recall_k - Recall_{k-1}) \times Precision_k \tag{10}$$

where k indexes the rank in the ordered predictions based on model confidence scores, and $Precision_k$ and $Recall_k$ are the precision and recall calculated considering predictions down to rank k. Mean Average Precision (mAP) is then computed by averaging the AP values across all L emotion categories, as per Equation (11):

$$mAP = \frac{1}{L} \sum_{l=1}^{L} AP_l \tag{11}$$

A higher mAP indicates better performance in correctly identifying and ranking the relevant emotion labels for each instance. While inference might involve selecting a specific operating threshold (e.g., where precision equals recall, as determined from the validation set or precision-recall curve analysis), the mAP metric itself evaluates the overall ranking quality independent of a single threshold. For benchmarks like EMOTIC, mAP is frequently used because it evaluates the model's ability to correctly identify relevant labels and rank them appropriately, which is crucial in MLC where users might only be interested in the top-k predictions (Majumder et al., 2020).

4.8 Chapter Summary

This section detailed the methodology employed in the dissertation. It began with the data preprocessing steps, outlining how context, body, and face image streams were generated from the EMOTIC dataset and prepared for model input using transformations and normalisation (Section 4.1, Figure 3). The proposed multi-modal framework architecture, involving parallel feature extraction, fusion, and classification, was presented (Section 4.2, Figure 4). The candidate feature extraction backbones (CNNs and Transformers) considered were described with justifications (Section 4.3). A comprehensive taxonomy of the investigated feature fusion strategies was provided, illustrating the different approaches with diagrams and explanations (Section 4.4, Figures 5-14). The handling of the multi-label classification task, including the evaluated loss functions (BCE, Discrete Loss, Focal Loss, ASL, Dice Loss, MLLSC) and their formulations (Equations (1)-(6)), was detailed (Section 4.5). Data augmentation techniques, covering both image-level transformations with examples (Section 4.6.1, Figures 15-19) and dataset-level resampling with SMOTE and ADASYN (Section 4.6.2), were

explained. Finally, the key evaluation metrics (Precision, Recall, F1 Score, mAP), defined by Equations (7)-(11), used to assess model performance were presented with detailed explanations (Section 4.7).

5 Experiments, Results and Discussion

This section presents the experimental evaluation of the multi-modal framework developed for context-aware emotion classification on the EMOTIC dataset, as detailed in Section 4. To systematically evaluate the impact of various design choices and identify an effective configuration, a five-stage experimental approach was adopted. This involved sequentially optimising key components of the system: feature extraction backbones, feature fusion methods, loss functions, image-level augmentation, and dataset-level augmentation, using the best performing configuration from the preceding stage as the baseline for the next. This methodical process allows for a clearer attribution of performance changes to specific components and directly addresses the research objectives outlined in Section 1 concerning the investigation and comparison of these techniques.

Each subsection that follows (5.2 to 5.6) corresponds to one experimental stage, providing a description of the specific experiment methodology, along with the presentation of quantitative results (primarily mAP) and a detailed discussion analysing the findings and their implications for the overall framework design. Section 5.7 summarises the optimal configuration derived from these experiments, and Section 5.8 compares the final results against existing work. Finally, Section 5.9 concludes the chapter.

5.1 Experiment Settings

The experiments detailed in this section were conducted within a consistent computational environment and adhered to standardised procedures for data handling, training, and evaluation. All implementations were developed using Python, leveraging core libraries essential for deep learning research. Specifically, PyTorch served as the primary framework for defining, training, and evaluating the neural network models. The NumPy library was employed for efficient numerical operations, particularly for loading and handling the dataset arrays. For accessing diverse pre-trained backbone architectures, the TIMM (PyTorch Image Models) library was utilised. Evaluation metrics, including precision, recall, and the components required for calculating mean Average Precision (mAP), were computed using functionalities from the Scikit-learn library. Visualisations such as loss curves and confusion matrices were generated using Matplotlib.

Data handling was optimised for efficiency. Instead of processing the original EMOTIC image files and MATLAB annotations during each run, the data was loaded from pre-processed NumPy arrays. These arrays, containing extracted image patches for the context, body, and face streams along with corresponding labels, were generated beforehand using the official 'mat2py.py' script provided by the EMOTIC dataset creators (Kosti et al., 2017). This conversion process was performed separately for the predefined training, validation, and test splits established by the dataset organisers (Etesam et al., 2024), and this research strictly adhered to these splits (70% train, 10% validation, 20% test) to ensure methodological consistency and comparability with existing literature. Data pre-processing involved converting the loaded image arrays into

PyTorch tensors suitable for GPU computation. Pixel values were normalised using standard ImageNet mean and standard deviation statistics; distinct ImageNet-derived normalisation weights were applied specifically for inputs to Swin Transformer backbones, following recommended practices (Kim et al., 2023). Emotion category labels were similarly converted to tensor format.

All experiments were executed on a desktop system featuring an AMD Ryzen 9 9900X CPU, 32GB of system RAM, and an NVIDIA RTX 2060 Super GPU equipped with 8GB of VRAM. Model training and inference were accelerated using the GPU via PyTorch's CUDA integration. The 8GB VRAM capacity presented a practical constraint, limiting the maximum feasible batch size during training. A batch size of 32 was generally adopted as it proved manageable for all model configurations. This hardware limitation also influenced the decision to standardise training duration to 20 epochs for most comparative experiments, representing a balance between allowing sufficient time for model convergence and maintaining feasible overall experimentation time.

The model optimisation process consistently employed the Adam optimiser, selected for its adaptive learning rate capabilities and generally robust convergence properties in deep learning tasks. An initial learning rate of 1e-3 was used, typical for Adam. To refine learning over time, a StepLR learning rate scheduler was applied, configured to decrease the learning rate by a factor of 0.1 every 7 epochs. This annealing strategy aids convergence during the later phases of training. Consistent with standard practices for evaluating multi-label classification performance on the EMOTIC dataset (Etesam et al., 2024), mean Average Precision (mAP) served as the primary evaluation metric throughout all experimental stages. Performance on the test split was used for reporting performance results.

To account for stochasticity inherent in the training process, particularly due to random elements in data augmentation and parameter initialisation, experiments from Stage 4 onwards (Image-Level Augmentation and Dataset-Level Augmentation) were executed five times with different random seeds where feasible. While a fixed seed (42) was set for PyTorch and NumPy operations to enhance reproducibility, certain augmentation libraries or operations might introduce variability that cannot be fully controlled by a single seed. Therefore, the results presented for these later stages represent the mean performance (primarily mAP) across these five runs, providing a more robust estimate of the expected outcome for each configuration.

5.2 Stage 1: Backbone Architecture Combination

The primary goal of this initial stage was to determine the most effective combination of pre-trained deep learning models to serve as feature extraction backbones for the context, body, and face input streams, directly addressing Objective (i) which sought to evaluate the representational power of diverse architectures for these distinct visual cues.

Various combinations of candidate backbone architectures (EfficientNet-B7, MobileNetV3, Swin Transformer, DeiT), all pre-trained on ImageNet to leverage transfer learning,

were implemented and tested within the three-stream framework. Each specific combination (e.g., Swin for context, Swin for body, EfficientNet-B7 for face) was trained using a fixed baseline configuration for other system components to isolate the effect of the backbone choice:

- Fusion Method: Simple Concatenation (selected as a straightforward baseline for initial comparison)
- Loss Function: Discrete Loss (chosen provisionally for its potential weighting benefits)
- Image Augmentation: None
- Dataset Augmentation: None

The final classification layer (head) of each pre-trained backbone was replaced with an identity layer, allowing the network to output feature vectors representing the learned visual information prior to the fusion module. Performance was measured by the mean Average Precision (mAP) on the held-out test set after training for 20 epochs. The detailed results for all tested combinations are presented in Table 2.

Table 2: Backbone Combination Performance

Context	Body	Face	Precision	Recall	F1 Score	mAP (%)
Swin	Swin	EfficientNet-B7	0.4341	0.0839	0.0355	***31.35***
Swin	Swin	MobileNetV3	0.4814	0.0877	0.0445	30.94
Swin	EfficientNet-B7	EfficientNet-B7	0.5725	0.0870	0.0377	30.07
Swin	EfficientNet-B7	MobileNetV3	0.4018	0.0875	0.0400	29.84
Swin	DeiT	EfficientNet-B7	0.5042	0.0880	0.0417	30.03
Swin	DeiT	MobileNetV3	0.4496	0.0911	0.0482	29.96
EfficientNet-B7	Swin	EfficientNet-B7	0.4684	0.0822	0.0327	29.66
EfficientNet-B7	Swin	MobileNetV3	0.4394	0.0833	0.0348	29.39
EfficientNet-B7	EfficientNet-B7	EfficientNet-B7	0.3569	0.0851	0.0234	23.76
EfficientNet-B7	EfficientNet-B7	MobileNetV3	0.2864	0.0841	0.0253	23.33
EfficientNet-B7	DeiT	EfficientNet-B7	0.4370	0.0829	0.0282	26.84
EfficientNet-B7	DeiT	MobileNetV3	0.3739	0.0805	0.0281	26.30
DeiT	Swin	EfficientNet-B7	0.4449	0.0869	0.0404	29.75
DeiT	Swin	MobileNetV3	0.4335	0.0829	0.0394	29.92
DeiT	EfficientNet-B7	EfficientNet-B7	0.4629	0.0839	0.0298	26.94
DeiT	EfficientNet-B7	MobileNetV3	0.4775	0.0855	0.0375	27.09
DeiT	DeiT	EfficientNet-B7	0.3828	0.0815	0.0300	27.51
DeiT	DeiT	MobileNetV3	0.4087	0.0866	0.0416	27.24

Analysis of the results in Table 2 reveals clear performance differences between the various backbone combinations. The highest mAP score (31.35%) was achieved when employing the Swin Transformer for both the context and body streams, paired with EfficientNet-B7 for the face stream. This specific configuration significantly outperformed all other tested combinations, suggesting that this particular blend of architectures possesses a superior capability in capturing and representing the diverse visual

features relevant to emotion across the three modalities, at least under the baseline conditions of this stage.

A prominent trend emerging from the data is the general effectiveness of Transformer-based architectures (Swin and DeiT) when applied to the context and body streams, compared to the high-capacity CNN, EfficientNet-B7. Examining the table, configurations using Swin Transformer as the context backbone (rows 1-6) consistently yielded higher mAP scores than those using EfficientNet-B7 for context (rows 7-12), irrespective of the body and face backbone choices. For example, Swin+Swin+EffB7 (31.35%) outperforms EffB7+Swin+EffB7 (29.66%), and Swin+DeiT+MobileNetV3 (29.96%) outperforms EffB7+DeiT+MobileNetV3 (26.30%). A similar pattern holds for the body stream: using Swin or DeiT often resulted in better overall performance than using EfficientNet-B7, particularly when combined with a Transformer context backbone. This empirical evidence suggests that the architectural properties of Transformers, specifically their self-attention mechanisms enabling the modelling of long-range dependencies and providing a global receptive field (Khan et al., 2023), are particularly advantageous for interpreting the broad scene context and holistic body posture information crucial for this task. The Swin Transformer, with its hierarchical structure and efficient shifted window attention mechanism (Liu et al., 2022), appeared especially potent, being part of the top-performing combination.

Regarding the face stream, the results indicate a consistent advantage for using the higher-capacity EfficientNet-B7 over the lightweight, efficiency-focused MobileNetV3. Across nearly all context and body backbone pairings, substituting MobileNetV3 with EfficientNet-B7 for the face stream led to an improvement in mAP (e.g., Swin+Swin+EffB7 at 31.35% vs. Swin+Swin+MobileNetV3 at 30.94

Conversely, configurations relying heavily on EfficientNet-B7, particularly the combination using it for all three streams (EffB7+EffB7+EffB7), yielded the lowest mAP score (23.76%). This poor performance might suggest that simply employing a high-capacity CNN across all streams is suboptimal, perhaps due to feature redundancy between streams or the inability of the simple concatenation fusion method used in this stage to effectively integrate and leverage the specific types of features extracted by CNNs from context and body cues compared to Transformers.

Finally, it is crucial to consider the other metrics reported in Table 2. While precision values exhibit some variation across configurations, the recall and F1 scores remain consistently low for all combinations. This pattern strongly highlights the significant challenge posed by the inherent class imbalance of the EMOTIC dataset, especially when using only a basic loss function without specific balancing techniques like resampling. The models evidently struggle to correctly identify instances belonging to less frequent emotion categories, resulting in a high number of false negatives and consequently poor recall and F1 performance. This observation underscores the importance of using mAP as the primary evaluation metric, as it provides a more robust measure of overall ranking quality in multi-label classification, and it clearly signals the necessity of addressing the class imbalance problem in subsequent experimental stages.

Based on these comprehensive findings, the backbone combination demonstrating the highest mAP performance, namely:

- Context: Swin Transformer
- Body: Swin Transformer
- Face: EfficientNet-B7

was selected as the optimal configuration. This empirically validated combination was carried forward into Stage 2 to facilitate a focused evaluation of different feature fusion methods.

5.3 Stage 2: Feature Fusion Method

Having established the optimal backbone architecture combination (Swin-Swin-EfficientNetB7) in Stage 1, this second stage aimed to systematically evaluate and compare the performance of various feature fusion strategies, thereby addressing Objective (ii) which focused on exploring effective methods for combining features from the separate visual streams.

The framework was trained multiple times, consistently utilising the Swin-Swin-EfficientNetB7 backbone configuration identified previously. The key independent variable in this stage was the specific fusion module employed to integrate the feature vectors extracted from the context, body, and face streams. A comprehensive set of ten distinct fusion methods, representing different conceptual approaches as detailed in Section 4.4, were tested: Simple Concatenation (serving as the baseline from Stage 1), Weighted Fusion, Attention-Based Weighted Fusion, Transformation Fusion, Cross-Modal Transformer Fusion, Low-Rank Bilinear Fusion, Gated Residual Fusion, Hierarchical Fusion, Layer Normalisation Fusion, and Bottleneck Fusion. All other experimental parameters, including the Discrete Loss function and the absence of data augmentation, were kept constant to isolate the effect of the fusion strategy. Performance was primarily assessed using the test set mAP, with results summarised in Table 3.

Table 3: Fusion Method Performance

Fusion Method	Precision	Recall	F1 Score	mAP (%)
Simple Concatenation	0.4341	0.0839	0.0355	31.35
Weighted Fusion	0.4841	0.0833	0.0361	***31.39***
Attention-Based Weighted Fusion	0.4677	0.0852	0.0450	30.25
Transformation Fusion	0.4308	0.0839	0.0490	29.86
Cross-Modal Transformer Fusion	0.4285	0.0868	0.0456	29.94
Low-Rank Bilinear Fusion	0.0817	0.0747	0.0168	28.82
Gated Residual Fusion	0.4731	0.0776	0.0283	30.66
Hierarchical Fusion	0.4906	0.0829	0.0349	31.00
Layer Normalisation Fusion	0.5130	0.0851	0.0373	31.28
Bottleneck Fusion	0.4200	0.0894	0.0543	30.28

The results comparing the ten different feature fusion strategies, presented in Table 3, reveal relatively subtle differences in performance among the top contenders, but

also highlight some clear underperformers. Based on the primary metric, test set mAP, the Weighted Fusion method achieved the highest score (mAP = 31.39%). This represents a marginal improvement over the Simple Concatenation baseline used in Stage 1 (mAP = 31.35%). The Weighted Fusion approach introduces learnable scalar weights for each modality before summing their projected features (Pawłowski et al., 2023). This slight edge suggests that allowing the model to adaptively learn the relative importance of context, body, and face features provides a small but measurable advantage compared to simply concatenating them and relying solely on subsequent fully connected layers to disentangle their contributions.

Other fusion methods demonstrating comparable performance include Layer Normalisation Fusion (mAP = 31.28%) and Hierarchical Fusion (mAP = 31.00%). The effectiveness of Layer Normalisation Fusion, which applies LayerNorm independently to each stream's features before concatenation (Tarekegn et al., 2021), suggests that stabilising the feature magnitudes and potentially balancing the influence of different streams prior to combination is beneficial for training stability and achieving good performance. The decent result obtained with Hierarchical Fusion, which first fuses body and face features before combining them with context features (Zeng et al., 2023), indicates that adopting a structured, staged fusion approach reflecting a potential semantic hierarchy can also be an effective strategy.

Interestingly, several more complex fusion strategies did not yield superior performance within this experimental setup. Attention-Based Weighted Fusion (mAP = 30.25%), which uses a separate network to predict attention weights, and Bottleneck Fusion (mAP = 30.28%), which compresses features before concatenation (Nagrani et al., 2021), performed similarly but slightly below the top methods. This might indicate that the specific attention mechanism implemented was not sufficiently powerful or that the information compression in the bottleneck approach led to some loss of discriminative detail. Gated Residual Fusion (mAP = 30.66%), employing dynamic gating (Arevalo et al., 2017), also fell slightly short.

Methods such as Transformation Fusion (mAP = 29.86%), which applies separate linear transformations before concatenation, and the theoretically powerful Cross-Modal Transformer Fusion (mAP = 29.94%), designed to explicitly model inter-modal interactions using attention (Wang et al., 2023), unexpectedly underperformed compared to the simpler Weighted Fusion and Simple Concatenation. For Transformation Fusion, this might imply that applying independent, fixed transformations before concatenation was less effective than allowing subsequent shared layers to learn interactions from the richer, raw concatenated features. For the Cross-Modal Transformer, its relatively lower performance could suggest that such complex architectures might require more extensive hyperparameter tuning, different feature pre-processing strategies, or significantly more training data or epochs to fully realise their potential compared to simpler methods within the constraints of this 20-epoch baseline experiment.

Particularly noteworthy is the poor performance of Low-Rank Bilinear Fusion, which yielded the lowest mAP (28.82%) and also exhibited significantly worse precision, recall, and F1 scores. This method aims to capture second-order feature interactions efficiently using a low-rank approximation (Chu et al., 2021). Its detrimental effect on performance in this context strongly suggests that either the specific low-rank approx-

imation used resulted in significant information loss, or that modelling these types of second-order interactions is not beneficial, or even harmful, for this specific emotion recognition task with these features.

Consistent with the findings from Stage 1, the recall and F1 scores remain generally low across all tested fusion methods, further reinforcing the conclusion that the choice of fusion strategy, while influencing performance to some degree, does not by itself overcome the fundamental challenge posed by class imbalance when using the baseline Discrete Loss function. However, based on achieving the highest mAP score, albeit by a small margin, Weighted Fusion (mAP = 31.39%) was selected as the optimal fusion strategy. This configuration was carried forward to Stage 3 for a focused evaluation of different loss functions.

5.4 Stage 3: Loss Function

The objective of this third stage was to identify the most effective loss function for training the multi-modal network, specifically addressing the challenges of multi-label classification and severe class imbalance inherent in the EMOTIC dataset, as outlined in Objective (iii).

The experiment utilised the best-performing configuration identified from the preceding stages: Swin-Swin-EfficientNetB7 backbones combined with the Weighted Fusion module. The framework was trained repeatedly, with the only variation being the loss function employed to compute the training objective and guide optimisation. The loss functions compared, representing a range of approaches from standard baselines to specialised imbalance-aware and noise-robust methods as detailed in Section 4.5, included: Binary Cross-Entropy (BCE) Loss, Discrete Loss (with dynamic weighting), Focal Loss (Ridnik et al., 2021), Asymmetric Loss (ASL) (Ridnik et al., 2021), Dice Loss (Yeung et al., 2023), and MLLSC Loss (Ghiassi et al., 2023). All other parameters, including the chosen backbones, the Weighted Fusion method, the absence of data augmentation, and standard training settings (20 epochs, Adam optimiser), remained fixed. Performance was primarily evaluated using the test set mAP, with detailed results presented in Table 4.

Table 4: Loss Function Performance

Loss Function	Precision	Recall	F1 Score	mAP (%)
Binary Cross-Entropy Loss	0.5042	0.0734	0.0172	30.70
Discrete Loss	0.4841	0.0833	0.0361	***31.39***
Focal Loss	0.3261	0.0716	0.0144	30.54
Asymmetric Loss	0.3261	0.0719	0.0144	30.21
Dice Loss	0.3642	0.1312	0.1088	22.60
MLLSC Loss	0.0649	0.1104	0.0675	17.48

This stage focused on evaluating the impact of different loss functions using the optimal configuration established in the previous stages (Swin-Swin-EffNetB7 backbones with

Weighted Fusion, no augmentation). The results, presented in Table 4, clearly demonstrate that the choice of loss function significantly influences model performance, particularly in navigating the challenges posed by the multi-label nature and severe class imbalance of the EMOTIC dataset.

Discrete Loss, which employs a weighted squared error formulation with dynamic weighting based on the inverse log frequency of positive labels within each batch (Yasuda et al., 2024), achieved the highest mAP score (31.39%). This performance marginally surpassed that of the standard Binary Cross-Entropy (BCE) Loss (mAP = 30.70%). This suggests that, within this specific experimental setup, the combination of a squared error term and dynamic, batch-adaptive weighting provided a slight advantage in optimising the model for the mAP metric compared to the standard unweighted BCE approach, which treats all labels and samples equally.

Interestingly, the specialised imbalance-aware loss functions, Focal Loss (mAP = 30.54%) and Asymmetric Loss (ASL) (mAP = 30.21%), did not outperform the simpler BCE or Discrete Loss in terms of mAP. These losses are specifically designed to mitigate class imbalance by down-weighting the contribution of easy, well-classified examples (typically abundant negatives) using modulating factors controlled by focusing parameters (γ) (Ridnik et al., 2021). While theoretically appealing, their performance in this experiment suggests that achieving optimal results might be highly sensitive to the specific choice of these hyperparameters (γ_+, γ_-). Without extensive tuning, which was beyond the scope of this comparative stage, they failed to demonstrate a clear advantage over the less complex losses in terms of the primary mAP ranking metric under these specific training conditions.

Dice Loss exhibited a markedly different performance profile. Although it resulted in the lowest mAP score (22.60%), it achieved significantly higher Recall (0.1312) and F1 Score (0.1088) compared to all other tested loss functions. This behaviour aligns with the known characteristic of Dice Loss, which directly optimises the Dice Similarity Coefficient (a measure of overlap) and is thus inherently more robust to class imbalance, forcing the model to identify a greater proportion of positive instances, even for rare classes (Yeung et al., 2023). However, this strong focus on recall appears to negatively impact precision and, crucially, the overall ranking quality across different thresholds, leading to the poor mAP score. This highlights a potential trade-off: Dice Loss can improve detection of minority classes but may harm overall classification and ranking performance.

The MLLSC Loss, inspired by methods for handling noisy or missing labels by using model confidence thresholds (τ, τ') to correct the loss calculation (Ghiassi et al., 2023), performed very poorly (mAP = 17.48%). This outcome suggests that either the specific noise characteristics of the EMOTIC dataset do not align well with the assumptions underlying this loss function, or, more likely, that its performance is critically dependent on careful tuning of the confidence thresholds, which was not undertaken in this comparative evaluation.

In conclusion, while Dice Loss demonstrated potential for improving recall on minority classes, its detrimental effect on the primary mAP metric rendered it unsuitable as the main loss function for this task. Discrete Loss provided the best mAP performance

among the tested options, slightly surpassing standard BCE. Therefore, the Discrete Loss (with dynamic weighting) was selected as the optimal loss function to be carried forward into the subsequent experimental stages focusing on data augmentation. The persistently low recall and F1 scores observed across most loss functions (with the exception of Dice Loss) further underscored that addressing the severe class imbalance effectively likely requires a combination of an appropriate loss function and complementary techniques such as data augmentation or resampling.

5.5 Stage 4: Image-Level Augmentation

The objective of this fourth stage was to systematically evaluate the effect of applying different intensities and types of image-level data augmentation during the training process, directly addressing the part of Objective (iv) related to image-level augmentation strategies.

This stage utilised the best overall configuration identified from the preceding Stages 1-3: Swin-Swin-EfficientNetB7 backbones, Weighted Fusion, and Discrete Loss function. The experiment compared three distinct levels of image augmentation complexity, applied consistently across the context, body, and face input streams during data loading for the training set only, as detailed in Section 4.6.1. The levels were defined as:

- 'None': No image augmentation applied, except for the necessary resizing and centre cropping required for Swin Transformer inputs. This served as the baseline from Stage 3.
- 'Basic': Applied a minimal set of common augmentations: 'RandomHorizontalFlip' (with p=0.5) and 'ColorJitter' (adjusting brightness, contrast, saturation with standard factors). For Swin inputs, these were applied after the initial resize and crop.
- 'Full': Applied a more extensive and diverse set of augmentations, including geometric transformations (RandomAffine with rotation, translation, scale), photometric changes (ColorJitter), quality degradations (RandomGaussianBlur, RandomGaussianNoise), and advanced mixing techniques (RandomCutMixV2, RandomMixUpV2, applied primarily to context and body streams). The specific parameters followed common practices, as detailed in Section 4.6.1.

Recognising the stochastic nature of data augmentation, each configuration ('None', 'Basic', 'Full') was trained five times using different random seeds to ensure the robustness of the findings. The mean performance metrics across these five runs are reported in Table 5.

This stage investigated the impact of incorporating image-level data augmentation into the training process, using the best configuration established previously. Table 5 presents the mean performance results comparing no augmentation ('None'), a 'Basic' set of augmentations, and a more comprehensive 'Full' set.

Table 5: Image Augmentation Level Performance (Mean over 5 runs)

Image Augmentation Level	Precision	Recall	F1 Score	mAP (%)
None	0.4841	0.0833	0.0361	31.39
Basic	0.4987	0.0904	0.0456	***31.73***
Full	0.3968	0.0857	0.0377	31.68

The results clearly indicate that employing some form of image augmentation provides a tangible benefit compared to training without any augmentation. The 'Basic' augmentation level, consisting only of Random Horizontal Flip and basic Colour Jitter, achieved the highest mean mAP score (31.73%). This represents an improvement over the 'None' baseline (mean mAP = 31.39%). Although the gain is modest (approximately 0.34 percentage points in mAP), it consistently appeared across the multiple runs, suggesting that even these simple transformations help the model to generalise better. This improvement likely stems from the model becoming more robust to variations in horizontal orientation and minor lighting/colour changes, which are common in the diverse 'in-the-wild' images of the EMOTIC dataset (Shantharam and Schwenker, 2024). The slightly improved mean recall and F1 score observed for the 'Basic' level compared to 'None' further support the notion that basic augmentation aids in learning more invariant and generalisable features.

Interestingly, applying the more extensive 'Full' set of augmentations did not lead to further improvements and, in fact, resulted in a slightly lower mean mAP (31.68%) compared to the 'Basic' level. This suggests a point of diminishing returns, where adding more complex or aggressive transformations, including affine transformations, noise, blur, and mixing techniques like CutMix and Mixup (applied primarily to the context and body streams in the 'Full' setting), did not provide additional benefits and might even have slightly hindered overall performance within the fixed 20-epoch training schedule. Several factors could contribute to this: the more severe distortions introduced by the 'Full' set might have made it harder for the model to extract subtle emotional cues; the model might require significantly more training epochs to effectively learn from these complex transformations; or the specific combination and parameters of the 'Full' augmentations might not have been optimal for this particular task and dataset. The notably lower mean precision score observed for the 'Full' level compared to 'Basic' (0.3968 vs 0.4987) could also indicate that the more aggressive augmentations introduced increased confusion or ambiguity, leading to more false positive predictions.

Considering the trade-off between the complexity of the augmentation strategy and the resulting performance gain, the 'Basic' level emerges as the most effective and efficient approach in this experiment. It provides a demonstrable, albeit small, improvement in mAP over no augmentation, without the potential drawbacks or increased training demands associated with the 'Full' level. Therefore, 'Basic' image augmentation (specifically, Random Horizontal Flip and basic Colour Jitter) was selected as the optimal image-level augmentation strategy to be included in the configuration for the final stage of experimentation, which focuses on dataset-level augmentation.

5.6 Stage 5: Dataset-Level Augmentation

The objective of this final experimental stage was to assess the effectiveness of applying a feature-space resampling technique, specifically ADASYN, to mitigate the persistent challenge of class imbalance, comparing its impact against using only the optimal image-level augmentation identified in Stage 4. This directly addresses the part of Objective (iv) concerning dataset-level resampling methods.

This stage utilised the best overall configuration identified from all previous stages: Swin-Swin-EfficientNetB7 backbones, Weighted Fusion, Discrete Loss, and 'Basic' image-level augmentation (Random Horizontal Flip and Colour Jitter). The experiment compared two scenarios:

- Baseline: Training with the optimal configuration from Stage 4, incorporating only 'Basic' image augmentation.
- ADASYN: Applying the ADASYN (Adaptive Synthetic Sampling) algorithm (He et al., 2008) in the feature space during training, in addition to the 'Basic' image augmentation. As detailed in Section 4.6.2, ADASYN adaptively generates synthetic feature vectors for minority class instances, focusing on those deemed harder to learn (i.e., closer to the decision boundary).

ADASYN was chosen for evaluation over the simpler SMOTE algorithm (Chawla et al., 2002) because its adaptive nature, which prioritises generating samples in more challenging regions of the feature space, was hypothesised to be potentially more beneficial for navigating the complex decision boundaries and severe imbalance present in the EMOTIC dataset (He et al., 2008; Halim et al., 2023). The experiments for both scenarios (Baseline and ADASYN) were executed five times using different random seeds to ensure reliable results, and the mean performance metrics are reported in Table 6.

Table 6: Dataset Augmentation Performance (Mean over 5 runs)

Dataset Augmentation	Precision	Recall	F1 Score	mAP (%)
Baseline (Basic Image Aug Only)	0.4987	0.0904	0.0456	31.73
ADASYN (+ Basic Image Aug)	0.4455	0.0814	0.0408	***32.00***

The results presented in Table 6 show the impact of incorporating ADASYN feature-space resampling into the best configuration identified thus far. The application of ADASYN resulted in a further, albeit modest, improvement in the primary evaluation metric, mean Average Precision (mAP). The mean mAP increased from 31.73% for the baseline (which included only basic image augmentation) to 32.00% when ADASYN was added. As mAP is the standard benchmark metric for the EMOTIC task, reflecting the overall quality of label ranking across all emotion categories (Etesam et al., 2024), this increase suggests that ADASYN successfully provided an additional enhancement to the model's ability to correctly identify and rank relevant labels, likely

by better representing the minority classes in the feature space before the final classification layers. The adaptive nature of ADASYN, which focuses synthetic sample generation on harder-to-learn minority instances near decision boundaries (He et al., 2008), appears to have yielded a tangible, positive effect on the primary metric in this complex, imbalanced multi-label scenario.

However, it is crucial to acknowledge the observed trade-offs in the secondary metrics. The introduction of ADASYN led to a noticeable decrease in mean Precision (from 0.4987 down to 0.4455) and mean Recall (from 0.0904 down to 0.0814), which consequently resulted in a lower mean F1 Score (from 0.0456 down to 0.0408). This suggests a potential downside to the synthetic oversampling process. While ADASYN improved the overall ranking performance captured by mAP, the generated synthetic samples might have introduced some level of noise, ambiguity, or overlap in the feature space. This could have led the final classifier to make more false positive predictions (hence the lower precision) and potentially misclassify some true positive instances that it might have otherwise captured (hence the lower recall), compared to the baseline model trained without resampling. This finding highlights a potential conflict often encountered with oversampling techniques, particularly in complex, high-dimensional feature spaces: optimising for overall ranking quality (mAP) might come at the cost of per-class identification accuracy (Recall/F1). The synthetic samples might help define decision boundaries better for ranking purposes but could simultaneously make threshold-based classification more challenging.

Despite the observed decrease in these secondary metrics, the improvement in mAP which is the principal metric for this task and the standard for evaluation on the EMOTIC dataset and is considered the deciding factor in this staged optimisation process. The gain in mAP, however small, indicates that ADASYN provided a net positive contribution towards the main research objective of improving context-aware emotion recognition performance as measured by standard benchmarks. Therefore, ADASYN was adopted as part of the final optimal configuration identified by this research, accepting the observed trade-offs in Precision and Recall as a consequence of prioritising the optimisation of the primary mAP metric. The added computational complexity associated with implementing ADASYN (requiring feature extraction before resampling and potentially longer training times if applied dynamically) is deemed justifiable given the achieved improvement in the key performance indicator.

5.7 Summary of Experimental Findings and Optimal Configuration

The systematic, five-stage experimental process detailed in this section (5.2 through 5.6) served to methodically evaluate and refine key components of the proposed multimodal framework for emotion classification on the EMOTIC dataset. Each stage built upon the best performing configuration from the previous one, allowing for a structured exploration of design choices and their impact on performance, primarily measured by mean Average Precision (mAP).

Stage 1 (Section 5.2, Table 2) focused on feature extraction backbones, concluding that the combination of Swin Transformer for the context stream, Swin Transformer for the body stream, and EfficientNet-B7 for the face stream yielded the highest initial mAP of 31.35%. This highlighted the suitability of Transformer architectures for context/body cues and a powerful CNN for facial details.

Stage 2 (Section 5.3, Table 3) evaluated various feature fusion strategies. Weighted Fusion emerged as marginally superior to other methods, including simple concatenation and more complex approaches like cross-modal transformers, achieving a slightly improved mAP of 31.39%. This suggested that adaptive weighting of modalities offered a small benefit.

Stage 3 (Section 5.4, Table 4) compared different loss functions designed to handle multi-label classification and class imbalance. The Discrete Loss function, incorporating dynamic weighting based on inverse log frequency within batches, maintained the highest mAP (31.39%) among the tested options, outperforming standard BCE and specialised losses like Focal Loss and ASL under the experimental conditions. Dice Loss improved recall but significantly harmed mAP.

Stage 4 (Section 5.5, Table 5) investigated the impact of image-level data augmentation. Applying a 'Basic' set of augmentations (Random Horizontal Flip and Colour Jitter) provided a consistent, albeit modest, improvement, raising the mean mAP to 31.73%. More complex ('Full') augmentations did not yield further benefits, indicating diminishing returns.

Finally, Stage 5 (Section 5.6, Table 6) assessed the effect of dataset-level augmentation via feature-space resampling. Incorporating ADASYN to adaptively oversample minority classes resulted in a further slight increase in the primary metric, reaching a final mean mAP of 32.00%, although this came at the cost of reduced precision and recall.

Therefore, based on the systematic optimisation process prioritising the principal mAP metric at each stage, the final optimal configuration identified through this comprehensive experimental evaluation (addressing Objective (v)) is defined as follows:

- **Backbones:** Context Stream = Swin Transformer, Body Stream = Swin Transformer, Face Stream = EfficientNet-B7 (all pre-trained on ImageNet).
- **Fusion Method:** Weighted Fusion (with learnable scalar weights for each modality).
- **Loss Function:** Discrete Loss (weighted squared error with dynamic batch-wise inverse log frequency weighting).
- **Image Augmentation:** Basic Level (Random Horizontal Flip with p=0.5, Colour Jitter with standard factors).
- **Dataset Augmentation:** ADASYN (applied in the feature space before the fusion layer).

This configuration represents the best-performing model developed within the scope, constraints, and methodology of this project. The staged approach proved valuable in isolating the impact of different components and guiding the selection process based on empirical evidence. While the final mAP of 32.00% underscores the inherent difficulty of the EMOTIC dataset and the task of context-aware emotion recognition, this systematically optimised configuration provides a robust baseline and valuable insights into effective architectural and training choices.

5.8 Comparison with Existing Work

To effectively contextualise the performance achieved by the optimal configuration developed through this research, this section compares its final mAP score with those reported by various existing methods on the EMOTIC dataset benchmark. Table 7 presents a summary of these results, highlighting the core techniques and reported mAP scores of several key benchmark approaches alongside the result from this work.

Table 7: Comparison of Overall Performance on EMOTIC (mAP)

Method Name	Key Technique(s)/Streams	Backbone(s)	Reported mAP (%)
EMOTIC Baseline (Kosti et al., 2017)	Two-branch CNN (Body, Context)	CNN (Low-Rank)	27.38
DRM (Chen et al., 2023)	Structured Knowledge & Relations	ResNet-50	26.48
High-Level Context (de Lima Costa et al., 2023)	Enhanced Context Representation	ResNet-50	30.02
TEKG (Chen et al., 2023)	Structured Knowledge & Relations	ResNet-50	31.36
EmotionCLIP (Etesam et al., 2024)	CLIP Linear Probe	CLIP (ViT variant)	32.91
EmotiCon (Depth) (Mittal et al., 2020)	Multi-Context, Depth Maps	CNN	35.48
This Work	Three-stream (Face, Body, Context)	Swin, EfficientNet-B7	32.00

The optimal configuration identified in this dissertation achieved a final mean Average Precision (mAP) of **32.00%**. As evidenced in Table 7, this result represents a substantial improvement over the original EMOTIC baseline (27.38%) (Kosti et al., 2017) and also surpasses several more recent methods, including the DRM configuration (26.48%) (Chen et al., 2023) and the High-Level Context approach (30.02%) (de Lima Costa et al., 2023). Furthermore, this work achieves a slightly higher mAP than the TEKG configuration (31.36%) (Chen et al., 2023), which notably employs external knowledge graphs and relation modelling on top of standard ResNet-50 backbones.

This level of performance is particularly noteworthy given the methodology employed. The framework developed here relies fundamentally on the three visual streams directly available within the EMOTIC dataset itself: context, body, and face. The success in reaching an mAP of 32.00% stems primarily from the **systematic optimisation of standard deep learning components**. Through the staged experimental process, effective choices were made for backbone architectures (leveraging the strengths of Swin Transformers for context/body and EfficientNet-B7 for face), feature fusion (identifying Weighted Fusion as effective), loss function design (utilising a custom weighted Discrete Loss), and data augmentation strategies (combining basic image augmentation with ADASYN resampling). This demonstrates that significant performance gains can be achieved on this complex task by carefully tuning a well-structured multi-modal architecture using established techniques, without resorting to external data sources

or highly specialised modules.

It is acknowledged that the achieved mAP is slightly below that of EmotionCLIP (32.91%) (Etesam et al., 2024), which utilises linear probing on features from the large pre-trained vision-language model CLIP, and notably lower than the state-of-the-art EmotiCon (Depth) method (35.48%) (Mittal et al., 2020). However, it is crucial to recognise the increased complexity inherent in these higher-performing approaches. EmotionCLIP leverages the vast knowledge implicitly encoded within CLIP after pre-training on massive web-scale image-text datasets. EmotiCon (Depth) achieves its superior performance by incorporating an entirely separate modality (depth maps, requiring additional data or estimation) and employing sophisticated techniques to explicitly model social interactions and multiple context types.

In contrast, the framework presented in this dissertation achieves its strong result by effectively maximising the information extracted solely from the provided context, body, and face image streams through careful architectural design and training optimisation. This makes the approach potentially more adaptable and less reliant on external resources or complex pre-processing steps compared to some state-of-the-art methods. Therefore, achieving a competitive mAP of 32.00%, surpassing several established benchmarks including knowledge-enhanced ones, through the systematic optimisation of a foundational three-stream visual framework can be considered a significant success for this research project.

A more granular comparison involves examining the Average Precision (AP) for individual emotion categories. It is important to note that such detailed per-category results are not always reported in the literature, limiting the scope of direct comparison. The following analysis compares the results of this work against the original EMOTIC baseline (Kosti et al., 2017) and the DRM/TEKG configurations (Chen et al., 2023), as these studies provided the necessary per-category AP data.

The per-category AP scores in Table 8 further reinforce the strength of the optimised framework developed in this work. It achieves state-of-the-art performance among the compared methods for a remarkable number of categories, including **Affection, Anger, Annoyance, Disapproval, Fatigue, Happiness, Pain, Sadness, Sensitivity, and Suffering**. The substantial improvements observed for key negative emotions like Sadness (41.19% vs 18.30% for TEKG) and Suffering (40.09% vs 17.91% for TEKG), as well as Anger and Annoyance, are particularly significant. This suggests that the specific combination of Swin/EfficientNet backbones, weighted fusion, the custom discrete loss, and ADASYN resampling is highly effective at discriminating these often challenging affective states directly from visual cues. The top performance on Happiness is also notable, indicating effective recognition of this frequent category.

While the model lags behind the knowledge-enhanced TEKG method on several other categories (e.g., Anticipation, Confidence, Engagement, Esteem, Excitement, Peace, Pleasure), particularly those that might be considered more socially complex or context-dependent, the ability to outperform established methods on nearly 40% of the emotion categories using only optimised visual processing is a strong testament to the effectiveness of the approach. The struggles with extremely rare or subtle categories like Embarrassment and Surprise are shared across most methods, highlighting persistent

Table 8: Comparison of Per-Category Average Precision (AP, %)

Category	EMOTIC (Kosti et al., 2017)	DRM (Chen et al., 2023)	TEKG (Chen et al., 2023)	**This work**
Affection	27.85	32.99	41.89	***43.71***
Anger	9.49	10.63	11.67	***18.34***
Annoyance	14.06	14.35	16.56	***19.58***
Anticipation	58.64	58.06	67.26	59.28
Aversion	7.48	7.01	9.50	9.49
Confidence	78.35	76.06	84.68	78.97
Disapproval	14.97	13.77	15.32	***18.45***
Disconnection	21.32	26.20	38.53	31.75
Disquietment	16.89	17.46	22.14	21.27
Doubt/Confusion	29.63	17.42	25.26	21.11
Embarrassment	3.18	2.34	4.60	2.55
Engagement	87.53	85.58	90.12	87.04
Esteem	17.73	16.26	24.79	18.55
Excitement	77.16	69.70	78.95	72.66
Fatigue	9.70	13.44	15.74	***17.73***
Fear	14.14	6.10	8.76	10.62
Happiness	58.26	65.86	74.13	***77.63***
Pain	8.94	8.65	8.58	***12.85***
Peace	21.56	24.16	32.98	29.02
Pleasure	45.46	42.12	52.50	49.43
Sadness	19.66	23.06	18.30	***41.19***
Sensitivity	9.28	7.25	9.90	***15.67***
Suffering	18.84	21.91	17.91	***40.09***
Surprise	18.81	7.99	12.54	9.03
Sympathy	14.71	11.96	20.30	15.72
Yearning	8.34	8.13	12.52	10.30
mAP (%)	27.38	26.48	31.36	***32.00***

challenges within the dataset itself.

Overall, the comparison demonstrates that the systematic optimisation strategy employed in this research successfully yielded a model that not only achieves a competitive overall mAP but also exhibits superior performance on a significant subset of emotion categories compared to baseline and even knowledge-enhanced methods, validating the project's approach and success.

5.9 Chapter Summary

This chapter detailed the systematic five-stage experimental process undertaken to evaluate and optimise the multi-modal framework for emotion classification on the EMOTIC dataset. By sequentially evaluating backbone architectures, fusion methods, loss functions, image augmentation strategies, and dataset resampling techniques, an optimal configuration was identified. This configuration, utilising Swin Transformers for context and body streams, EfficientNet-B7 for the face stream, Weighted Fusion, Custom Discrete Loss, basic image augmentation, and ADASYN resampling, achieved a final mean Average Precision (mAP) of **32.00%** on the test set.

The experimental results demonstrated the benefits of using Transformer backbones for context and body features, the effectiveness of simple yet adaptive fusion like Weighted Fusion, the importance of selecting an appropriate loss function (with the

custom Discrete Loss performing best here), and the positive impact of both basic image augmentation and adaptive feature-space resampling (ADASYN) in improving the primary mAP metric. The achieved mAP score represents a significant accomplishment, validating the effectiveness of the systematically optimised three-stream visual approach. It surpasses several established benchmarks, including recent knowledge-enhanced methods that rely on external information sources (Chen et al., 2023).

This work successfully demonstrates that substantial performance on the challenging EMOTIC benchmark can be attained by focusing purely on maximising the information gleaned from the fundamental context, body, and face visual streams through careful optimisation of standard deep learning components. While state-of-the-art methods achieve higher scores through greater complexity (e.g., incorporating depth maps (Mittal et al., 2020)), the competitiveness of the 32.00% mAP obtained here underscores the value and success of the systematic optimisation strategy employed within this project. Furthermore, the per-emotion analysis revealed particular strengths of this optimised visual framework, achieving superior performance on nearly 40% of the emotion categories compared to baseline and knowledge-enhanced methods. Overall, the staged methodology proved effective in navigating the numerous design choices and arriving at a well-performing and efficiently designed configuration based on empirical evidence gathered within the project's scope.

6 Conclusions and Future Work

This dissertation embarked on the complex challenge of context-aware, multi-label emotion recognition, specifically utilising the EMOTIC dataset. The primary aim, as stated in Section 1, was to design, implement, and comprehensively assess a robust multi-modal deep learning framework capable of effectively integrating visual information from context, body, and face streams to predict apparent discrete emotions, while simultaneously addressing the inherent difficulties posed by the dataset, such as class imbalance and potential label noise. Through the systematic methodology and staged experimental process detailed in Sections 4 and 5, this aim has been successfully fulfilled. All five core objectives outlined in Section 1 – investigating backbones, exploring fusion strategies, evaluating loss functions, assessing augmentation techniques, and evaluating overall performance – were systematically addressed, leading to the development and validation of an optimised framework.

6.1 Summary of Findings and Contributions

The research yielded several key findings and contributions, summarised below:

6.1.1 Optimal Configuration Identification

Findings:

- A systematic five-stage experimental evaluation identified an optimal configuration for the multi-modal framework.
- This configuration consists of Swin Transformer backbones for context and body streams, an EfficientNet-B7 backbone for the face stream, a Weighted Fusion mechanism, a custom dynamically weighted Discrete Loss function, basic image-level augmentation (Random Horizontal Flip, Colour Jitter), and ADASYN feature-space resampling.

Contribution:

- Provided empirical validation for specific architectural and training choices within a three-stream visual framework for the EMOTIC dataset.
- Established a well-performing baseline configuration derived through methodical optimisation, demonstrating the effectiveness of combining specific Transformer and CNN backbones with adaptive fusion and targeted imbalance mitigation.

6.1.2 Performance Evaluation

Findings:

- The optimised framework achieved a mean Average Precision (mAP) of 32.00% on the held-out EMOTIC test set.
- This performance surpasses the original EMOTIC baseline (Kosti et al., 2017) and several recent methods, including knowledge-enhanced approaches like TEKG (Chen et al., 2023).
- The framework demonstrated state-of-the-art performance among compared methods on nearly 40% of the individual emotion categories, showing particular strength in recognising negative affects (e.g., Sadness, Suffering, Anger, Annoyance) and Happiness.

Contribution:

- Demonstrated that a competitive level of performance on the challenging EMOTIC benchmark can be achieved by systematically optimising a standard three-stream visual architecture using established deep learning techniques, without reliance on external data or highly complex modules.
- Validated the success of the project's approach by achieving results comparable to, and in some aspects superior to, more complex methods involving knowledge graphs or simple VLM probing (Chen et al., 2023; Etesam et al., 2024).
- Highlighted the specific emotional categories where the optimised visual approach excels.

6.1.3 Component Analysis Insights

Findings:

- Transformer architectures (specifically Swin) proved highly suitable for processing context and body streams, likely due to their ability to capture global dependencies.
- EfficientNet-B7 was effective for extracting features from the more localised face stream.
- Simple, adaptive fusion mechanisms like Weighted Fusion outperformed several more complex strategies (e.g., Cross-Modal Transformer, Low-Rank Bilinear) within the experimental constraints.
- The custom Discrete Loss with dynamic weighting offered a slight advantage over standard BCE and other tested imbalance-aware losses (Focal, ASL) in terms of mAP. Dice Loss improved recall but severely degraded mAP.
- Basic image augmentation (flips, jitter) provided a consistent, modest performance improvement, while more complex augmentations showed diminishing returns.
- Dataset-level augmentation using ADASYN yielded a further marginal increase in mAP, though it negatively impacted precision and recall metrics.

Contribution:

- Provided specific, data-driven insights into the relative effectiveness of different backbones, fusion methods, loss functions, and augmentation techniques for the

task of multi-modal emotion recognition on EMOTIC.
- Offered guidance for future research by highlighting effective component choices (e.g., Swin for context/body, weighted fusion) and identifying areas where standard complex approaches might underperform without extensive tuning or longer training (e.g., cross-modal transformers).
- Underscored the persistent challenge of class imbalance and the trade-offs involved in mitigation strategies (e.g., ADASYN improving mAP vs. recall/precision).

6.1.4 Methodological Contribution

Findings:

- A systematic, staged experimental design was successfully employed to navigate the numerous design choices involved in building the multi-modal system.
- Each stage built upon the optimised configuration from the previous one, allowing for methodical comparison and selection of components.

Contribution:

- Demonstrated a structured and effective methodology for optimising complex deep learning frameworks by isolating and sequentially evaluating key components.
- Provided a clear and traceable path for identifying the final optimal configuration based on empirical evidence gathered at each stage.

6.2 Critical Analysis and Limitations

Despite the successful fulfilment of the project's aim and objectives and the competitive performance achieved, a critical analysis reveals several limitations inherent to the study and the broader challenge:

- **Dataset Challenges:** The EMOTIC dataset's inherent characteristics, such as significant class imbalance (Audibert et al., 2024) and potential label noise due to subjective annotation (Costa et al., 2023), fundamentally limit achievable performance and affect the reliability of evaluation. While mitigation strategies (weighted loss, ADASYN) were employed, they did not fully resolve the imbalance issue, as evidenced by persistently low recall scores for many methods. The static nature of the images also precludes the use of temporal dynamics (Wang et al., 2024).
- **Methodological Constraints:**
  - *Training Duration:* The fixed 20-epoch training schedule, necessitated by computational constraints (8GB VRAM), might have been insufficient for complex models or augmentation strategies (e.g., Cross-Modal Transformer, 'Full' augmentation) to reach optimal convergence. Longer training could potentially unlock better performance for some configurations.

- *Hyperparameter Tuning:* Extensive hyperparameter tuning for loss functions (e.g., gamma in Focal/ASL, thresholds in MLLSC) or resampling methods (k in ADASYN) was outside the scope of the staged comparison, potentially impacting their relative performance observed in Stage 3 and Stage 5.
- *Ablation Study Scope:* While the staged approach evaluated components sequentially, a detailed ablation study quantifying the precise contribution of each modality (context, body, face) within the final optimal configuration was not explicitly performed, limiting definitive conclusions about their individual importance in the final model.

- **Scope:** The research focused solely on the visual modalities provided by the EMOTIC dataset. Real-world emotion perception often involves integrating information from other channels, such as auditory cues (prosody, speech content) or physiological signals, which were not considered in this work.

- **Performance Gaps:** While achieving 32.00% mAP is a strong result demonstrating the success of the optimisation strategy, it falls short of the absolute state-of-the-art results (e.g., 35.5% by EmotiCon depth-based variant (Mittal et al., 2020)). This suggests that advanced techniques incorporating richer contextual reasoning (e.g., explicit modelling of social dynamics, scene semantics using GCNs or depth maps) or leveraging the power of very large pre-trained multimodal models more deeply offer further advantages. The per-emotion analysis also revealed difficulties in recognising certain nuanced social emotions compared to more basic ones, indicating limitations in the model's semantic understanding compared to more complex approaches. The trade-off observed with ADASYN (improving mAP but reducing precision/recall) also highlights the complexity of optimising for multiple objectives simultaneously.

6.3 Future Work

Building upon the findings and limitations of this research, several promising avenues for future work emerge:

- **Advanced Fusion and Modality Handling:** Explore more sophisticated fusion architectures, potentially revisiting attention or transformer-based methods with longer training durations or targeted hyperparameter tuning. Investigate techniques explicitly designed to handle noisy or missing modalities (e.g., robust fusion networks that can adapt when a stream like 'face' is occluded, modality imputation).

- **Enhanced Imbalance and Noise Handling:** Experiment with more advanced class imbalance techniques beyond ADASYN, such as distribution-aware losses (e.g., LDAM Loss), ensemble methods focusing on minority classes, or curriculum learning approaches that present easier examples first. Implement and evaluate methods specifically designed for learning with noisy labels (LNL) in multi-label settings, potentially combining them with robust loss functions.

- **Incorporating Temporal Dynamics:** Extend the framework to process video sequences instead of static images. This would allow the model to leverage crucial dynamic cues from facial expressions, body movements, and evolving context, which are known to be significant for human emotion perception (Richoz et al., 2018) and could lead to substantial performance improvements.
- **Leveraging Large Vision-Language Models (VLMs):** Move beyond simple linear probing of VLM features (as in EmotionCLIP (Etesam et al., 2024)). Explore fine-tuning state-of-the-art VLMs on the EMOTIC task or integrating their semantic understanding capabilities more deeply into the fusion process, perhaps by using VLM embeddings as additional input streams or employing cross-modal attention between visual features and VLM-generated contextual descriptions.
- **Explainability and Interpretability:** Implement explainability techniques (e.g., attention map visualisation like Grad-CAM for CNNs/Transformers, feature attribution methods like SHAP) to gain deeper insights into which visual cues (specific regions in context, body, or face) and modalities the model relies on for its predictions across different emotions. This would aid model debugging, enhance trustworthiness, and provide valuable insights into the model's reasoning process.
- **Cross-Dataset Generalisation:** Evaluate the robustness and generalisability of the developed framework by testing its performance on other context-aware emotion recognition datasets (e.g., CAER-S (Lee et al., 2019)) without retraining, or by exploring domain adaptation techniques.
- **Extended Training and Optimisation:** Conduct experiments with significantly longer training durations (e.g., 50-100 epochs) to ensure full convergence, particularly for more complex configurations. Perform more systematic hyperparameter optimisation for critical components like loss function parameters (e.g., γ in Focal/ASL) and augmentation strategies using techniques like grid search or Bayesian optimisation on the validation set.

6.4 Concluding Remarks

This dissertation successfully addressed the aim of developing and evaluating a multimodal deep learning framework for context-aware emotion recognition using the EMOTIC dataset. Through a rigorous and systematic staged experimental process, an optimised configuration was identified, achieving a competitive mAP of 32.00%. This result validates the effectiveness of the chosen approach, demonstrating that significant performance can be achieved by carefully optimising standard deep learning components such as backbones, fusion, loss, and augmentation when applied to the fundamental visual streams of context, body, and face.

The research highlights the strengths of Transformer architectures for contextual and body analysis, the utility of adaptive fusion and loss weighting, and the benefits of combining image-level and dataset-level augmentation. Notably, the framework surpasses

several established benchmarks, including knowledge-enhanced methods, showcasing the power of focused optimisation within a constrained visual setting. While acknowledging the inherent challenges of the dataset and the superior performance of highly complex state-of-the-art methods that incorporate external data or modalities, this work stands as a successful demonstration of achieving robust and competitive results through methodical design and evaluation. The findings provide valuable insights for the field and lay a solid foundation for future research directions aimed at further advancing the capabilities of machines to understand human emotions in their rich, natural context.

References

Akhand, M., Roy, S., Siddique, N., Kamal, M. A. S., and Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep cnn. *Electronics*, 10(9):1036.

Alayón, S., Hernández, J., Fumero, F. J., Sigut, J. F., and Díaz-Alemán, T. (2023). Comparison of the performance of convolutional neural networks and vision transformer-based systems for automated glaucoma detection with eye fundus images. *Applied Sciences*, 13(23):12722.

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., and Khan, F. S. (2023). Transformers in remote sensing: A survey. *Remote Sensing*, 15(7):1860.

Arevalo, J., Solorio, T., Montes-y Gómez, M., and González, F. A. (2017). Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Audibert, A., Gauffre, A., and Amini, M.-R. (2024). Multi-label contrastive learning: A comprehensive study. *arXiv preprint arXiv:2412.00101*.

Aviezer, H., Bentin, S., Dudarev, V., and Hassin, R. R. (2011). The automaticity of emotional face-context integration. *Emotion*, 11(6):1406.

Calvo, R. A. and Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, J., Yang, T., Huang, Z., Wang, K., Liu, M., and Lyu, C. (2023). Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*, 53(4):4201–4217.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., and Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712.

Cheng, S. and Zhou, G. (2020). Facial expression recognition method based on improved vgg convolutional neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(07):2056003.

Chu, J., Cai, J., Li, L., Fan, Y., and Su, B. (2021). Bilinear feature fusion convolutional neural network for distributed tactile pressure recognition and understanding via visualization. *IEEE Transactions on Industrial Electronics*, 69(6):6391–6400.

Costa, W., Talavera, E., Oliveira, R., Figueiredo, L., Teixeira, J. M., Lima, J. P., and Teichrieb, V. (2023). A survey on datasets for emotion recognition from vision: Limitations and in-the-wild applicability. *Applied Sciences*, 13(9):5697.

de Lima Costa, W., Talavera, E., Figueiredo, L. S., and Teichrieb, V. (2023). High-level context representation for emotion recognition in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 326–334.

Elharrouss, O., Akbari, Y., Almadeed, N., and Al-Maadeed, S. (2024). Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. *Computer Science Review*, 53:100645.

Etesam, Y., Yalçın, Ö. N., Zhang, C., and Lim, A. (2024). Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4769–4776. IEEE.

Fathalla, R. (2020). Emotional models: Types and applications. *International Journal of Synthetic Emotions (IJSE)*, 11(2):1–18.

Feng, K. and Chaspari, T. (2020). A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science*, 2:9.

Franzoni, V., Biondi, G., Perri, D., and Gervasi, O. (2020). Enhancing mouth-based emotion recognition using transfer learning. *Sensors*, 20(18):5222.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Ghiassi, A., Birke, R., and Chen, L. Y. (2023). Multi label loss correction against missing and corrupted labels. In *Asian Conference on Machine Learning*, pages 359–374. PMLR.

Guo, Y., Ge, H., and Li, J. (2023). A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Frontiers in Computer Science*, 5:1159063.

Halim, A. M., Dwifebri, M., and Nhita, F. (2023). Handling imbalanced data sets using smote and adasyn to improve classification performance of ecoli data sets. *Building of Informatics, Technology and Science (BITS)*, 5(1):246–253.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., and Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241:122666.

Jiao, T., Guo, C., Feng, X., Chen, Y., and Song, J. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, 80(1).

K, M. P., R, D., R, G., Shasteeswaran, S., Tharmiya, R., and S.K, R. (2024). Emotion detection in facial expressions and speech using deep hybrid learning. *International Journal of Research Publication and Reviews*.

Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., and Farooq, U. (2023). A survey of the vision transformers and their cnn-transformer based variants. *Artificial Intelligence Review*, 56(Suppl 3):2917–2970.

Kim, H. E., Maros, M. E., Miethke, T., Kittel, M., Siegel, F., and Ganslandt, T. (2023). Lightweight visual transformers outperform convolutional neural networks for gram-stained image classification: An empirical study. *Biomedicines*, 11(5):1333.

Kopalidis, T., Solachidis, V., Vretos, N., and Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, 15(3):135.

Korsmit, I. R., Montrey, M., Wong-Min, A. Y. T., and McAdams, S. (2023). A comparison of dimensional and discrete models for the representation of perceived and induced affect in response to short musical sounds. *Frontiers in Psychology*, 14:1287334.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). Emotic: Emotions in context dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 61–69.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766.

Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152.

Li, S. and Tang, H. (2024). Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

Limami, F., Hdioud, B., and Oulad Haj Thami, R. (2024). Contextual emotion detection in images using deep learning. *Frontiers in Artificial Intelligence*, 7:1386753.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., and Zheng, W. (2023). The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, 9:e1400.

Majumder, A., Dutta, S., Kumar, S., and Behera, L. (2020). A method for handling multi-class imbalanced data by geometry based information sampling and class prioritized synthetic data generation (gicaps). *arXiv preprint arXiv:2010.05155*.

Marsella, S., Gratch, J., Petta, P., et al. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46.

Matsuda, Y.-T., Fujimura, T., Katahira, K., Okada, M., Ueno, K., Cheng, K., and Okanoya, K. (2013). The implicit processing of categorical and dimensional strategies: an fmri study of facial emotion perception. *Frontiers in human neuroscience*, 7:551.

Mensink, T. and Mettes, P. (2023). Infinite class mixup. *arXiv preprint arXiv:2305.10293*.

Min, S., Yang, J., and Lim, S. (2024). Emotion recognition using transformers with random masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4860–4865.

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14234–14243.

Mobbs, R., Makris, D., and Argyriou, V. (2025). Emotion recognition and generation: A comprehensive review of face, speech, and text modalities. *arXiv preprint arXiv:2502.06803*.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213.

Pawłowski, M., Wróblewska, A., and Sysko-Romańczuk, S. (2023). Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5):2381.

Peng, Y., Wang, D. Z., Patwa, I., Gong, D., and Fang, C. V. (2015). Probabilistic ensemble fusion for multimodal word sense disambiguation. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 172–177. IEEE.

Pereira, R., Mendes, C., Ribeiro, J., Ribeiro, R., Miragaia, R., Rodrigues, N., Costa, N., and Pereira, A. (2024). Systematic review of emotion detection with computer vision and deep learning. *Sensors*, 24(11):3484.

Praveen, R. G. and Alam, J. (2024). Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4803–4813.

Richoz, A.-R., Lao, J., Pascalis, O., and Caldara, R. (2018). Tracking the recognition of static and dynamic facial expressions of emotion across the life span. *Journal of vision*, 18(9):5–5.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. (2021). Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91.

Shantharam, R. M. and Schwenker, F. (2024). Ml-based pain recognition model using mixup data augmentation. *Applied System Innovation*, 7(6):124.

Shen, Z. (2024). A comparative study of hybrid cnn and vision transformer models for facial emotion recognition. In *2024 11th International Conference on Dependable Systems and Their Applications (DSA)*, pages 401–408. IEEE.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., and Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., et al. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *Journal of Medical Systems*, 48(1):84.

Tarekegn, A. N., Giacobini, M., and Michalak, K. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Terven, J., Córdova-Esparza, D.-M., and Romero-González, J.-A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction*, 5(4):1680–1716.

Turkstra, L. S., Hosseini-Moghaddam, S., Wohltjen, S., Nurre, S. V., Mutlu, B., and Duff, M. C. (2023). Facial affect recognition in context in adults with and without tbi. *Frontiers in Psychology*, 14:1111686.

Urnisha, N. N., Bithi, S. I., Rafee, M. M. S., Remon, N. I., Hasan, M. M., and Chowdhury, P. (2024). A transfer learning approach for facial emotion recognition using a deep learning model. *International journal of research and scientific innovation*, 11(4):274–284.

Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., Li, C., and Quan, D. (2023). Multimodal transformer augmented fusion for speech emotion recognition. *Frontiers in Neurorobotics*, 17:1181598.

Wang, Y., Yan, S., Liu, Y., Song, W., Liu, J., Chang, Y., Mai, X., Hu, X., Zhang, W., and Gan, Z. (2024). A survey on facial expression recognition of static and dynamic emotions. *arXiv preprint arXiv:2408.15777*.

Winterbottom, T., Xiao, S., McLean, A., and Al Moubayed, N. (2022). Bilinear pooling in video-qa: empirical challenges and motivational drift from neurological parallels. *PeerJ Computer Science*, 8:e974.

Yang, D., Yang, K., Li, M., Wang, S., Wang, S., and Zhang, L. (2024). Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12447–12457.

Yasuda, Y., Miyazaki, T., and Goto, J. (2024). Weighted asymmetric loss for multi-label text classification on imbalanced data. *Journal of Natural Language Processing*, 31(3):1166–1192.

Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B., and Yang, G. (2023). Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging*, 36(2):739–752.

Yu, J., Liu, Y., Fan, R., and Sun, G. (2024). Mixcut: A data augmentation method for facial expression recognition. *arXiv preprint arXiv:2405.10489*.

Zeng, Z., Liu, H., Chen, F., and Tan, X. (2023). Compensated attention feature fusion and hierarchical multiplication decoder network for rgb-d salient object detection. *Remote Sensing*, 15(9):2393.

Zhang, M.-L. and Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Zhang, Y., Ding, W., Xu, R., Hu, X., and De Raedt, L. (2022). Visual emotion representation learning via emotion-aware pre-training. In *IJCAI*, pages 1679–1685.