POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Attention mechanism and Visual Transformer models for Facial Emotion Recognition in natural settings

Tesi di Laurea Magistrale in
Computer Science and Engineering

Author: **Giacomo Da Re**

Student ID: 976587
Advisor: Prof. Andrea Bonarini
Academic Year: 2022-2023

Abstract

Emotions are the most immediate and natural form of communication among human beings. The joy of an infant smiling at their parents, the fear of a child in the dark night, the aversion to a disliked food – these are languages common to all peoples across all ages. While cultural differences might affect their expression, the truth of emotions lies in their spontaneity.

To fully embrace this concept, the following work presents a neural network model that classifies emotions based on facial images, favoring a natural context. The proposed architecture relies on the latest technology in computer vision, namely the attention mechanism, which allows the model to focus on specific parts of the input to maximize the chances of correct classification. This cutting-edge technology has been explored and tested in various foundational models, leading to the development of a dual-channel solution that analyzes the image in two distinct ways: one extracting general features from the input image and the other focusing on fixed points of the face for an even more stable and precise classification.

The field of facial emotion recognition is set to become increasingly significant in various domains, ranging from human-computer interaction to potential applications in video games or the development of smart cities. This work aims to pave the way for the development of modern and high-performing models.

Keywords: Facial Emotion Recogniton, Attention mechanism, Visual Transformer, In the wild context, AffectNet

Abstract in lingua italiana

Le emozioni sono il primo e più spontaneo veicolo di comunicazione tra esseri umani. La gioia di un neonato che sorride ai propri genitori, la paura di un bambino per il buio della notte, il disgusto di un alimento che non apprezziamo, sono linguaggi comuni a tutti i popoli in ogni tempo. Per quanto ci possano essere differenze culturali nell'esprimerle, ciò che le rende vere è la loro spontaneità.

Per seguire in toto quest'ultimo punto, il seguente lavoro presenta un modello di rete neurale che, classificando in base a immagini di volti l'emozione che il soggetto prova in quell'istante, prediliga il contesto naturale. L'architettura proposta si appoggia alla più moderna tecnologia disponibile nell'ambito del computer vision, ovvero il meccanismo di attenzione; esso permette di concentrarsi su parti specifiche dell'input da utilizzare per massimizzare le probabilità di una corretta classificazione. Questa moderna tecnologia è stata esplorata e provata in numerosi modelli di base, fino ad arrivare alla realizzazione di una soluzione a due canali che permetta di analizzare l'immagine in due modi distinti: il primo andando a estrarre caratteristiche generali dell'immagine in input, il secondo concentrandosi su punti fissi del volto per una classificazione ancora più stabile e puntuale.

Il campo del riconoscimento delle emozioni facciali sarà sempre più importante in vari ambiti, a partire dall'interazione uomo-computer, fino alla possibilità di utilizzarlo nei videogiochi o nella realizzazione di città intelligenti. Il seguente lavoro si propone di aprire una strada allo sviluppo di modelli moderni e prestanti.

Parole chiave: Riconoscimento delle emozioni facciali, Meccanismo di attenzione, Visual Transformer, Contesto naturale, AffectNet

Contents

1 | Introduction

In an era where robots and artificial intelligence (AI) are progressively intertwining with human experience, the exigency for machines to comprehend and respond to human emotional states is markedly significant [53]. One of the most universal and significant methods that people communicate their emotions and intentions is through the medium of their facial expressions. The idea of a future in which interactions with robots mirror the natural and intuitive dynamics of human-human interactions is gradually becoming a tangible reality. The basis for the realization of this vision revolves around the axis of emotion recognition, a specialized but booming subfield of AI, loaded with the potential to dramatically increase human-machine interactions.

Many modern intelligent systems implement facial analytics in images and videos [45], delving into aspects such as age, gender, and ethnicity. This underscores the comprehensive nature and importance of facial understanding beyond just emotional states. The scope of emotion recognition transcends merely enhanced interaction quality (although it was conceived as one of the first purposes of its use [38]), extending its tendrils across diverse sectors such as healthcare, education, entertainment, and security, thus underscoring its seminal role in the progressive evolution of AI.

Nonetheless, the journey towards mastering adept emotion recognition capabilities remains arduous. Particularly in uncontrolled, real-world scenarios, or contexts "in the wild," traditional computer vision models often grapple with capturing the nuanced expressions and understated emotional cues intrinsic to human communication. The fluctuating lighting conditions, diverse facial orientations, and myriad expression variations further convolute the emotion recognition landscape. Although the dawn of deep learning has infused a propelling momentum within the field, the quest for models adept at accurately and reliably deciphering emotions in such heterogenous and unstructured settings remains unquenched. This glaring void accentuates the imperative for innovative methodologies capable of unraveling the complexity and subtlety embedded within human emotions.

This study endeavors to bridge this chasm by leveraging the capabilities of Visual Trans-

formers, a nascent architectural paradigm that orchestrates attention mechanisms to parse images. Distinct from their convolutional counterparts, Visual Transformers possess the ability to focus on disparate segments of an image and decipher the contextual relationships interlinking them, thereby heralding promise for superior emotion recognition, especially in challenging "wild" environments.

By training various models and scrutinizing its performance through various metrics, we aspire to distill insights into the effectiveness of Visual Transformers within the domain of emotion recognition in the wild. Moreover, the exploration of attention mechanisms within our models unveils a portal into comprehending how machines can be schooled to "pay attention" to pertinent cues whilst interpreting human emotions amidst a myriad of distractors inherent in real-world settings.

As we navigate the trajectory towards engendering emotionally intelligent machines, the revelations of this study not only enrich the burgeoning corpus of knowledge in emotion recognition but also incite further ventures into innovative architectural designs aimed at refining human-machine interaction. This study strives to edge closer towards a future where machines can intuitively fathom and harmonize with the emotional tapestry of human interactions, thereby nurturing a more harmonious human-machine symbiosis in the wild.

1.1. Structure of the thesis

A summary description of the structure of the thesis and its chapters follows.

Chapter 2 delves into the foundational concepts critical to understanding the interplay between human emotions and their computational recognition. In Section 2.1, "Understanding Emotion," the thesis unpacks the layered definitions and theories surrounding emotions. Section 2.2, "The Spectrum of Human Emotions," explore models that classify emotions into discrete categories and those that view emotions across a continuum. Section 2.3 provides an overview of emotion recognition technologies, with a particular emphasis on facial emotion recognition and the challenges of recognizing emotions 'in the wild'. The chapter concludes with discussions on the practical applications of these technologies and their social and ethical implications, respectively in Section 2.4 and 2.5, setting the stage for the application of these theoretical underpinnings in practical computational systems.

Chapter 3, "Attention mechanism in Facial Expression Recognition," delves into cutting-edge deep learning methods for detecting emotions from faces. It starts with "Vision

Transformers: revolutionizing Computer Vision," in Section 3.1, discussing how these models use attention mechanisms to focus on important features in facial recognition tasks. Section 3.1.2, "What are Vision Transformers (ViTs)?" introduces the ViT model, explaining its function and potential advantages in emotion recognition. The subsequent section, 3.1.3, "Advantages of Vision Transformers," discusses the strengths of ViTs over traditional models. Finally, Section 3.1.4, "ViT: Bridging Attention and FER," discusses how Vision Transformers can enhance the field of facial emotion recognition (FER) by providing nuanced analysis of facial expressions, potentially leading to more accurate emotion detection.

Chapter 4 delves into the "Methods and Materials" used in this research, starting with an extensive look at the "Datasets," particularly "AffectNet," in Section 4.1.1. The "Experimental Setup" is detailed in Section 4.2, covering the framework used, model training configurations, data adjustments, and sampling methods. Section 4.3 outlines the "Evaluation Metrics" that assess model performance, while Section 4.4's "Target Description" zeroes in on "Emotion Recognition on AffectNet." Proposed solutions are briefly discussed in Section 4.5. In section 4.6 and 4.7, after an accurate description of the two models, ViT and Swin ViT results are addressed. The chapter concludes with Section 4.8's "Two Stream Model" and its results, and a comparison of the model's performance in Section 4.9, offering a rounded view of the methodologies and their effectiveness.

Chapter 5 presents the "Conclusions" of the study, beginning with Section 5.1, "Contributions of the Thesis," which recaps the significant findings and the novel insights provided by the research. This section acknowledges the unique contributions made to the field of emotion recognition through the implementation of Visual Transformers. Following this, Section 5.2, "Limitations and Future Perspectives," offers a critical reflection on the study's limitations, discussing the challenges encountered during the research process and a perspective on the potential future directions for emotion recognition technology.

2 | Theoretical background

2.1. Understanding Emotion

Emotions are universally acknowledged as pervasive forces in human life, significantly influencing our decisions, behaviors, and interpersonal interactions. They are continually triggered by an array of stimuli and situations, molding our cognitive processes and behavioral responses [3, 18]. This notion is not only a fundamental premise in countless papers across various disciplines, including psychology, medicine, marketing, and management, but also a well-documented conclusion from extensive empirical studies.

However, the "ubiquity" of emotions, while widely accepted, is not thoroughly understood in the context of our everyday experiences. Past research has successfully outlined the causes and effects of emotions under controlled circumstances, yet our comprehension of emotional life in natural settings remains notably superficial. The methodologies that have shaped our current understanding are often limited by their scope, primarily employing narrow demographic samples and relying on retrospective self-reporting. These approaches are inherently fraught with limitations, including potential memory biases and the absence of real-time emotional tracking, thereby creating a significant gap in our understanding of authentic, everyday, emotional experiences.

In understanding the multi-dimensional nature of emotions, it is crucial to recognize that "Emotions are mental states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure. There is no scientific consensus on a definition. Emotions are often intertwined with mood, temperament, personality, disposition, or creativity." [58]. This comprehensive view highlights the intricate interplay of various internal and external factors that contribute to our emotional states.

Furthermore, a crucial aspect that current research methods often overlook is the concurrent experience of both negative and positive emotions. Studies indicate that individuals frequently report the simultaneous presence of these seemingly contradictory emotional states, adding a layer of complexity to our emotional landscape [26].

The recognition that emotions are integral to human existence marks the beginning of a deeper inquiry beyond laboratory settings, necessitating a sophisticated exploration into the real-world tapestry of emotional experiences. This involves understanding not just the frequency of various emotions, but also unraveling the complexities of their interrelationships, co-occurrences, and exclusivity within the human experience network. Such an exploration transcends academic interest, holding significant potential for practical applications across diverse fields like mental health, education, and commerce, thereby enriching our comprehension of the human psyche and potentially revolutionizing intervention strategies.

2.2. The Spectrum of Human Emotions

Throughout history, scholars and scientists have endeavored to understand the intricate mechanisms of human emotions. Various computational and theoretical models have emerged, each proposing unique interpretations and classifications of emotional processes. In general, there are two main categories of emotion models in affective computing, namely discrete emotion model and dimensional emotion model (or continuous emotion model).

2.2.1. Discrete Emotion Model

The concept of the discrete emotion model stands as a cornerstone in the field of affective computing and psychology, owing its prominence to the clarity and precision it provides in characterizing human emotions. This model, fundamental in the study and understanding of human affect, posits that there are a fixed number of basic, distinct emotions, often universal across the human experience. These primary emotions are characterized by unique expressions, physiological responses, and neural mechanisms, serving as the building blocks for the more complex, nuanced emotional states we experience. In the realm of emotion recognition and analysis, especially in technologically driven applications, the discrete emotion model proves particularly valuable. The model proposed in this work will use a training dataset [35] based on the discrete emotional model

Ekman's Discrete Model

Ekman's Discrete Model (figure 2.1 holds a central position, particularly with its postulation of universal emotions. This model, a seminal contribution by Paul Ekman, suggests that there are six basic emotions - anger, disgust, fear, happiness, sadness, and surprise - that are universally experienced by humans regardless of cultural background [15]. These emotions are not just universal, but also discrete, meaning they are distinct categories of

emotions, not just points along a continuous spectrum [47].

The development of Ekman's model was predicated on the hypothesis that human emotions are universal across races and cultures, a theory rooted in the idea of natural selection. This suggests that humans are biologically equipped with a set of emotional responses that have evolved over time due to their adaptive value in dealing with fundamental life tasks. These basic emotions prepare individuals to react quickly to stimuli, often before conscious thought is possible, indicating the presence of what Ekman calls "automatic appraisers" [16].

Ekman's research extended beyond the identification of these universal emotions, delving into the realm of facial expressions. He posited that these basic emotions are innately tied to specific facial expressions, providing a nonverbal language of emotion that's understood across cultural boundaries [16].

Further research has extended this model, proposing more than 20 basic emotions. These include states like amusement, contempt, contentment, embarrassment, excitement, guilt, pride, relief, satisfaction, pleasure, and shame [10].

Figure 2.1: Images used by Ekman for his experiments. Each one represents a different emotion.

Plutchik's Psychoevolutionary Theory

Plutchik's model, presented in [39] and further detailed in his wheel model (figure 2.2), posits the existence of eight primary bipolar emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. These emotions, fundamental to human experience, are deeply tied to evolutionary survival behaviors. Each emotion exists in relation to others, forming a complex interplay; for example, joy and sadness are diametrically opposed, while anticipation may intensify into vigilance.

This wheel model, also known as the componential model, visualizes these relationships and the different intensity of emotions. In this model, stronger emotions are represented at the center of the wheel, while their weaker counterparts lie at the periphery, indicating differing levels of emotional intensity. This arrangement not only illustrates the primary emotions but also allows for the derivation of more nuanced emotional states through their combinations and intensities.

Furthermore, Plutchik's theory underscores the adaptive function of emotions, emphasizing their role in evolutionary survival. It suggests that emotions have evolved to handle essential life tasks, guiding behaviors that enhance an organism's reproductive success. They serve as a functional medium, helping organisms, including humans, adapt to various environmental challenges.

In the context of sentiment analysis, these discrete emotions are often categorized into three polarities: positive, negative, and neutral, each serving a specific role in interpersonal communication and decision-making processes. The model's allowance for ambivalent sentiments, those containing elements of more than one basic emotion, is particularly notable. Techniques like ambivalent sentiment handling [55] have been developed to analyze these multi-level sentiments, enhancing the accuracy of binary classifications and providing a more detailed understanding of complex emotional states.

Plutchik's model advocates for an evolutionary perspective, suggesting that the primary emotions extend beyond the human experience, manifesting in different forms and serving adaptive functions across various species.

Tomkins' Theory of Nine Affects

One of the pivotal debates among theorists revolves around categorizing emotions into two main groups: basic emotions and complex emotions. Basic emotions, also referred to as primary or fundamental emotions, are often metaphorically compared to primary colors like blue, red, and yellow, whose blend spawns an array of other shades. In a similar

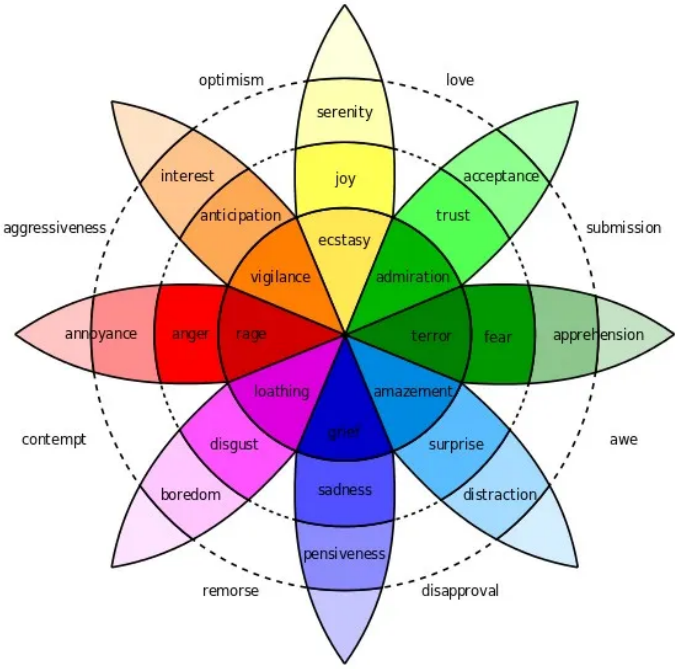

Figure 2.2: Plutichk's emotional wheel model

vein, complex emotions, also known as secondary emotions, are construed as derivations born from the intermingling of various basic emotions. While this bifurcation into fundamental and complex emotions is a widely held notion, consensus eludes the specifics of classification criteria, and consequently, the emotions to be classified under each category.

Silvan Tomkins was among the pioneers to suggest a biological basis for distinguishing primary emotions from secondary ones, a proposition he introduced as part of his affect theory [101, 102]. In Tomkins' framework, an 'affect' is deemed an innate biological response, an amplification of various stimuli, propelled by different intensities and neural firing patterns, which captivates attention and spurs action. Conversely, an 'emotion' is the conscious recognition of an affect, amalgamated with memories of analogous past experiences.

Underpinning this classification, Tomkins also formulated a behavioral blueprint for individuals, positing that optimal well-being is attained through the maximization of positive affects and the minimization of negative ones. Fundamental to this pursuit is the unabridged expression of these affects, facilitating their recognition and comprehension by others.

2.2.2. Dimensional emotion model

Russell's Circumplex Model of Affect

Contrary to the discrete emotion models, Russell's Circumplex Model of Affect proposes that emotions are not distinct entities, each supported by specific and independent neural structures and pathways. Instead, emotions are the cognitive interpretations of core neural sensations that are the product of two independent neurophysiological systems [40].

In his seminal work, Russell [44] introduced the concept of a continuous, bi-dimensional emotion space model, a significant departure from the categorical approach. This model, known for its circumplex structure, employs two axes: Valence, or the degree of pleasantness or unpleasantness, and Arousal, or the degree of activation or deactivation. These axes create a Cartesian coordinate system where each emotion can be represented as a point defined by its valence and arousal coordinates.

One of the key studies underpinning this model involved participants sorting 28 emotion-laden words based on perceived similarity. Russell then employed a statistical technique to group these ratings, finding that emotions clustered in a circle, indicating more of a spectrum than distinct categories [44]. This spectrum was characterized by two bipolar dimensions: valence (ranging from unpleasant to pleasant) and arousal (varying from low to high). These dimensions are independent and bipolar, meaning that valence and arousal are uncorrelated and each encompasses a range of emotions [44].

The model is visually represented by four quadrants (as reported in figure 2.3, each corresponding to a different combination of valence and arousal. For instance, emotions in the first quadrant are characterized by high arousal and positive valence (e.g., happiness), while those in the third quadrant are associated with low arousal and negative valence (e.g., sadness). The second quadrant contains emotions having high arousal and negative valence (e.g., anger), and the fourth quadrant includes emotions with low arousal and positive valence (e.g., calmness) [11].

Russell's model emphasizes that emotional experiences known as "mixed emotions" cannot be extreme opposites in terms of valence or arousal, as they are located close to each other in the same quadrant. This aspect of the model suggests a complexity in human emotions, where they are not just discrete categories but can be blends of different feelings.

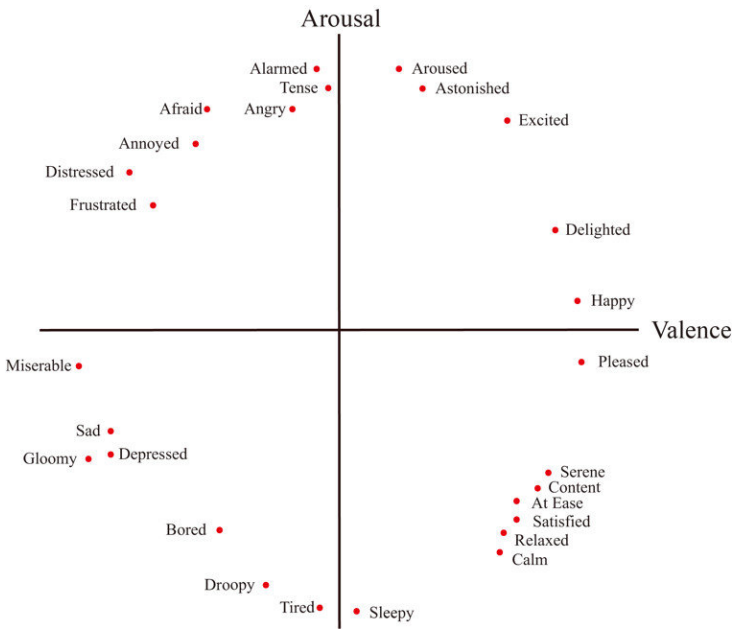

Figure 2.3: Russell's circumplex model.

2.3. Emotion recognition: an overview

The challenge and allure of emotion recognition lie in its intricacy. Various modalities can be employed, from facial expressions and body postures to voice tonality. Facial expressions are among the most immediate and universal forms of non-verbal communication. Within moments, a slight eyebrow raise, a shift in mouth angle, or a piercing gaze can convey emotions ranging from joy to sorrow and surprise to anger. Humans have long relied on these expressions to swiftly gauge another's emotional state.

However, they don't always tell the full emotional story. Physiological signals like EEG, EMG, and GSR offer deeper insights, revealing subconscious emotions untouched by conscious control [46]. So, relying solely on facial expressions to decode emotions can be misleading. Barrett et al. [1] assert that faces aren't straightforward emotional barcodes. They believe that expressions, rather than being explicit emotional signals, are context-dependent and require a more comprehensive interpretation, integrating elements like the surrounding situation, body posture, and tone of voice. Further expanding this perspective, Witkower et al. [59] delves into body movements as crucial conduits of emotional communication. They posit that gestures, gait, and posture may hold keys to understanding feelings as much as facial cues, if not more. In essence, while faces are vital emotional indicators, a holistic understanding of emotion necessitates a broader scope, encompassing both physiological signs and body language.

Regardless, the focus of this study is to grapple with the nuanced task of Facial Emotion Recognition (FER) as it unfolds within the unpredictability of everyday life, something that has proven to be inherently complex. Despite considerable efforts, the realm of FER has not fully matured, with current methodologies still grappling with the subtleties and variability inherent in spontaneous facial expressions—far removed from the uniformity and control of laboratory conditions. This research is an attempt to navigate these complexities by exploring novel approaches to FER, thereby contributing to the body of knowledge in a meaningful way. It acknowledges the foundational work of scholars such as Barrett and Picard, who have pointed out the multifaceted nature of emotion recognition. By focusing exclusively on FER, this investigation delves into a broad spectrum of theoretical and practical insights, with the aspiration of elevating the capabilities of current technologies. The ultimate goal is to enhance the sophistication with which we interpret emotional cues, thereby advancing the reliability and accuracy of such recognitions in the varied tapestry of real-world settings.

2.3.1. Facial emotion recognition

The journey of facial emotional recognition (FER) spans various fields, from psychology and neuroscience to technology and artificial intelligence. Its roots can be traced back to Darwin's early work on facial expressions, which he believed were universal across all cultures. Over time, with advancements in technology and understanding of the human mind, methodologies evolved significantly.

As previosly said, Paul Ekman's groundbreaking work in the late 20th century focused on the universality of certain emotions and their corresponding facial expressions. He proposed that basic emotions are innate and consistently expressed and recognized across diverse cultures. Through his research, particularly with diverse societies, he found consistent facial expressions that transcended cultural boundaries, challenging then-prevailing beliefs that such expressions were culturally learned.

In the late 20th century, Wallace V. Friesen and Paul Ekman laid the groundwork by identifying universal emotions and their corresponding facial expressions, leading to the Facial Action Coding System (FACS) [17] which became a cornerstone in emotion recognition research. It is based on years of psychological investigation and experimentation and still the most widely used as the robust method for describing facial behaviors. Its use has spread outside of the psychological and clinical fields. FACS consists of a set of 44 visually discriminable independent Action Units (AUs), reported in figure 2.4. Expressions can be described by combining multiple AUs. Each expression is given a score that

is made up of the list of AUs.

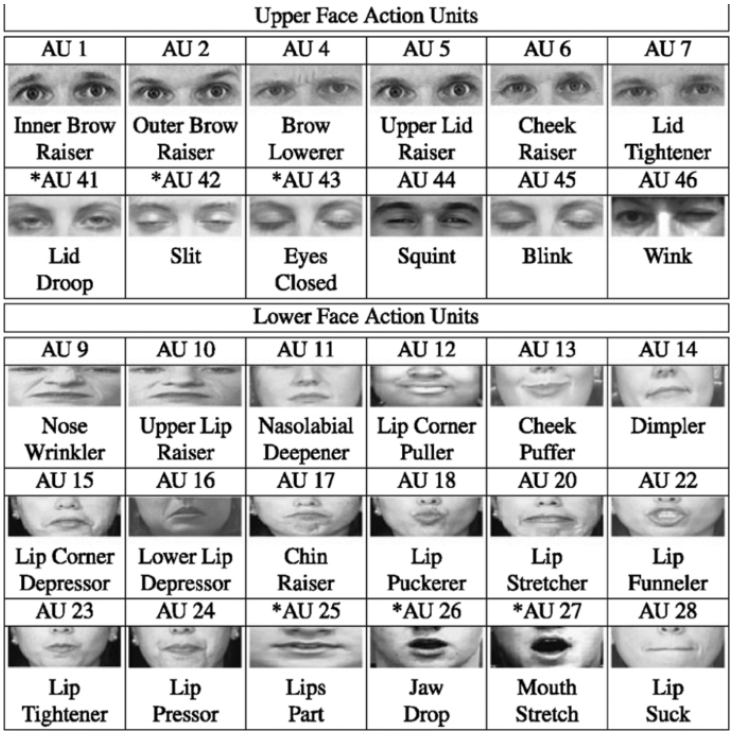

Figure 2.4: List of action units.

Let's explore a use case of FACS. In figure 2.5, the Action Units (AUs) associated with the emotions of disgust and fear are described.

For the emotion "disgust", the combination of AU9 and AU10 encodes its expression. Similarly, the combination of AU1, AU2, AU5, AU7, and AU20 encodes the emotion 'fear'

Figure 2.5: Detected AU for disgust and fear.

The model proposed in this paper will deal with the detection of 7 emotions plus the neutral one. In table 2.1 they will be listed and, for each of them, the corresponding coding system will be given following the action units proposed by the scientific community.

Emotion	Action Units
Happiness	$6 + 12$
Sadness	$1 + 4 + 15$
Surprise	$1 + 2 + 5B + 26$
Fear	$1 + 2 + 4 + 5 + 7 + 20 + 26$
Anger	$4 + 5 + 7 + 23$
Disgust	$9 + 15 + 17$
Contempt	$R12A + R14A$

Table 2.1: Listed AU for 7 basic emotions.

Instead of predetermining facial cues, the model's attention system is designed to autonomously identify pertinent clues within each image. This adaptive process seeks to pinpoint the relevant indicators required to predict the accurate emotion, harnessing the power of attentional networks to discern without imposing predefined facial action constraints. The goal is to allow the model to capture emotional subtleties that might be overlooked or underestimated in traditional action unit detection methods.

2.3.2. In the wild context

Emotion detection, while advanced in controlled environments, encounters a myriad of challenges when applied to "in the wild" settings. The controlled environment often provides a standardized backdrop where emotions are elicited under specific conditions, allowing for more predictable and clear emotional readings. In stark contrast, real-world, or "in the wild" scenarios, present spontaneous and nuanced emotional expressions. These expressions may deviate from textbook definitions, making them harder to identify and classify.

Additionally, the unpredictability of environmental factors in the real world plays a pivotal role. Factors such as lighting conditions, unexpected shadows, or sudden movements can dramatically alter the detection of emotional cues. The absence of a consistent environment means that each instance requires its own unique interpretation, without the luxury of a standard set.

Culture and individual differences further compound these challenges. Emotions and their expressions are deeply rooted in cultural norms and individual experiences. What may be recognized as a sign of happiness in one culture may be interpreted differently in another. This cultural relativity of emotional cues mandates a broader understanding and a more inclusive model for accurate detection.

Moreover, real-world settings often present situations where emotions aren't singular but layered. A person's face might reflect a mix of joy and surprise, or sadness intermingled with elements of disgust. These complex emotional amalgamations demand a sophisticated approach that can discern the subtleties and layers of human emotions.

Occlusions and obstructions are another hurdle. In real-world contexts, faces might not always be fully visible. They may be partially hidden behind objects, hair, or even other people. These partial views introduce uncertainty and can skew the accuracy of detection.

Furthermore, the role of context is paramount. The surroundings, the situation, or even the preceding events can influence how an emotion is expressed and perceived. Without this contextual understanding, a simple facial expression like tears may be misinterpreted, given that tears can signify joy, sadness, or mere physical discomfort.

Lastly, the transient nature of emotions means they can change rapidly. In dynamic real-world situations, capturing the evolution of an emotion becomes crucial. A static interpretation might miss out on these temporal shifts, leading to inaccurate conclusions.

In essence, while commendable progress in emotion detection within controlled settings has been obtained, the "in the wild" domain remains a labyrinth of complexities. Navigating it requires a comprehensive approach that is attuned to the multifaceted nature of human emotions and their contexts.

Consider this example.

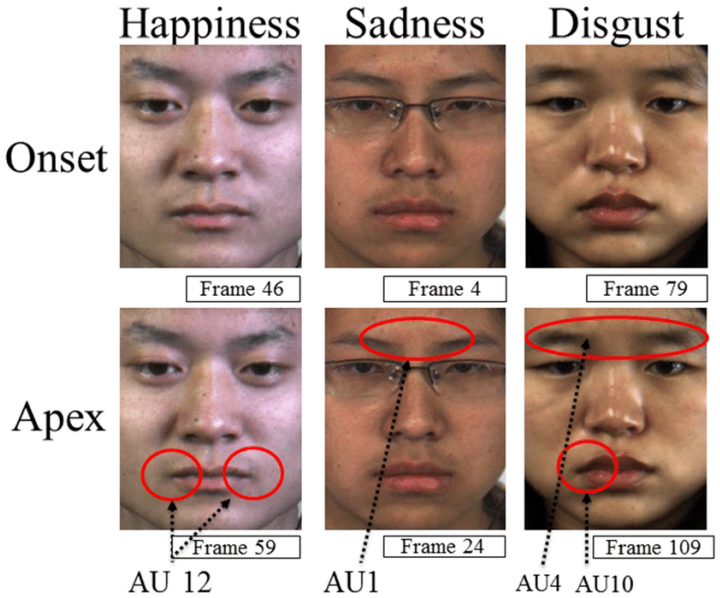

Figure 2.6: Images taken from CASME II dataset.

Taking a glance at the CASME [60] dataset image, one immediately recognizes the con-

trolled, laboratory setting, designed to elicit pure, undiluted emotional expressions. Such settings are instrumental to obtain high-quality data where external factors are minimized, offering a sterile examination of facial emotions. It is also easy to recognise the action units, as can be seen in the proposed image (see Figure 2.6).

Figure 2.7: Images taken from AffWild2 dataset.

On the other hand, the AffWild [24] dataset image showcases the complexity and unpredictability of real-world settings. In-the-wild captures often involve different lighting conditions, partial face occlusions, and a myriad of background distractions (as it turns out in figure 2.7. Emotions here are spontaneous, intertwined with environmental influences and other nuanced human behaviors.

2.4. Practical Applications of Emotional Recognition

With the increasing integration of technology into daily routines, there is a rising demand for machines equipped with the ability to not only comprehend commands but also to discern human emotions. Such a capability holds the potential to reshape interactions with devices, platforms, and interpersonal communications. This section will explore the many practical applications of emotional recognition, highlighting its transformative impact across different industries and its significant implications.

- **Emotion Recognition in Urban Design:** The emotional quality of public spaces plays a pivotal role in influencing the well-being and satisfaction of its users. Leveraging technology like physiological sensors and advanced machine learning can offer insights into how different urban designs resonate emotionally with individuals. For instance, Li et al. utilized ensemble learning to evaluate emotional responses in various urban public spaces [30].
- **FER in Human-Computer Interaction (HCI):** Expanding the horizons of HCI,

FER facilitates more intuitive and emotionally attuned interfaces, fostering richer interactions between humans and machines.

- **Emotion-aware Gaming:** The integration of FER in gaming environments has given rise to games that adapt to players' emotions, creating a personalized and immersive gaming experience.
- **Revolutionizing Marketing with Affective Computing:** The integration of AI and emotional intelligence holds transformative potential for various industries, notably in marketing. Through affective computing, there's an opportunity to enhance customer interactions, innovate products, and fine-tune market research to be more responsive to user emotions. A practical exploration of this synergy can be seen in the work on emotionally intelligent machines in marketing [51].
- **Personalized Learning with FER:** By recognizing students' emotional states, e-learning platforms can adjust content delivery in real-time, ensuring optimal engagement and comprehension.
- **Healthcare and Therapy:** FER plays a pivotal role in monitoring patients' emotional states, assisting healthcare providers in understanding non-verbal cues, especially for patients with communication difficulties.

Drawing insights from these seminal works, we embark on exploring the myriad applications of FER, each promising to redefine the landscape of human-machine interactions and more.

2.5. Social and Ethical Implications

The proliferation of affective computing and emotion recognition technologies promises revolutionary changes across diverse sectors, from urban design to marketing. However, the intertwining of such advanced capabilities with societal structures inevitably introduces complex ethical dilemmas.

Privacy emerges as a primary concern. The backbone of these technologies often involves the collection and analysis of personal data, which, in the absence of rigorous oversight, can raise alarming questions about individual consent and the potential for misuse. Are individuals fully aware of how their emotional data is being utilized, and can they truly consent to its collection?

Bias and discrimination further complicate the landscape. Machine learning models, pivotal in affective computing, derive their understanding from training data. If this

data reflects societal biases, the resulting models can inadvertently perpetuate or even amplify these biases. Such a scenario may lead to discriminatory consequences, where certain societal groups find themselves at a disadvantage due to misinterpretations of their emotional states.

Beyond bias, the potential for emotional manipulation looms large. With insights into an individual's emotional predispositions, there is a potential avenue for entities to exploit these emotions for commercial gain, political influence, or other purposes. This ability to 'tune into' human emotions and potentially manipulate them introduces ethical concerns that are yet to be fully grappled with.

Depersonalization is another unintended consequence. The allure of technology that can seemingly understand and respond to human emotions might overshadow the irreplaceable value of genuine human connection. An over-reliance on emotion recognition technology might diminish the depth and authenticity of human interactions, leading to a society where genuine empathy becomes scarce.

Lastly, the issue of accountability remains unresolved. In instances where emotion recognition technologies err, determining responsibility becomes paramount. The ramifications of these errors can be profound, influencing critical aspects of life, from employment to personal relationships.

In conclusion, the potential of affective computing and emotion recognition is undeniable. Yet, as with all transformative technologies, a measured and ethically informed approach is essential. Balancing innovation with respect for individual rights and societal well-being should remain at the heart of future developments.

3 | Attention mechanism in Facial Expression Recognition

In the domain of Deep Learning (DL), numerous solutions have been proposed for Facial Expression Recognition (FER) tasks. The backbone networks of DL-based FER primarily stem from established pre-trained ConvNets, such as VGG [49], VGG-face [36], ResNet [21], and GoogLeNet [50]. Given the diverse network architectures, DL-based FER can be categorized into three main branches: ConvNet learning for FER, ConvNet-RNN learning for FER, and adversarial learning for FER. With the advent of attention mechanisms in recent years, modern networks have begun to incorporate these to spotlight critical information in facial images, enhancing the precision and performance of FER systems. These attention-driven models are particularly adept at discerning distinctive features and reducing ambiguities, marking them as a cutting-edge evolution in the FER landscape.

3.1. Vision Transformers: revolutionizing Computer Vision

3.1.1. Attention mechanism and Transformer

Deep neural networks (DNNs) have brought transformative advancements to a range of AI applications. Starting with the foundational work of multi-layer perceptrons (MLP) by [42], the field rapidly evolved with the advent of convolutional neural networks (CNNs) tailored for image tasks [25, 27] and recurrent neural networks (RNNs) for processing sequential data [43]. More recently, the attention mechanism has further enriched the capabilities of these networks, enabling more refined contextual understanding and model interpretability.

The attention mechanism, introduced in the paper "Attention Is All You Need" [52], revolutionized the field of deep learning. It allows models to focus on different parts of

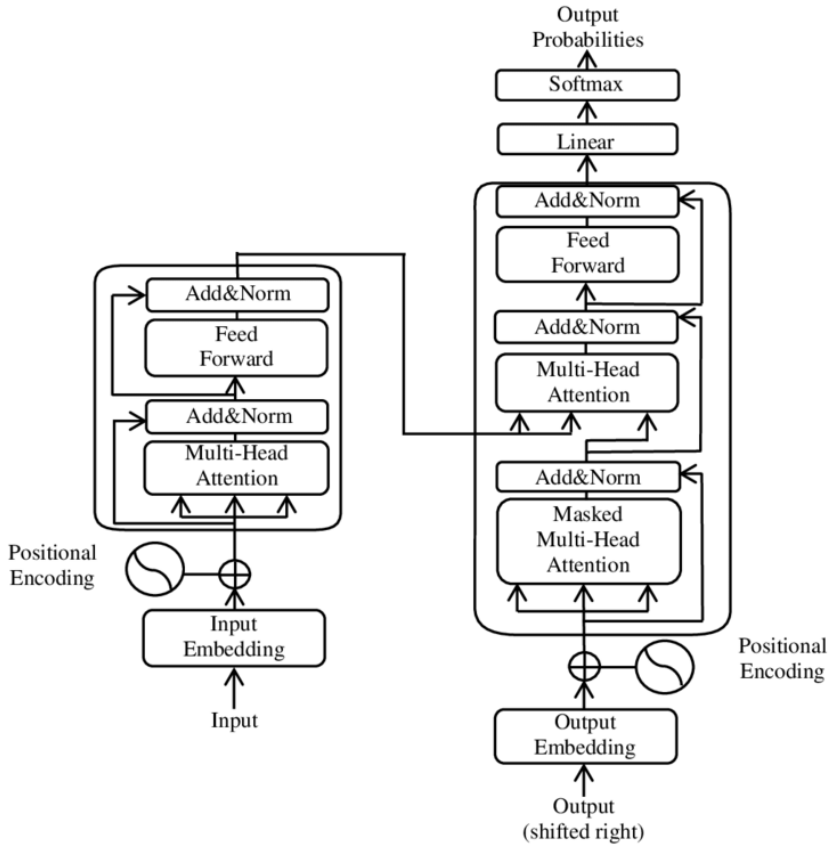

Figure 3.1: Transformer architecture proposed by Vaswani et al. in [52].

the input data, depending on their relevance, essentially enabling the model to "attend" to specific information while downplaying others. This mechanism works by computing a weighted sum of values based on their compatibility with a given query. The strength of this compatibility is determined by an attention score, which is then normalized using a softmax function (see Equation 3.1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

Building upon this foundational concept, the Transformer architecture (figure 3.1) was introduced. Unlike traditional sequence-to-sequence models that rely on recurrent or convolutional layers, Transformers are entirely based on attention mechanisms, not requiring recurrence. The Transformer consists of an encoder and a decoder, each including multiple layers of multi-head self-attention and position-wise feed-forward networks. This design allows Transformers to handle long-range dependencies in data effectively.

Its success in NLP led to its adoption in computer vision (CV). Unlike CNNs, which em-

phasize local features, Vision Transformers (ViTs) are adept at global context awareness due to their self-attention mechanism, offering a more comprehensive image understanding [14]. ViTs scale effectively with data and lack a fixed pattern, offering flexibility for task-specific architectures.

3.1.2. What are Vision Transformers (ViTs)?

For many years, Convolutional Neural Networks (CNNs) have stood as the cornerstone of state-of-the-art computer vision systems, spanning various tasks from image classification to object detection and segmentation. Yet, this status quo began to shift with the emergence of Vision Transformers (3.2).

Instead of operating on pixels directly as CNNs, the Visual Transformer (Vit) first divides the input image into fixed-size patches. Each patch is then linearly embedded into a flat vector and optionally combined with a positional embedding to retain spatial information. The collection of these embedded patches forms a sequence, akin to a sentence in NLP.

An extra learnable embedding, often termed the "class" embedding, is prepended to this sequence. This sequence is then processed by the Transformer encoder, which uses self-attention mechanisms to weigh the importance of different patches relative to one another. After several layers of such transformations, the output corresponding to the "class" embedding is passed to a multi-layer perceptron (MLP) head, which produces the final classification output.

Transformers have been applied to various CV tasks, including image classification by Chen et al. [8], and Dosovitskiy et al.'s ViT model [14], object detection [5], segmentation [62]. Yet, this status quo began to shift with the emergence of Vision Transformers (ViTs), which are now contesting and occasionally outperforming CNNs in different vision applications. These better performances can be seen in Figure 3.3, taken from [14].

3.1.3. Advantages of Vision Transformers

One of the most significant benefits of ViTs is their global context awareness (CIT). In contrast to CNNs, which generally focus on local features and aggregate them in a hierarchical manner, ViTs inherently understand global contexts in images, thanks to their self-attention mechanism. Furthermore, transformers are designed to scale effectively with increasing data and computational power. This scalability means that ViTs often thrive with larger datasets and more extensive model sizes [14]. Additionally, the lack of a fixed architectural pattern in ViTs, unlike the typical convolution-pooling layers in

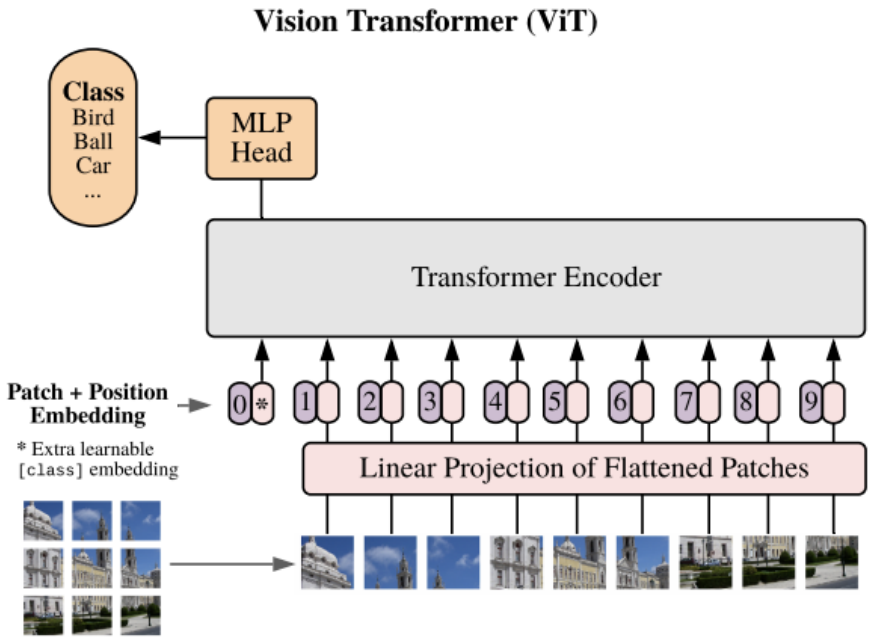

Figure 3.2: Vision transformer architecture proposed by Dosovitskiy et al. in [14].

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Figure 3.3: Comparison with state of the art on popular image classification benchmarks as described by Dosovitskiy et al. in [14].

CNNs, offers immense flexibility in tailoring architectures to specific tasks.

In Figure 3.4 each image is accompanied by an overlaid heatmap indicating areas of attention and significance.

This visualization can be seen as a representation of the attention mechanism in action, where the network highlights regions in the image it deems important for its decision-making. Unlike traditional Convolutional Neural Networks (CNNs), which extract hierarchical features through convolutional layers and gradually build understanding from

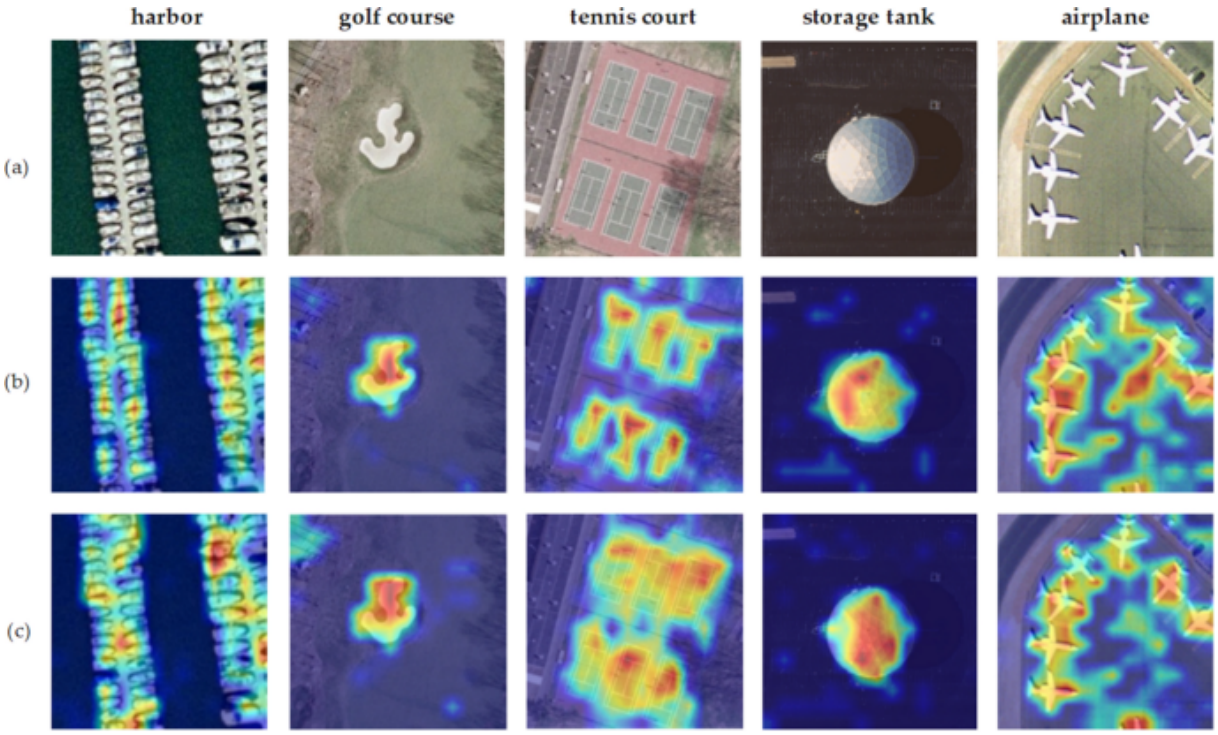

Figure 3.4: Attention mechanism in action.

simple to complex patterns, attention mechanisms allow models to dynamically focus on specific parts of an image. This focus can potentially offer insights into what the model "thinks" is crucial for its prediction, by highlighting these regions.

CNNs employ spatial hierarchies and are primarily designed to recognize patterns in a translation-invariant manner. Their layers progressively capture abstract information. On the other hand, the attention mechanism, as visualized in this image, offers a more intuitive understanding of which parts of the input data the model is emphasizing, providing a more interpretable machine learning approach.

3.1.4. ViT: Bridging Attention and FER

Attention mechanisms are designed to highlight and emphasize the most informative regions in facial images, effectively helping models to focus on the crucial features relevant for recognizing emotions. This mechanism serves as a bridge between visual transformers and FER by providing models the ability to selectively focus on pivotal facial features, enhancing their performance in emotion recognition tasks. Various attention mechanisms are proposed to distinguish distinctive facial features that play a pivotal role in FER [28, 56]. In the ever-evolving realm of FER, the Vision Transformer (ViT) stands out in various models and tasks. Li et al. [29] addressed FER challenges with the Mask

Vision Transformer (MVT), aimed at removing complicated backdrops and fixing label inaccuracies. Additionally, Wang et al. [41] brought attention to pose and occlusion problems, introducing datasets with enhanced annotations. To refine feature recognition, Wen et al. developed the Distract your Attention Network (DAN) [23].

POSTER++

It is important to mention a network architecture that will often be used in this work as a reference model. The architectural design of POSTER++ [34], showed in Figure 3.5 directly inspired the proposed model .

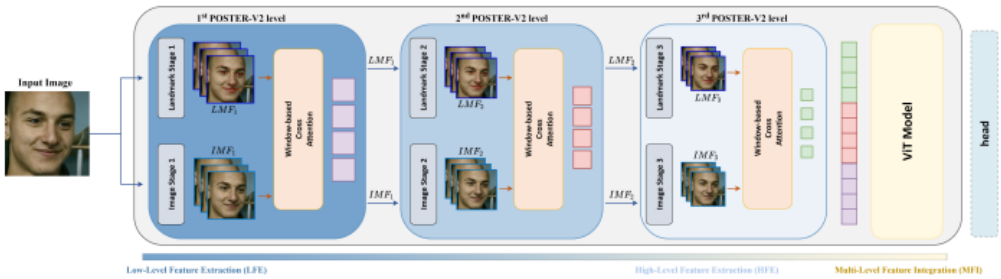

Figure 3.5: POSTER++ architecture proposed by Mao et al. in [34]. The various parts of the network with their functionalities are clearly visible, as well as the 3 levels of depth.

The POSTER++ model proposes a facial landmark detector and image backbone. Central to its design is the window-based cross-attention mechanism, structured to facilitate linear computation. Within this system, image features are systematically segregated into multiple non-overlapping windows, while the landmark features are restructured to synchronize with the image feature windows, ensuring seamless interplay. This intricate design philosophy culminates in the window-based multi-head cross-attention (W-MCSA).

Another notable feature of POSTER++ is its approach to feature extraction. The model obtains multi-scale features both from the facial landmark detector and the image backbone. These multi-scale features, once identified, are amalgamated using a streamlined vision transformer. The consolidated features then traverse through vanilla transformer blocks, utilizing the attention mechanism to manage long-range dependencies across different scales.

POSTER++ introduces a novel outlook on scale sensitivity in facial expression recognition (FER). This unique standpoint, synergized with the deployment of the vanilla transformer block for feature amalgamation, underscores POSTER++'s dedication to innovation.

Expanding upon the principles and innovations of POSTER++, the ARBEx [57] method-

ology has emerged as an advanced proposition. Grounded in the mechanics of POSTER++, ARBEx integrates a distinctive reliability balancing mechanism to navigate the intricate challenges presented by Facial Expression Learning datasets, which traditionally label each sample singularly. Drawing inspiration from various studies, ARBEx adopts a nuanced approach, emphasizing the refinement of label distributions using a label correction paradigm. At the outset, a label distribution is formulated, harnessing the embedding "e" seamlessly within the MLP network, a hallmark of its POSTER++ foundation. Subsequent to this, ARBEx's reliability balancing feature employs specialized label correction techniques, ensuring the stabilization of the primary distribution. This meticulous process culminates in heightened predictive precision, anchored by more robust and reliable labeling.

4 | Methods and materials

From what has been presented in past sections, Facial Expression Recognition (FER) in unconstrained, real-world environments, often referred to as "in the wild", presents a myriad of challenges that can significantly affect the recognition accuracy. Navigating through these challenges necessitates the correct selection and utilization of datasets that are representative of such unconstrained scenarios. However, the current landscape of datasets available for "in the wild" emotion recognition is somewhat limited. Furthermore, only a handful of these datasets delve into the realm of multi-modal learning, which integrates data from multiple sources or modalities, adding another layer of complexity to the task. The work will therefore focus on a FER task.

Central to the solution proposed in this work is the adoption of a discrete output model, mapping closely to Paul Ekman's universally recognized set of emotions. This approach not only offers clarity in terms of the emotional categorizations but also streamlines the recognition process.

Venturing into the latest advancements in computer vision, this work's overarching objective is to construct a model profoundly rooted in attention mechanisms. By leveraging these paradigms, the intention is to foster a more intuitive and effective FER system, tailored for "in the wild" scenarios.

In this chapter, the methods and materials underpinning this endeavor will be thoroughly presented, offering insights into the foundational strategies, dataset considerations, and innovative attention-based modeling techniques employed.

4.1. Datasets

Continuing from our exploration of facial expression recognition (FER) technology, it is clear that comprehensive datasets are the lifeblood of meaningful advancements in the field. For algorithms to be both robust and generalizable, they require vast and varied data for training and validation. However, despite the pivotal role of "in the wild" datasets — those capturing spontaneous facial expressions in real-world scenarios — there remains

a significant gap in their availability.

The reasons behind this gap are multifaceted. For one, the increasing awareness and concerns surrounding privacy, especially with GDPR-like regulations coming to the fore, make the collection of facial data a potential legal minefield. Consent becomes paramount, and without explicit permissions, obtaining and utilizing such data for research can fall into murky ethical waters.

An added layer of complexity is the geographical skew in dataset origins. A significant proportion of accessible datasets emanate from China, presenting a dual challenge. Firstly, the potential cultural, social, and even physiological variances can introduce biases in any algorithm trained solely on these datasets. If FER technology is to be truly global, it cannot lean predominantly on data from a single region.

Further deepening this issue is the challenge of representation within these datasets. Not only is there an ethnic skew, but certain datasets also display imbalances in terms of age, gender, and even socio-economic indicators. For instance, a FER system trained predominantly on younger faces might struggle to accurately interpret the nuanced expressions of an older individual. Similarly, cultural norms and social behaviors can influence the way emotions are expressed, meaning a system might misinterpret a facial gesture from one culture if trained predominantly on data from another.

At the core of it, datasets play a crucial role in driving progress in Facial Expression Recognition. However, the current landscape of these datasets highlights a pressing requirement for wider representation, adherence to ethical data collection standards, and a genuine commitment towards ensuring global inclusivity. It is imperative to understand that only by addressing these concerns we can truly aspire to craft FER systems that stand as equitable, precise, and universally relevant in diverse settings.

In table 4.1 there's an overview of facial expression datasets in the wild, and after that a little description of each dataset it had been considered.

In-the-Wild	**Year**	**Data Size**	**Expression category**
FER2013 [19]	2013	35,857 gray images	SBE
EmotioNet [2]	2016	1,000,000 images	Compound expressions
AffectNet [35]	2017	450,000 images	SBE (plus contempt)
RAF-DB [31]	2017	29,672 images	SBE Compound expressions
DFEW [22]	2020	12,059 clips	SBE
AffWild2 [24]	2019	546 videos (2.6 milion frames)	SBE plus valence and arousal

Table 4.1: List of emotion recognition datasets in the wild. Seven Basic Emotions (SBE): anger, disgust, fear, happy, sad, surprise, and neutral.

- **EmotioNet [2]:** Consistes of one million images of facial expressions retrieved from the Internet using words derived from "feeling" in WordNet. Faces in these images were detected and annotated with Action Units (AUs) using Kernel Subclass Discriminant Analysis (KSDA). Images were categorized into 23 emotion categories based on the AUs. Additionally, 10% of the database (100,000 images) were manually annotated by expert coders. The dataset offers a comprehensive resource for the FACS model in wild settings, though it lacks a dimensional model of affect.
- **FER2013 [19]:** Introduced in the ICML 2013, this dataset consists of 35,887 images representing the six basic expressions plus the neutral expression. The images were acquired using the Google image search API, resized to 48x48 pixels, and converted to grayscale (see Figure 4.1). Human labelers vetted the images to ensure quality. Currently, FER-2013 stands as one of the largest publicly available facial expression databases in wild settings, suitable for training Deep Neural Networks (DNNs). However, it has some limitations such as unregistered faces and limited representation of certain expressions.

Figure 4.1: Samples from FER2013.

- **DFEW [22]:** Dynamic Facial Expression in the Wild (DFEW) consists of over 16,000 video clips segmented from thousands of movies with various themes. Professional crowdsourcing is applied to these clips, and 12,059 clips have been selected and labelled with one of 7 expressions (six basic expressions plus neutral).
- **RAD-DB [31]:** The Real-world Affective Faces Database (RAF-DB) is a significant facial expression collection featuring nearly 30,000 images from the Internet (see a sample in Figure 4.2). Each has been meticulously annotated by around 40 crowdsourced annotators. The database stands out for its diversity, encompassing variations in age, gender, ethnicity, and head poses, and accounting for challenges

like varied lighting and occlusions. RAF-DB provides a 7-dimensional expression distribution vector for each image, contains two subsets focusing on basic and compound emotions.

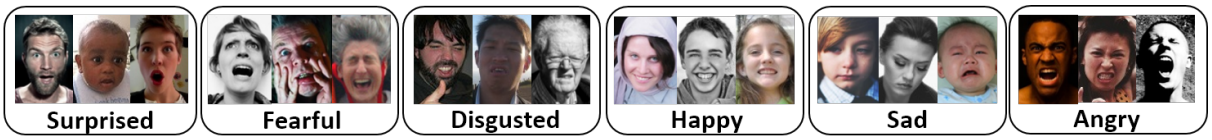

Figure 4.2: Samples from RAF-DB.

- **AffWild2 [24]:** AffWild2 is a pioneering database in affective computing, addressing the gaps found in previous in-the-wild datasets. It features frame-by-frame annotations for seven basic facial expressions, twelve action units, and emotions like valence and arousal. The database consists of 564 videos with approximately 2.8 million frames, covering 554 diverse subjects across various ages, ethnicity, and backgrounds. Its uniqueness is further emphasized by its integration of audiovisual data and action unit annotations, making AffWild2 an unmatched resource in capturing genuine, real-world emotional scenarios.
- **AffectNet [35]:** Given its importance in this work, this dataset will be explored in a separate section.

4.1.1. AffectNet

AffectNet is one of the most extensive manually annotated databases specifically designed for facial expression recognition in the wild. The evolution of facial affect analysis underscores the deep interconnection between the capabilities of technology and the richness of the data it utilizes. The realm of affective and the related machine learning techniques are significantly bolstered by the availability of comprehensive and diverse annotated datasets. A shining example of this interdependency is the inception of AffectNet [35].

AffectNet wasn't conceived merely as another dataset in the computing world; it embodies a vision. It aims to serve as a monumental repository for researchers diving deep into facial expression recognition, particularly in settings that are uncontrolled and replicate real-world scenarios. What sets AffectNet apart is its foundational robustness, derived not just from its sheer size but from its diverse composition. The dataset encapsulates a rich tapestry of human emotions across a plethora of ethnic backgrounds and age groups, a crucial target in this field. See a sample in Figure 4.3

The genesis of AffectNet can be traced back to a meticulous process of collating emotion

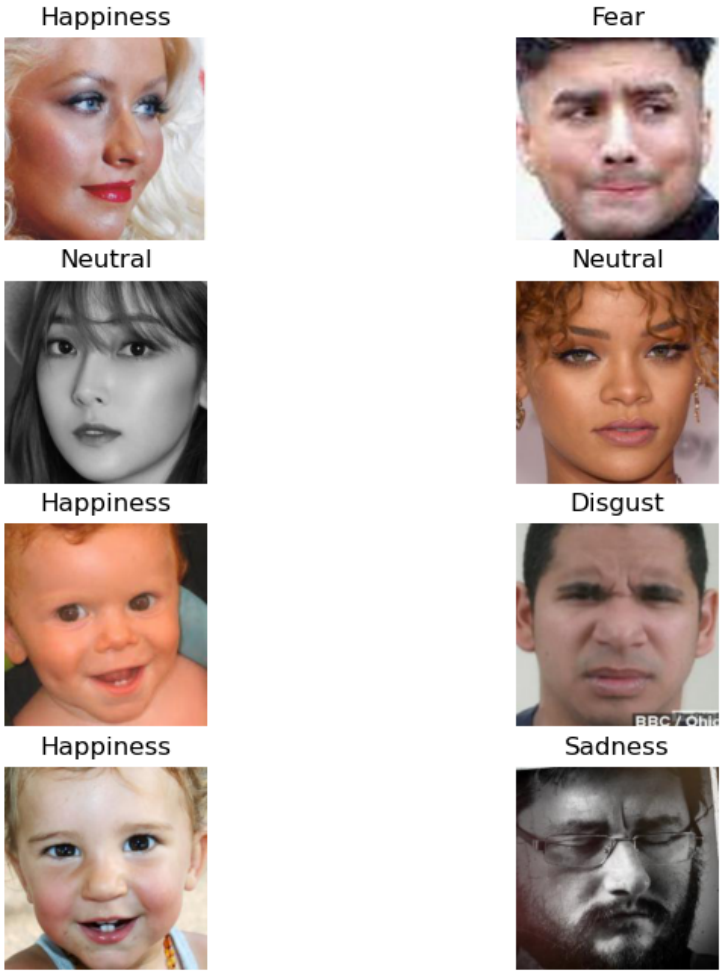

Figure 4.3: Samples from AffectNet.

descriptors from a range of attributes – from gender to age to ethnicity. This effort culminated in 362 distinctive English phrases, each capturing intricate emotional nuances. However, recognizing the universality of emotion that transcends linguistic boundaries, these phrases were translated into several languages such as Spanish, Portuguese, German, Arabic, and Farsi. This wasn't merely a translation: bilingual experts ensured each term resonated emotionally, reflecting the cultural nuances and richness of each language [35].

Harnessing the vastness of the internet, a collection of 1.8 million unique URLs was compiled using 1250 emotion-centric search queries across major search engines like Google,

Bing, and Yahoo. This endeavor wasn't just about data gathering. The mission was to encapsulate the intricate spectrum of human emotions. A deliberate effort was made to ensure the integrity of this data collection by avoiding potential pitfalls, such as cartoons and stock images.

These URLs were subsequently transformed into distinguishable facial images using OpenCV's face recognition capabilities. To augment this data, Microsoft's Cognitive Face API provided additional insights on these images, ranging from demographics like age and gender to intricate details like makeup or glasses presence, facial occlusions, and head orientations.

The annotations of AffectNet form its essence. Instead of relying on crowdsourced data, experienced annotators from the University of Denver labeled 450,000 images, distinguishing between well-defined emotional states and nuanced valence and arousal parameters.

AffectNet's categorical model considers eleven discrete categories: Neutral, Happyness, Sadness, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face. Figure 4.4 details the number of images in each category. Notably, a vast proportion of the images returned from search engines were labeled as happy or neutral, possibly hinting at a societal preference to share positive images. The dataset also offers a selection of images across categories, underscoring its diversity and breadth.

Expression	Number
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
Contempt	5,135
None	35,322
Uncertain	13,163
Non-Face	88,895

Figure 4.4: Annotated images for each class.

Figure 4.5 illustrates the distribution of emotion labels in AffectNet. The histogram clearly reveals an imbalance in the quantity of labeled data for various emotions. "Happiness" stands out prominently, far surpassing other categories. "Neutral" follows closely, while emotions such as "Sadness," "Surprise," and "Anger" have fewer data points. Specifically, emotions like "Fear," "Disgust," and "Contempt" are sparsely represented, highlighting a dearth of data for these emotions.

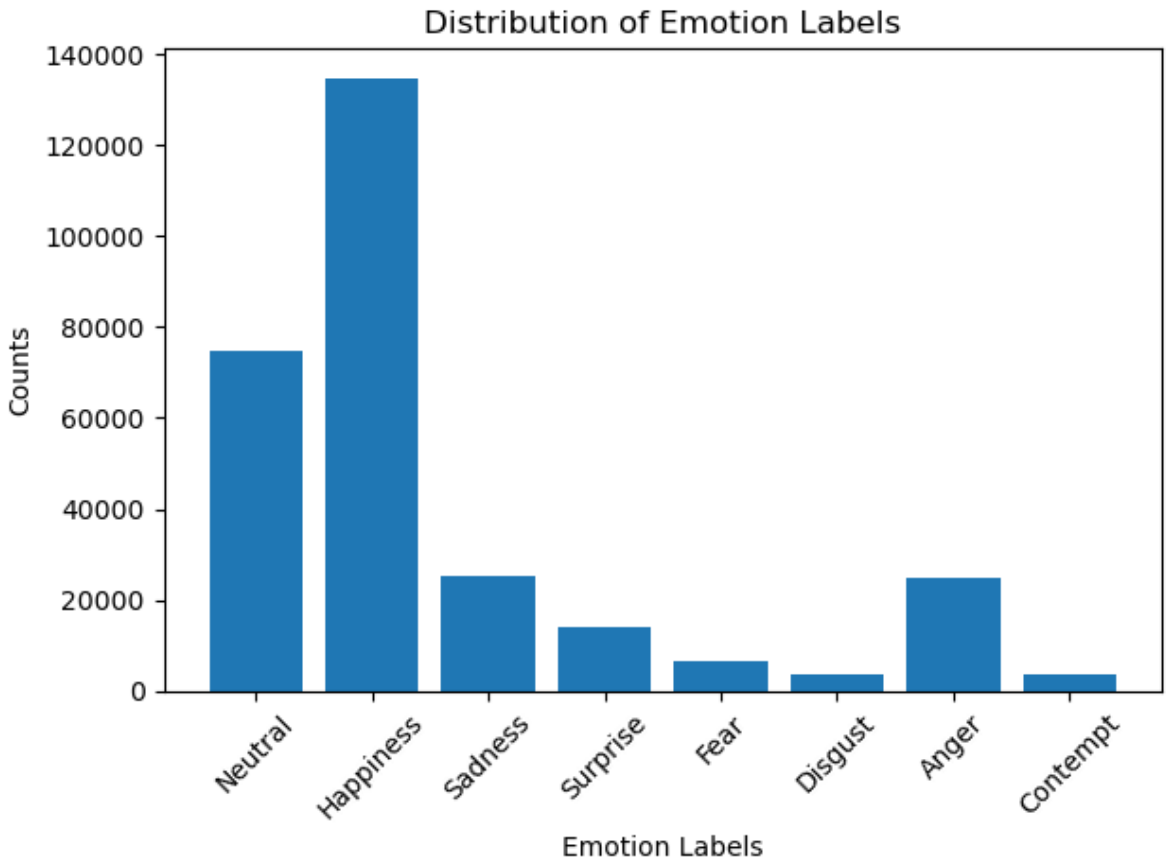

Figure 4.5: Class distribution in AffectNet.

This disproportion might be attributed to the challenges in obtaining genuine images that portray rarer emotional states, especially in real-life contexts. It is conceivable that individuals favor sharing images reflecting positive emotions like happiness over negative ones, leading to this dataset imbalance.

If used for training without modification, a model may become biased towards more common emotions, such as happiness, neglecting less prevalent ones. This may reduce accuracy in identifying underrepresented emotions in practical applications, potentially affecting the efficacy of emotion recognition systems.

Figure 4.6 presents a table showcasing the relationship between "Query Expression" and "Annotated Expression". Rows represent the Annotated Expression, while columns correspond to the Query Expression. Both rows and columns are labeled with abbreviations that likely denote different emotions or states, such as "HA", "SA," "SU", and so forth. The table's cells contain numerical values, presumably percentages, which indicate the correspondence between a specific queried expression and how it was annotated. For ex-

		Query Expression						
		HA	SA	SU	FE	DI	AN	CO
Annotated Expression	NE*	17.3	16.3	13.9	17.8	17.8	16.1	20.1
	HA	**48.9**	**27.2**	**30.4**	**28.6**	**33**	**29.5**	**30.1**
	SA	2.6	15.7	4.8	5.8	4.5	5.4	4.6
	SU	2.7	3.1	16	4.4	3.6	3.4	4.1
	FE	0.7	1.2	4.2	4	1.5	1.4	1.3
	DI	0.6	0.7	0.7	0.9	2.7	1.1	1
	AN	2.8	4.5	3.8	5.6	6	12.2	6.1
	CO	1.3	0.9	0.4	1.1	1.1	1.2	2.4
	NO	5.4	8.7	4.8	8.1	8.8	9.3	11.2
	UN	1.3	3.1	4.3	3.1	4.1	3.7	2.7
	NF	16.3	18.6	16.7	20.6	16.9	16.8	16.3

Figure 4.6: Annotated categories for queried emotions.

ample, when "HA" was queried, 48.9% of the time it was annotated as HA, but 16.3% of the time it was annotated as NE.

AffectNet was used as the reference dataset for the developed model, thus basing the performance of the architecture on the results obtained from its use.

4.2. Experimental setup

To ensure that a machine learning or deep learning model performs effectively, it is vital to set up a comprehensive experimental framework. In this section, we delineate the key components that constitute the experimental setup for the project.

4.2.1. Used framework

For the development of this project, **PyTorch** [37] was chosen as the primary deep learning framework. PyTorch, developed by Facebook's AI Research lab, stands out due to its dynamic computation graph, which provides a high degree of flexibility during model development. This is especially beneficial for experimenting with novel architectures and making on-the-fly changes.

Some of the core reasons for choosing PyTorch include:

- **Simple Syntax:** PyTorch's Pythonic nature makes it relatively easy to understand and write, even for those who might be newer to deep learning.
- **Rich Ecosystem:** PyTorch comes with a vast array of utilities, pre-trained models, and community-contributed tools, which expedite the development process.
- **Strong Community Support:** The PyTorch community is active and growing, ensuring that any issues, queries, or doubts can be swiftly addressed through forums,

discussions, and documentation.

Moreover, the project leveraged various PyTorch-specific libraries and extensions, such as `torch.nn` for neural network modules and `torch.optim` for optimization routines. The choice of PyTorch not only streamlined the development process but also provided a robust foundation for future extensions and modifications to the project.

4.2.2. Model Training Configuration

Training a neural network needs careful selection and tuning of several hyper-parameters. These parameters are instrumental in guiding how the model learns from the data. The choices made in this context are based on both empirical evidence and practical experience. Below is a breakdown of the primary training parameters used:

- **Batch Size:** Following various experiments and considering the available GPUs (NVIDIA GeForce GTX 1080 Ti), a default batch size of 32 was established for training, whereas a batch size of 8 was selected for validation. This configuration strikes a balance between computational efficiency and the model's capacity to generalize.
- **Loss Function:** Cross-entropy loss, often termed as log loss or logistic loss, is pivotal for classification tasks. It measures the difference between predicted probabilities and actual class labels. This metric penalizes predictions based on their deviation from the true labels, with a logarithmic penalty—large deviations result in higher penalties and vice versa. During model training, the goal is to minimize this loss. Ideally, a perfect model would have a cross-entropy loss of 0, meaning predictions align flawlessly with true labels. Given its efficacy in highlighting discrepancies, cross-entropy is a preferred choice for classification challenges.

$$CrossEntropy = -\sum_{i} t_i \log(f(s)_i) \tag{4.1}$$

  $CrossEntropy$ represents the cross entropy loss for a given classification task. Notice that t_i is the true label of the sample, while $f(s)_i$ is the Softmax probability of the i^{th} class.
- **Optimizer:** The Stochastic Gradient Descent (SGD) optimizer was utilized to update the network's weights. It is one of the most common and well-established optimization techniques. Learning rate and weight decay are two crucial hyper-parameters for this optimizer. The learning rate determines the step size at each

iteration when updating the weights. Initially, a value of 1×10^{-4} was used, which can also be represented as $1e-4$. Conversely, weight decay, set at $1e-4$ (as cited in [34]), aids in preventing the model from overfitting by incorporating a penalty into the loss function.

- **Number of Epochs:** The number of training epochs is variable, as the early stopping mechanism was implemented. While the maximum limit was set to 50 epochs, with each epoch representing a complete pass over the training dataset, the actual number may be less due to early stopping. As we will observe later, the model tends to overfit easily, making the use of early stopping frequently necessary. This mechanism ensures the model halts training once it ceases to benefit from additional epochs, thus helping to mitigate the risk of overfitting.

4.2.3. Data Adjustments

Data Augmentation

Sample augmentation is a technique that artificially amplifies the training set by generating altered replicas of the dataset using existing data. This method facilitates the rapid recognition of significant attributes from the data. Augmentation ensures that the model is provided with a diverse set of data points, thereby achieving robust training data. In FER problems, common data pre-processing and augmentation steps encompass image resizing, scaling, rotating, padding, flipping, cropping, color augmentation, and image normalization.

In this project, a multitude of image processing techniques have been employed:

Image Resizing and Scaling The method of bi-linear interpolation is employed for image resizing. Linear interpolation is applied in both the x and y directions to adjust the image size. This technique is iteratively employed until the desired size is attained [**?**]. Given a function fs whose value needs to be estimated at coordinates (x, y), and when the function values at vertices $Q11 = (x1, y1)$, $Q12 = (x1, y2)$, $Q21 = (x2, y1)$, and $Q22 = (x2, y2)$ are known, the interpolation equation is:

$$fs(x,y) = \frac{1}{(x2-x1)(y2-y1)} \begin{bmatrix} x2-x & x-x1 \end{bmatrix} \begin{bmatrix} fs(Q11) & fs(Q12) \\ fs(Q21) & fs(Q22) \end{bmatrix} \begin{bmatrix} y2-y \\ y-y1 \end{bmatrix} \quad (4.2)$$

This operation continues until the entire image has been appropriately resized.

Color Jitter The "transforms.ColorJitter()" operation introduces variations in brightness, contrast, saturation, and hue of the image. The specified parameters, such as brightness=0.3, contrast=0.3, etc., dictate the strength of the jitter applied. This transformation is particularly beneficial for enhancing the diversity of the dataset, ensuring the model's robustness against variations in lighting and color.

Image Conversion to Tensor The conversion of the transformed image back into a tensor format is a critical step in preparing the data for processing with PyTorch. This is due to PyTorch's design, which requires data to be in tensor format for its computation graph. Tensors are a specialized data structure that are similar to arrays and matrices, and they facilitate efficient operations on the GPU. These operations include mathematical computations that are essential for training neural networks. By converting images to tensors, we ensure that they can be batched together into a single tensor for more efficient computation, and we also make use of PyTorch's GPU acceleration capabilities, leading to faster processing and training times.

Random Flipping Operations The random flip operation can be applied either horizontally or vertically on an input image, based on a predetermined probability. For a given torch tensor $A \in R^{m \times n}$, the matrix $A = A_{ij}$, where i represents the row and j represents the column.

The horizontal flip transformation can be represented as $A \rightarrow A_i(n + 1 - j)$. In this operation, the columns of A are inverted such that the first column of A aligns with the last column of $A_i(n + 1 - j)$ and vice versa.

Conversely, the random vertical flip operation indicates the potential flipping of the input image vertically. The vertical flip transformation can be symbolized as $A \rightarrow A_{m+1-i,j}$. Here, the rows of A are inverted, making the first row of A align with the last row of $A_{m+1-i,j}$ and the other way around.

Normalization The image undergoes normalization using the "transforms.Normalize()" function. The mean and standard deviation values provided, namely mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225], are derived from the ImageNet dataset [12]. These values represent the average and standard deviation of pixel values across each of the RGB channels calculated over the entire ImageNet dataset. The normalization ensures that the pixel values in the image have a standard distribution, which is pivotal for stabilizing the training process and ensuring faster convergence. When leveraging a pre-trained model on ImageNet, it's common practice to use its normalization parameters,

as the model anticipates inputs to be normalized in this particular manner. Adopting these normalization values helps to maintain the distributive properties of the original data when using a pre-trained model, leading to improved performance and convergence during training.

Random Erasing The ‘transforms.RandomErasing(p=0.2)‘ operation introduces random occlusion within the image. With a probability of 0.2, a randomly chosen rectangle in the image is filled with random pixel values. This technique, known as Random Erasing, enhances the model’s robustness by training it to recognize objects even when parts of them are occluded or missing.

In Figure 4.7 there’s two examples of applied transformations to a given input image.

4.2.4. Sampling

As previously observed, obtaining well-balanced “in-the-wild” facial expression recognition (FER) datasets is a challenging endeavor, and AffectNet is no exception. The samples with which networks are trained often mirror this imbalance, making it crucial to employ sampling techniques as a mitigation strategy.

The emotion labels for the dataset are assigned as:

- 0: Neutral
- 1: Happiness
- 2: Sadness
- 3: Surprise
- 4: Fear
- 5: Disgust
- 6: Anger
- 7: Contempt

In Figure 4.8, it is evident that certain emotional labels, notably “Happiness” and “Neutral”, dominate within this batch. This skew reflects the intrinsic imbalance within the AffectNet dataset. Such disparities can bias the model to favor predictions of prevalent labels, which might impair its performance on underrepresented classes. To mitigate this challenge, a weighted sampler strategy was employed.

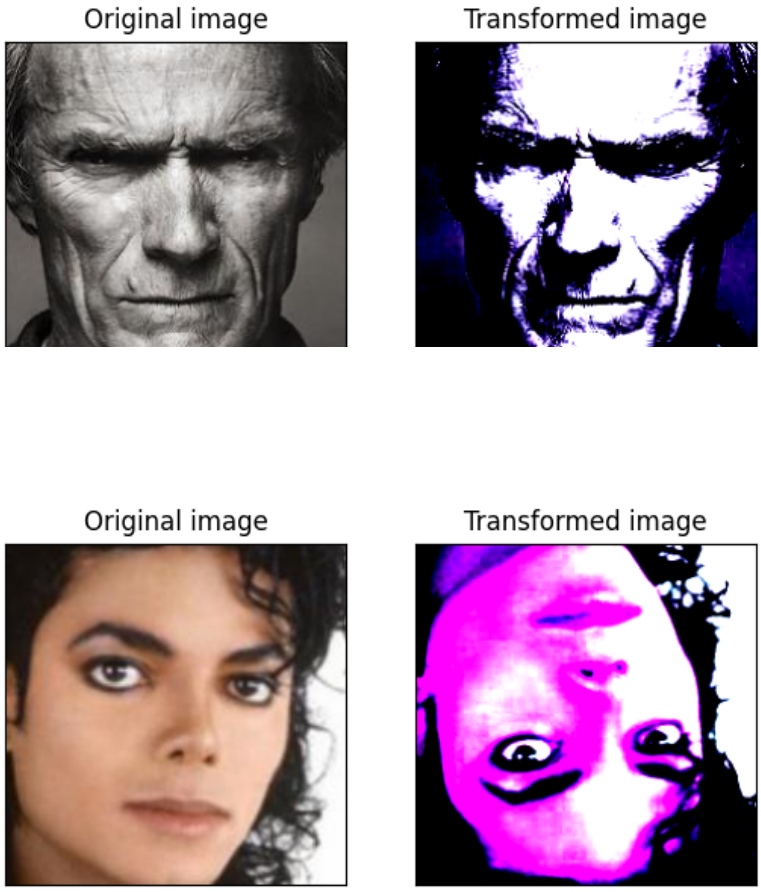

Figure 4.7: Examples of transformation on input images.

A possible solution to this challenge is assigning a weight to each item in the dataset. These weights are typically determined by the inverse frequencies of the labels, meaning that underrepresented classes will have a higher probability of being selected during sampling. In essence, for each batch, the sampler selects items not uniformly, but according to these predefined weights. By leveraging this method, it ensures that during training, each class has a more balanced representation in the batches, counteracting the inherent biases in the dataset and promoting a model that generalizes better across all classes.

To achieve this goal, `torch.utils.data.WeightedRandomSampler` from PyTorch was

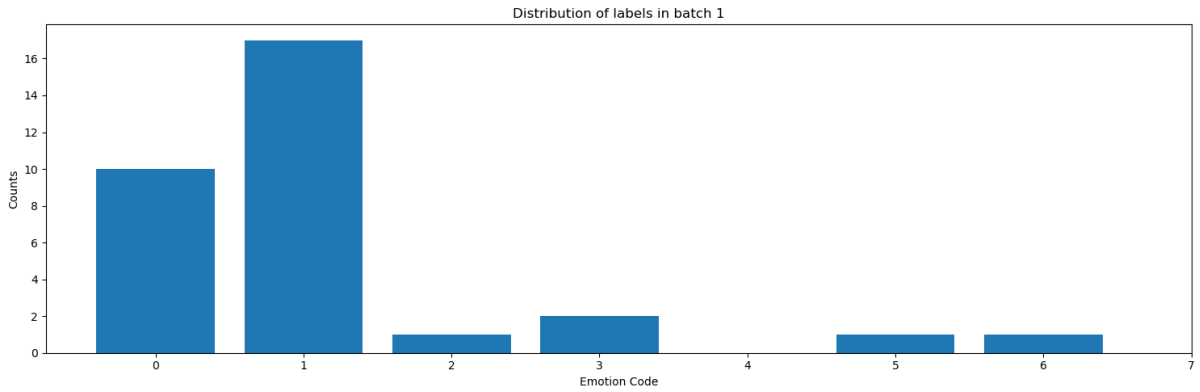

Figure 4.8: Batch distribution without sampling.

used. The intention is to harmonize the probabilities of each class's selection during batch creation, ensuring a consistent label distribution throughout training and enhancing the model's ability to generalize over the entire class spectrum.

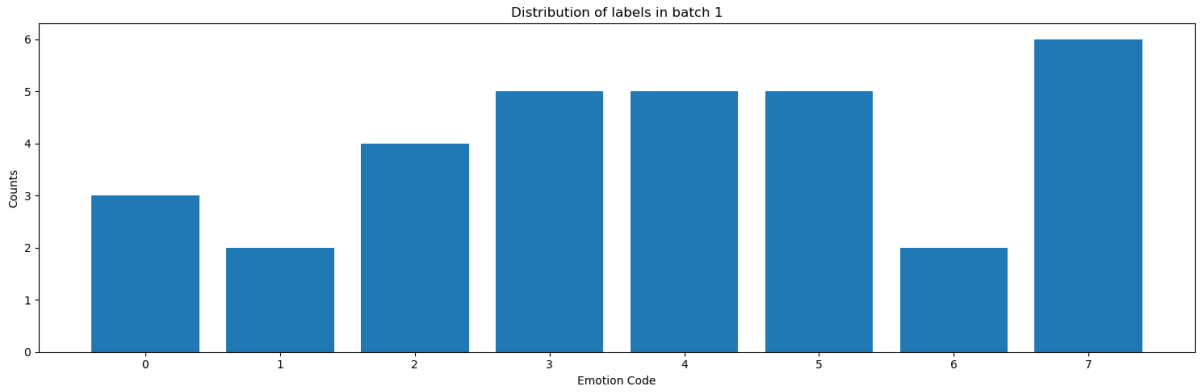

Figure 4.9: Batch distribution with sampling.

4.3. Evaluation Metric

Throughout the experiments, accuracy is utilized as the primary evaluation metric, which is a fundamental concept that measures the correctness of predictions made by a model.

In the case of multi-class classification, accuracy measures the proportion of correct classifications (n_{correct}) and the total number of classified terms (n_{total}). The equation is:

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{4.3}$$

4.4. Target Description

4.4.1. Emotion Recognition on AffectNet

The primary objective of this work is to recognize and classify human emotions from visual input like images or video frames, utilizing the AffectNet dataset, which provides annotations in the form of specific emotion labels for each image.

- Neutral : 0
- Happiness : 1
- Sadness : 2
- Surprise : 3
- Fear : 4
- Disgust : 5
- Anger : 6
- Contempt : 7

All the various proposed models output a feature vector. This vector undergoes a final classification layer, resulting in a vector where each position corresponds to the likelihood of a specific emotion. The emotion associated with the maximum value in this vector is predicted as the dominant emotion for the given image. A visual representation of an output vector for a given input image is shown in Figure 4.10.

As previously said, emotion is a complex construct that encompasses a myriad of subjective feelings, physiological responses, and expressive behaviors. One way to make sense of this complexity is by representing emotions in a multi-dimensional space, commonly referred to as the "emotion space".

The concept of emotion space is derived from psychological theories that aim to categorize and quantify emotions in terms of their primary components. In such a space, each dimension or axis signifies a distinct emotion. The most elementary models consider emotions along two axes: valence (ranging from negative to positive feelings) and arousal (ranging from calmness to excitement). However, for more detailed tasks like emotion recognition in the AffectNet dataset, a higher dimensional space is required to represent each specific emotion label.

In the context of the model being used, the output vector can be perceived as a point in

Figure 4.10: Emotion score for a given image. In this example, the input image is classified as 0: Neutral.

this 8-dimensional emotion space. Each element of this vector represents the strength or score for the corresponding emotion. The proximity of this point to any particular axis (emotion) determines the likelihood of that emotion.

For instance, an output vector with a high value at the "Happiness" dimension would indicate a strong prediction towards the emotion of happiness. The overarching objective is to populate this 8-dimensional emotion hyperspace with an 8D vector generated by the model. By mapping the image data onto this hyperspace, the model can effectively perform emotion predictions based on the positioning and magnitude of the vectors within

this space.

Figur 4.11 offers a visual representation of this 8-dimensional emotion hyperspace concept. The graph showcases how four different samples, denoted by distinct colored lines, fare across eight emotion dimensions, ranging from "Emotion 0" to "Emotion 7." The y-axis quantifies the emotion score, spanning from approximately -0.010 to 0.025.

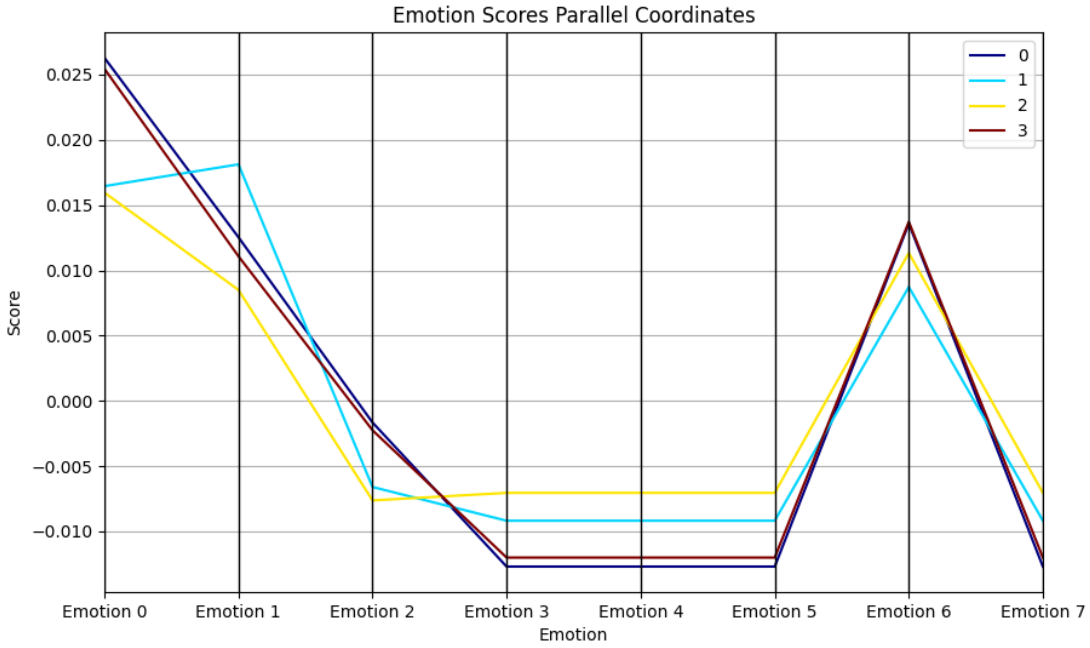

Figure 4.11: Parallel coordinates plot visualizing emotion scores across eight distinct dimensions for four separate samples, highlighting variations and patterns in emotional intensity predictions.

Observing the graph, we can see how different samples compare across the multiple emotion dimensions. For instance, the yellow line, representing sample 2, peaks sharply at “Emotion 6”, indicating a strong prediction or intensity towards that particular emotion for that sample. On the other hand, the light blue line, representing sample 1, remains relatively stable near the zero mark, suggesting a more neutral or balanced emotional profile.

This visual representation provides valuable insights into the model’s predictions, as it shows how the output vectors (samples) are situated within the emotion hyperspace. The relative positions of these lines across the axes give an intuitive understanding of the predicted emotions and their respective intensities.

4.5. Proposed solutions

Within the advancements in facial analysis, greatly propelled by deep learning innovations and expansive datasets, a particular oversight emerges. The prominent convolution-based pre-trained models, such as VGGFace [4] (Figure 4.12) and FaceNet [48], while exhibiting exceptional performance in facial recognition, are distinct from the direction in which current research trends are heading. These convolutional architectures, distinguished for their excellence in facial recognition, offer a solid base that could be further harnessed for a more comprehensive spectrum of facial analysis tasks.

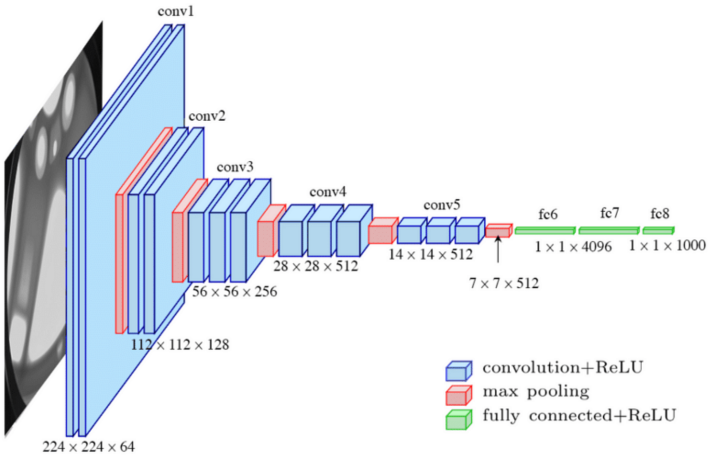

Figure 4.12: Standard VGG architecture.

The solution presented herein pivots on the power of attention mechanisms. By weaving together a diverse tapestry of embedding techniques, this approach aims to encapsulate the richness of facial features and expressions. The exploration begins with more established architectures, such as the classic ViT and Swin ViT, laying down a solid groundwork. Through this series of experiments, we arrive at what can be described as the real focus of this research: the introduction of a model that exploits the dynamics of cross-window attention. This method promises greater granularity in capturing facial nuances.

This proposed solution stands as a testament to this vision, aiming to bridge existing gaps while charting new territories in the realm of affective computing.

This architecture will be discussed in a separate subsection and it unfolds through a series of developmental phases. It is a three stages architecture. Each layer has a two stream architecture. The two streams deal with two different tasks, extracting different expressive features of the face:

- **Landmark stage:** extracts the salient points of the face in order to direct the

image stage to the points to be observed.

- **Image stage:** deals with extracting higher-level features from the image stage.

These features will then be passed through a simple window-cross-attention layer. The idea behind window cross attention is to combine local and global features by limiting the scope of attention to a fixed-size window around each position.

Downstream of the three stages, there is a visual transformer (ViT) to integrate the information and extract the probability vector to enable classification.

Implementation, conceptual details and experiments will be presented in the following subsections.

4.6. Visual Transformer (ViT)

The first model tested, already briefly described in section 3.1.2, is the visual transformer (ViT).

4.6.1. ViT description

Given the Figure 3.2, the following section will review a more pointed and in-depth description of how it is composed.

- **Patch Embedding Layer**: This layer is responsible for converting the input image into a series of smaller, manageable pieces known as patches. Each patch is then linearly transformed into a higher-dimensional space, which serves as the input to the subsequent transformer layers.
- **Transformer Encoder Layers**: At the heart of the ViT are the transformer encoder layers. Figure 4.13 graphically shows its architecture. Each layer consists of:

  The architecture of a **Transformer Encoder**, depicted in Figure 4.13, can be formally described as follows:

  - **Embedded Patches:** The input to the Transformer Encoder is a set of linearly transformed image patches. These patches are flattened and projected into an N-dimensional space to form the embedded patches.
  - **Normalization (Norm):** Before being processed by the multi-head attention mechanism, the embedded patches undergo normalization to stabilize the learning process.

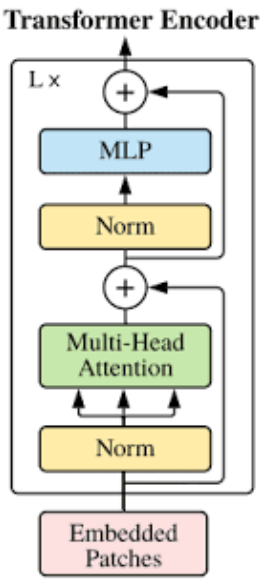

Figure 4.13: Classic Transformer Encoder architecture proposed by Vaswani et al. in [52].

- **Multi-Head Attention (Figure 4.14):** This module performs the attention operation in parallel across multiple heads, allowing the model to focus on different parts of the input sequence simultaneously. Each head computes an output that is a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.
- ***Addition Operation:*** Following the multi-head attention, the output is combined with the original input through a residual connection, followed by another normalization step. This addition operation helps to avoid the vanishing gradient problem and encourages feature reuse.
- **Second Normalization (Norm):** A second normalization layer is applied after the addition operation to prepare the output for the subsequent feed-forward neural network.
- **Multi-Layer Perceptron (MLP):** This is a feed-forward neural network that further processes the output from the attention mechanism. It typically consists of two linear transformations with a non-linear activation function in between.
- ***Addition Operation:*** The output of the MLP is again added back to the input through a residual connection, which is a standard component in Transformer architectures to facilitate deeper stacking of layers.

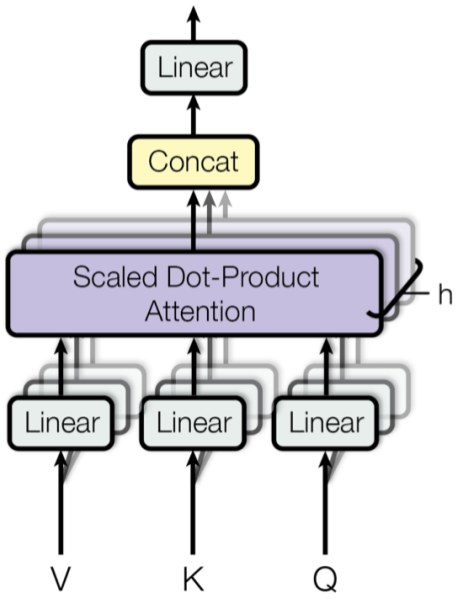

Figure 4.14: Illustration of the Multi-Head Attention mechanism. The module processes input through multiple attention mechanisms in parallel, denoted by Q, K, and V which are then passed through linear transformations. The outputs are combined using concatenation, followed by another linear transformation. The scaled dot-product attention, indicated here, is a typical choice for these mechanisms, although other forms of attention could also be applied.

  - **Output (L x):** The final output of the Transformer Encoder is a sequence of vectors, each corresponding to the input patches, enriched by the contextual information aggregated through the attention and MLP mechanisms.

- **Normalization and Regularization**: Throughout the architecture, normalization techniques are employed to ensure that the activations remain well-scaled and stable throughout the training process. Regularization methods like dropout are implemented to improve the robustness of the model and mitigate overfitting, which is especially important for models with a large number of parameters.

- **Classification Head**: The final layer of the ViT is a linear projection that maps the encoded image representations to the desired output classes. This layer typically has as many units as there are classes in the classification task. For the scope of this thesis work, the output layer is designed to accommodate **8 classes**, which corresponds to the specific classification requirements of the study.

4.6.2. ViT Results

In the exploration of the Vision Transformer's (ViT) capabilities for the task at hand, the configuration parameters were aligned with those specified in [14]. This ensured that the foundational aspects of the network's architecture were preserved.

More in detail, two versions of ViT will be considered: the basic one and the large one. In table 4.2 configuration parameters are specified:

Model	Layers	Hidden size	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M

Table 4.2: Comparison of ViT-Base and ViT-Large model configurations as described in [14].

In the configuration of the Vision Transformer (ViT), key hyperparameters are set to define the architecture's size and complexity. The specific parameters chosen can result in different variants of ViT, such as ViT-Base or ViT-Large, each suitable for different scales of data and computational resources. The parameters for this particular setup are as follows:

- **Image Size:** The input images will be 224×224 pixels.
- **Patch Size:** Each image is divided into patches of 16×16 pixels, as shown in Figure 4.15.

Figure 4.15: Input image divided in 16×16 pixels.

- **Number of Layers:** The depth of the transformer, variable between basic (12)and large (24) ViT.

- **Number of Heads:** The multi-head attention mechanism allows the model to attend to different parts of the image simultaneously. It is variable between basic (12) and large (16) ViT.
- **Hidden Dimension:** The size of the hidden layers is dependent on the type of ViT, whether it is basic or large, and it influences the width of the model.
- **MLP Dimension:** The size of the feedforward layers within the transformer blocks is 4096.
- **Dropout:** A dropout rate of 0.4 is used for regularization to prevent overfitting.
- **Attention Dropout:** The attention scores are also regularized with a dropout rate of 0.4.
- **Number of Classes:** The model is configured to classify images into 8 distinct categories, the AffectNet targets.

A series of trials were conducted to caliber the ViT's performance across different scenarios. These trials encompassed a variety of setups to identify configurations that maximized the model's efficacy.

ViT-B without pre-training

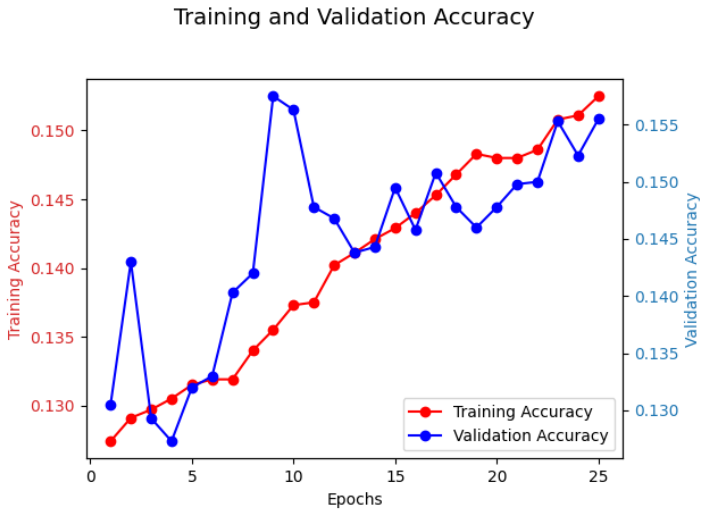

Figure 4.16: Training and validation accuracy for the ViT base model without pre-training over 25 epochs.

The trends displayed in Figures 4.16 and 4.17 provide insights into the learning progression of a ViT base model across 25 epochs. In Figure 4.16, the red line tracks the model's

training accuracy, steadily increasing as the model learns from the training dataset. Conversely, the validation accuracy, represented by the blue line, exhibits considerable variance, implying difficulties in the model's generalization capabilities.

Similarly, Figure 4.17 outlines the trajectory of training and validation loss. The training loss (red line) depicts a declining trend, indicative of the model's learning from the training data. In contrast, the validation loss (blue line) demonstrates considerable fluctuation within a narrow range, suggesting potential overfitting and a struggle to learn generalized patterns; it is clear from the image that, after the 10th epoch, it begins inexorably to rise.

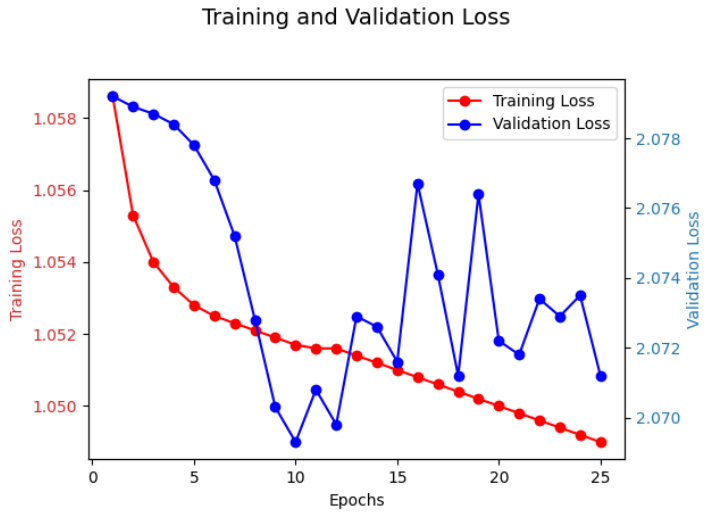

Figure 4.17: Training and validation loss for the ViT base model without pre-training over 25 epochs.

Despite the learning curves' fluctuations, the model does not demonstrate substantial overfitting by the training's conclusion, as indicated by the absence of a pronounced upward trend in validation loss. However, the relatively low accuracy levels underscore the model's limited capacity for generalization, emphasizing the necessity for model refinement, enhanced data diversity, or improved data preprocessing.

The lowest validation loss observed was 2.0789, and the highest validation accuracy reached was 0.1575, both metrics indicating there is significant room for improvement in the model's performance.

For all the reasons just listed, the idea of moving forward with ViT without pre-training was immediately shelved.

Vit-B with pre-training

Figures 4.18 and 4.19 illustrate the nuanced behavior of a ViT base model's training process over 25 epochs. In Figure 4.18, we observe an incremental climb in training accuracy (depicted in red), indicative of the model's capability to learn from the training set. However, the validation accuracy (in blue) shows only modest gains with noticeable volatility, suggesting that the model is struggling to generalize to new data.

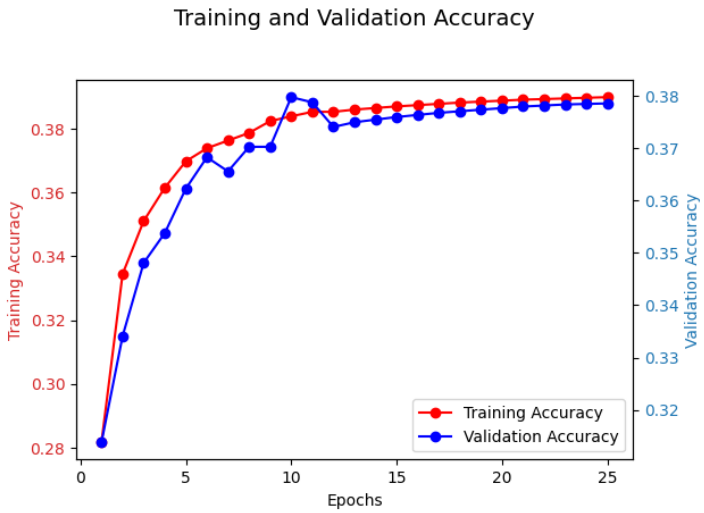

Figure 4.18: Training and validation accuracy for the ViT base model with pre-training over 25 epochs.

Looking at Figure 4.19, the red line representing the training loss follows a downward trajectory, which is a positive indication of the model's learning. Nevertheless, the validation loss, plotted in blue, reveals slight fluctuations, indicating a challenging generalization process. While these oscillations are less pronounced than in previous iterations, they still highlight the model's sensitivity to the training data and its limited generalization to unseen data.

Despite slight improvements, the performance of the model does not mark a significant departure from previous outcomes. The maximum validation accuracy and loss recorded were marginally better than earlier trials but not substantial enough to suggest a breakthrough in the model's performance. The peak validation accuracy stands at 0.3853, with the lowest validation loss at 1.2074, underscoring that while there are marginal improvements, the overall generalization capability of the model remains underwhelming.

Given these observations, it is evident that the ViT base model without pre-training shows only a fractional enhancement in its ability to generalize. This calls for further

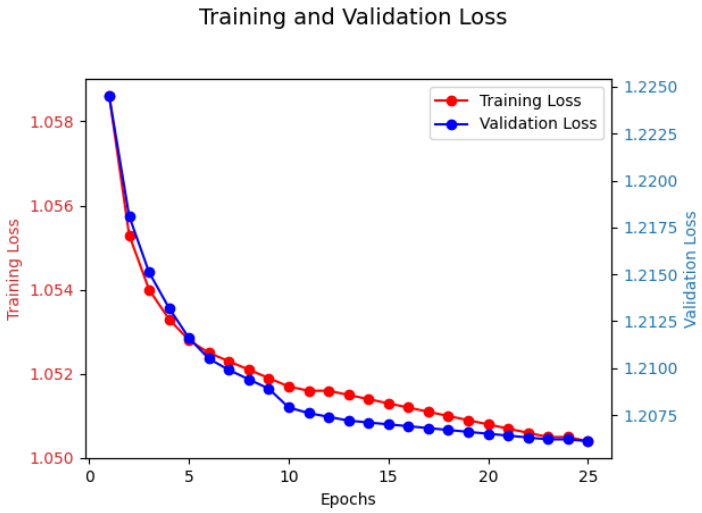

Figure 4.19: Training and validation loss for the ViT base model with pre-training over 25 epochs.

model optimization, potentially through other pre-training, to bolster its performance.

ViT-L with pre-training

As shown in Figure 4.20, the ViT-Large model's training accuracy (red) steadily increases over the epochs, indicating progressive learning and adaptation to the training dataset.

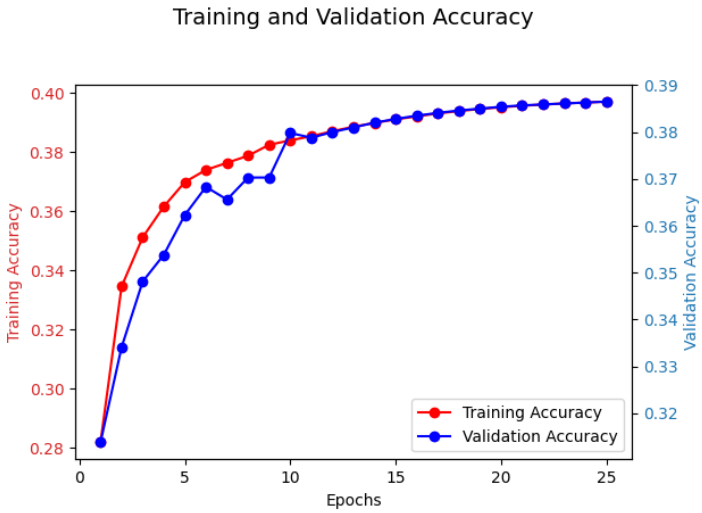

Figure 4.20: Training and Validation Accuracy over 25 epochs for the ViT-Large model with pre-training.

The validation accuracy (blue), while also trending upwards, has a plateau towards the

later epochs, suggesting a ceiling effect in the model's ability to generalize from the provided training data. The highest validation accuracy recorded is 0.3865, reflecting modest but not outstanding generalization performance. Figure 4.21 presents the training and validation loss for the same ViT-Large model. The decreasing trend in training loss (red) is a positive indicator of the model's learning, while the validation loss (blue) shows a downward trend with less variance compared to the training loss. This suggests that the model is starting to stabilize and potentially overfit less. The lowest recorded validation loss is 1.2048, which, while it shows improvement, still leaves room for enhancement in model performance. These figures cumulatively suggest that the ViT-Large model bene-

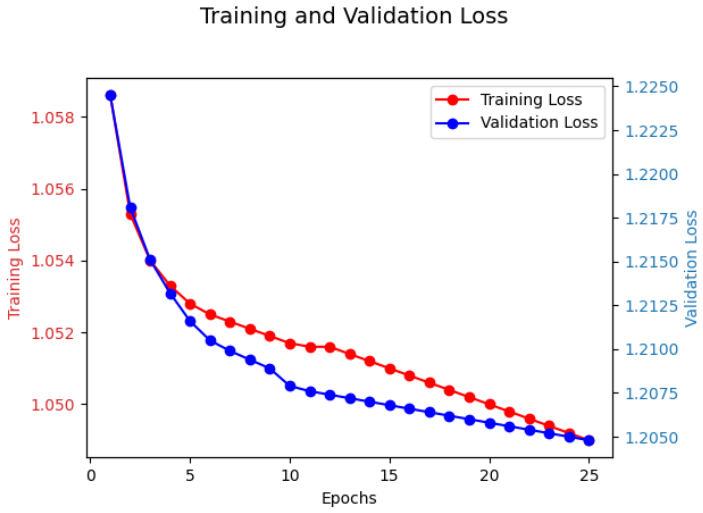

Figure 4.21: Training and Validation Loss over 25 epochs for the ViT-Large model with pre-training model.

fits from increased complexity and capacity, achieving slightly better performance than its base counterpart. Yet, the gains are not as significant as one might expect, underscoring the need for further model tuning or enhanced training strategies.

4.7. Swin ViT

Swin Transformer [33] emerges as a potent variant of the canonical Visual Transformer architecture. Swin ViT distinguishes itself by introducing a hierarchical structure that operates on shifted window partitions, thereby facilitating efficient modeling of local and global image features. This approach enables the model to adaptively process visual cues at multiple scales (Figure 4.22), which is critical for capturing the subtleties of human emotions in images. The following subsections will detail the underlying principles of Swin ViT, its adaptation for emotion recognition tasks, and the specific implementation

nuances that align with the goals of this research.

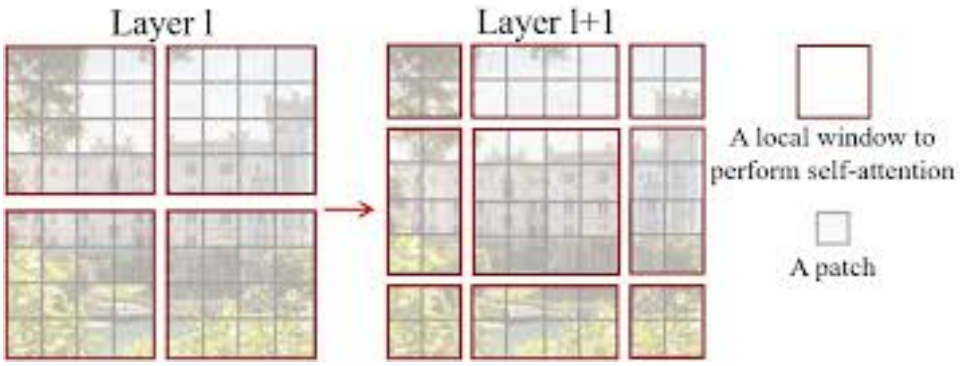

Figure 4.22: The Swin Transformer employs a novel shifted window scheme for self-attention calculations. In the left layer l, a standard window partitioning method is used within which self-attention is computed. In the subsequent layer $l+1$ (on the right), the partitioning is shifted to form new windows that cross the boundaries established in layer l, thereby allowing for inter-window connectivity. [33].

4.7.1. Swin ViT description

Swin ViT overview

The Swin Transformer stands out in the landscape of vision transformers due to its innovative approach to handling the inherent complexities of visual data. Unlike its predecessors such as the Vision Transformer (ViT), which laid the foundation for Transformer architectures in vision by treating non-overlapping image patches as tokens similar to words in NLP, Swin Transformer introduces a novel shifted windowing scheme that maintains local processing while capturing global context, thereby enhancing efficiency and scalability.

The original ViT, while groundbreaking, demonstrated optimal performance primarily on large-scale datasets like JFT-300M [14]. However, it has become evident that the ViT architecture has limitations as a general framework for dense vision tasks, particularly at high input resolutions, due to the quadratic increase in complexity with respect to image size [33].

Swin Transformer addresses these issues by producing hierarchical feature maps with a design that adapts to various scales and resolutions, as it can be seen in figure 4.23. This approach mitigates the computational burden, particularly in tasks requiring high-resolution feature maps, such as object detection and semantic segmentation. The architecture's efficiency is not just theoretical but is evidenced by its superior speed-accuracy trade-off in comparison to other models [33].

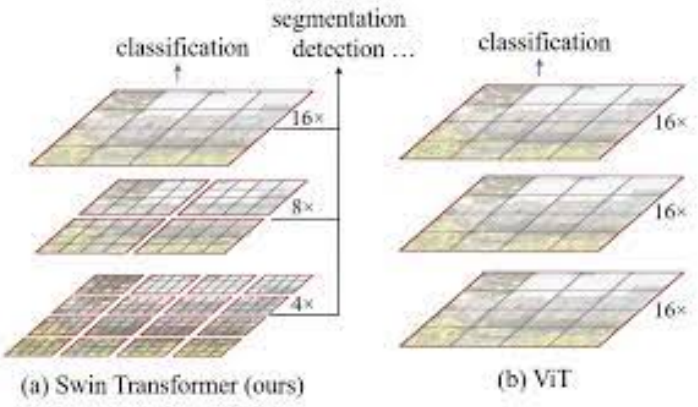

Figure 4.23: (a) Swin Transformer: Hierarchical features are formed through merging patches (gray) and localized self-attention windows (red), ensuring linear computational complexity. Suitable for various vision tasks. (b) Prior Vision Transformers: Single-resolution feature maps with global self-attention, resulting in quadratic computational complexity. [33].

Several concurrent works to the Swin Transformer explore variations in Transformer architecture for image classification, such as the works by Liu et al. [61], Touvron et al. [9], and Wang et al. [20]. However, the Swin Transformer has been empirically found to outperform these in terms of speed-accuracy trade-off for general-purpose vision tasks. Another concurrent work by Wu et al. [54] also attempts to construct multi-resolution feature maps using Transformers (Figure 4.24), but it retains the quadratic complexity in relation to image size. In contrast, Swin Transformer's complexity increases linearly with image size and operates locally, which is advantageous for modeling the highly correlated nature of visual signals.

Each of these models brings a unique perspective to the challenge of processing visual data with Transformer architectures, contributing to the continuous evolution of the field. The selection of the most appropriate attention mechanism often depends on the specific requirements of the task at hand, whether it's the need for fine-grained detail in object detection or the broad context required for scene understanding.

Incorporating the Swin Transformer into Facial Emotion Recognition (FER) within this work is based on the premise that its hierarchical, high-resolution image processing capabilities may offer nuanced insights into the complexities of human expressions.

It is posited that the Swin Transformer's architecture, with its ability to merge image patches at various depths (as can be seen in Figures 4.25), may be particularly suited to capture the full range of facial expressions that signify different emotions, from the most

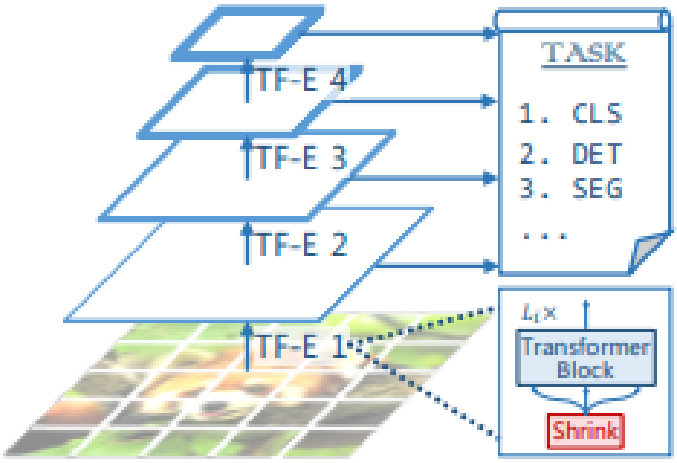

Figure 4.24: This schematic illustrates the multi-level architecture of a pyramid vision transformer, with each layer (TF-E 1 to TF-E 4) representing a stage of feature extraction for various tasks like classification (CLS), detection (DET), and segmentation (SEG).

conspicuous to the minutely subtle.

The architecture's shifted window mechanism, which cleverly combines local and global image features, is also anticipated to be beneficial. It is hoped that this will allow for a more precise discernment of marked and well-defined expressions, which are often the key to accurately identifying and classifying complex emotional states.

The subsequent sections will delve into the empirical findings to ascertain the validity of these expectations and to evaluate the actual performance of the Swin Transformer in the challenging domain of FER.

Architecture of Swin ViT

The Swin Transformer architecture, depicted in Figure 4.26, represents a significant shift in vision transformer design. It starts by dividing an input RGB image into fixed-size patches using a patch splitting module similar to ViT, treating each patch as a "token". The feature of each token is derived from the concatenation of the RGB pixel values within the patch. For instance, with a patch size of 4×4, the feature dimension is $4 \times 4 \times 3 = 48$. These raw features are then projected to a higher dimension, denoted by C, through a linear embedding layer.

The Swin Transformer confronts this challenge by partitioning the input image into a grid of non-overlapping windows and then computing self-attention within these local

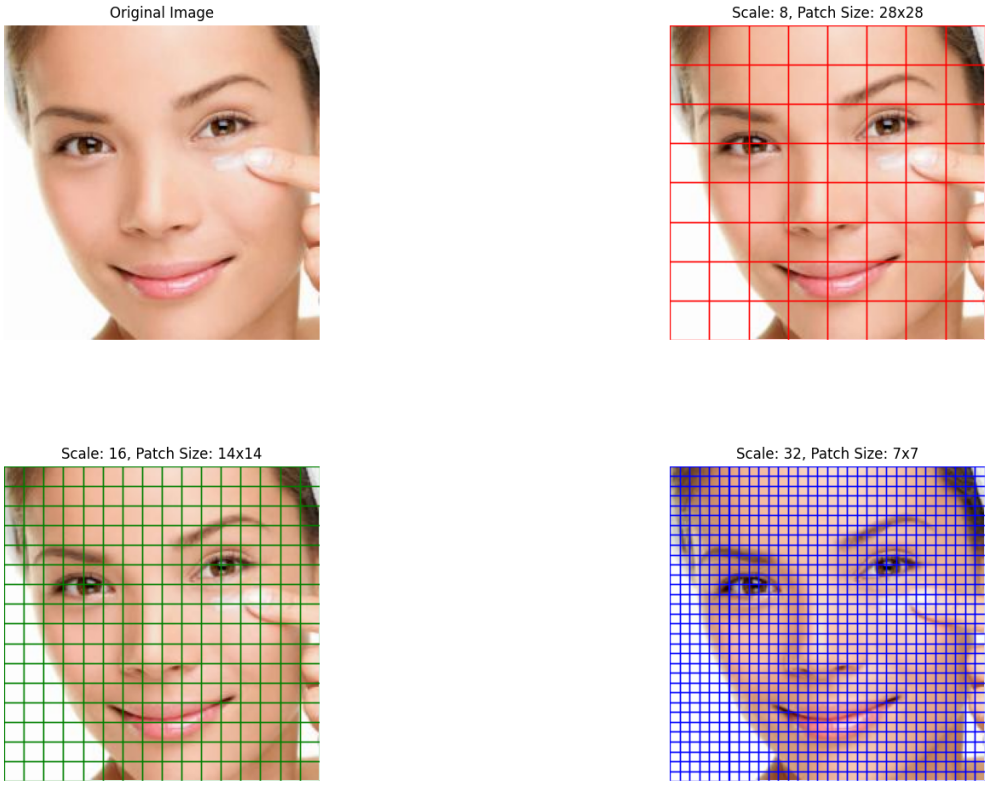

Figure 4.25: A composite figure with original and Swin Transformer patch divided images.

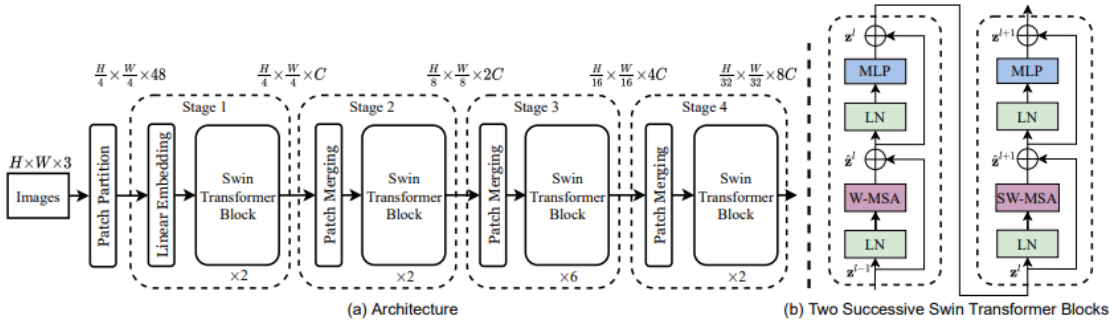

Figure 4.26: Swin ViT architecture.

windows. This design reduces computational complexity from quadratic to linear when compared to global self-attention, as shown in the following equations:

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C, \tag{4.4}$$

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC, \tag{4.5}$$

where hw represents the total number of image patches, C is the feature dimension of each patch, and M is the size of the window.

The Swin Transformer block, as implemented in Python, is reflected in the mathematical formalism of the model's alternating partitioning strategy. Each block performs a set of operations, beginning with layer normalization (`LayerNorm`), followed by a shifted window attention mechanism (`ShiftedWindowAttention`), and culminating with a multi-layer perceptron (MLP). The stochastic depth component (`StochasticDepth`) is employed as a regularization technique to improve training efficiency.

For example, a typical Swin Transformer block is structured as follows:

```
SwinTransformerBlock(
  (norm1): LayerNorm((128,), eps=1e-05, elementwise_affine=True),
  (attn): ShiftedWindowAttention(
    (qkv): Linear(in_features=128, out_features=384, bias=True),
    (proj): Linear(in_features=128, out_features=128, bias=True)
  ),
  ...
  (mlp): MLP(...)
)
```

This block corresponds to the equations that articulate the flow of features through the transformer's layers:

$$\begin{aligned} \hat{z}_l &= \text{W-MSA}(\text{LN}(\hat{z}_{l-1})) + \hat{z}_{l-1}, \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l, \\ \hat{z}_{l+1} &= \text{SW-MSA}(\text{LN}(z_l)) + z_l, \\ z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1}, \end{aligned} \tag{4.6}$$

In these equations, $\hat{z}_l$ and z_l are the outputs after the self-attention and MLP modules, analogous to the outputs from the `attn` and `mlp` components in the Python implementation. The operation W-MSA or SW-MSA is effectively performed by the `ShiftedWindowAttention` layer within the block.

The Swin Transformer's hierarchical structure, featuring a series of stages where each stage progressively merges patch tokens to form a pyramid of embeddings, is particularly noteworthy.

4.7.2. Swin ViT results

The guidelines for the configuration setup are thoughtfully proposed by Ze Liu in [33], ensuring an optimized balance between computational efficiency and model performance. Herein, we delineate the pivotal elements of the configuration:

- **patch_size**: Central to the Swin Transformer's initial processing phase, it governs the subdivision of the input image into discrete patches. In the given scheme, a 2×2 patch size is utilized, implying that the image is partitioned into patches of these precise dimensions prior to their propagation through the transformer's layers.
- **embed_dim**: This parameter stands for the dimensionality of the token embeddings, with the current configuration embedding each token in a 96-dimensional vector space, fostering a rich, expressive representation of the input data.
- **depths**: An array delineating the number of transformer blocks allocated to each stage within the model. The configuration specifies a sequential depth arrangement of 2, 2, 6, and 2 blocks, thereby structuring the transformer's depth in a hierarchical manner.
- **num_heads**: This array determines the multiplicity of attention heads within each transformer layer, enhancing the model's ability to focus on various segments of the input data simultaneously. The configuration progressively scales the number of heads from 3 to 24 across the stages, amplifying the model's attention capacity as

it delves deeper.

- **window_size**: A defining feature of the Swin Transformer, the **window_size** prescribes the dimensions of the window within which self-attention is computed. The chosen 2×2 window size indicates that self-attention mechanisms operate within these localized clusters of patches.
- **mlp_ratio**: This ratio benchmarks the scale of the MLP layer's hidden dimension against the token embeddings' dimensionality. With an **mlp_ratio** of 4, the model ensures that the MLP's hidden layer is quadruple the size of the embedding space, providing ample capacity for feature transformation.
- **dropout**: A regularization mechanism, the **dropout** rate is set to 0.4, strategically deactivating a subset of the input units at random during training, thus mitigating the risk of overfitting.
- **attention_dropout**: Complementing the overall dropout strategy, the **attention_dropout** rate is similarly assigned a value of 0.4, targeting the attention weights specifically and promoting a robust attention mechanism.
- **stochastic_depth_prob**: The probability assigned to the stochastic depth process, pegged at 0.1, incorporates an element of randomness in layer retention during training, which not only expedites the training process but also bolsters the model's generalization prowess.
- **num_classes**: Reflecting the model's versatility in classification tasks, the parameter is configured to accommodate 8 distinct categories, enabling the model to differentiate a diverse range of inputs effectively.

In the scope of the discussed application, a fine-tuning strategy has been implemented, representing a specialized form of transfer learning. This approach enables the adaptation of a model, initially trained on an extensive and varied dataset, to a more specialized task with minimal retraining. The model employs the **Swin_T_Weights** from the **IMAGENET1K_V1** collection, which are weights pretrained on the ImageNet-1K dataset, known for its breadth and diversity [12].

Throughout the fine-tuning phase, the parameters inherited from the pretrained layers were fixed, preserving the learned features from the ImageNet-1K dataset. The training efforts were concentrated on the final classification layers, which were tailored to classify eight specific categories pertinent to the task at hand. This strategic fusion of retained pre-existing knowledge with targeted adaptation underscores the model's adaptability and the overarching objective to harness cutting-edge technology for refined image recognition

tasks.

Figures 4.27 and 4.28 together provide a comprehensive picture of the model's learning trajectory over 24 epochs. In Figure 4.27 showcases a similar rapid advancement in the initial stages for both training and validation accuracy. The training accuracy, charted in red, swiftly ascends, suggesting effective learning, while the validation accuracy, plotted in blue, also climbs but plateaus after the first ten epochs, reaching the maximum value of 0.3416. This plateauing effect may signal that the model is reaching the limits of its ability to generalize based on the current training data. The leveling out of both the loss and accuracy curves points to a model that is well-fitted, without obvious signs of overfitting, where training accuracy would continue to increase, and validation accuracy would begin to decrease.

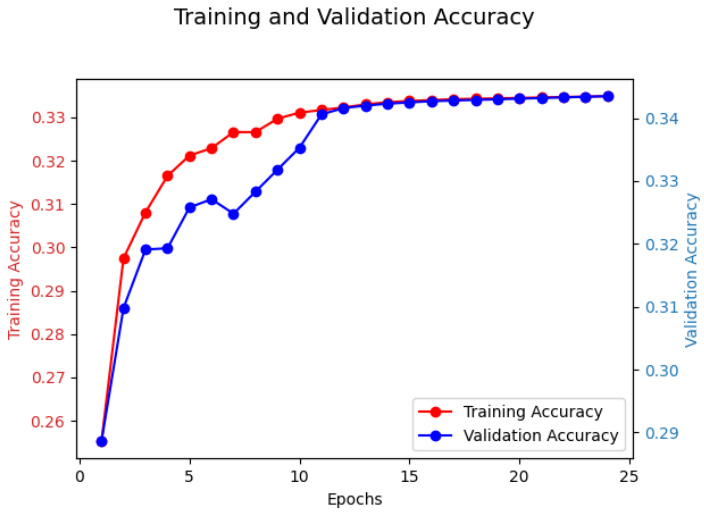

Figure 4.27: Training and Validation Accuracy over 24 Epochs. The graph displays the progression of the training and validation accuracy across the epochs, with the training accuracy shown in red and the validation accuracy in blue.

At the same time, in Figure 4.28, we observe a pronounced initial decrease in both training and validation loss, with the training loss depicted in red and the validation loss in blue. This sharp descent is indicative of the model quickly assimilating the patterns within the training data. As the epochs advance, this reduction in loss decelerates and levels off, signifying the model's convergence towards optimal parameters.

The slight undulations seen in the later epochs across both metrics may suggest areas for model tuning, such as adjusting the learning rate, increasing regularization, or modifying the model architecture. To push the model beyond its current generalization ceiling, exploring strategies like data augmentation, incorporating more data, or architectural

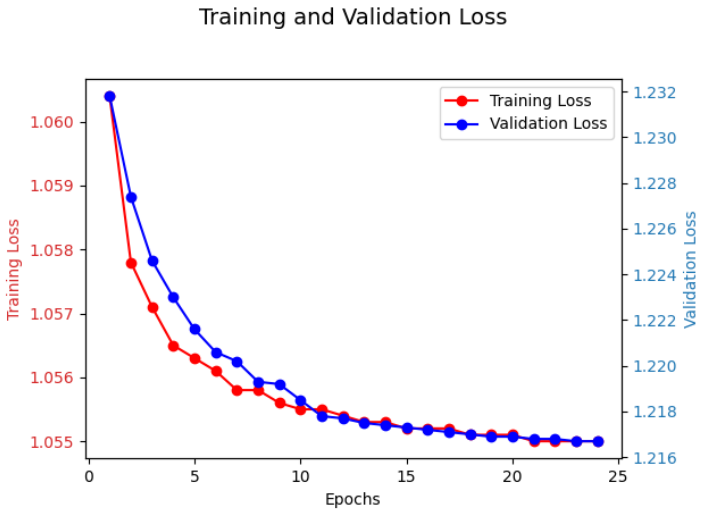

Figure 4.28: Training and Validation Loss across 24 Epochs. The graph delineates the downward trend of the training loss (in red) and validation loss (in blue) as the number of epochs increases.

changes may yield further improvements. Monitoring both loss and accuracy is crucial for diagnosing model performance and ensuring that the model not only learns effectively but also generalizes well to new, unseen data.

The initial conception of utilizing the Swin Transformer as a detector for emotions and higher-resolution patterns in images was based on its architectural promise to handle varying scales and complexities within visual data efficiently. The Swin Transformer's design to compute self-attention within local windows and its shifted windowing approach allows for capturing finer details, which is essential for tasks like emotion detection where subtle facial cues can be significant.

However, upon evaluation, the data suggests that the expectation for the Swin Transformer to outperform ViT in detecting emotions and high-resolution patterns has not been fully substantiated. While the Swin Transformer has shown proficiency in handling images, the anticipated leap in performance for emotion detection and resolution-specific tasks has not been evident.

Several factors may contribute to this outcome. The uniformity in performance across epochs, as indicated by the plateau in the accuracy and loss graphs, suggests that while the model is learning, it's not doing so at a pace or scale that outmatches existing benchmarks or expectations. This plateau may signify a potential underutilization of the Swin Transformer's capabilities or a need for further optimization in its application to emotion detection and high-resolution image processing.

The challenge of detecting emotions from images lies not only in recognizing patterns but also in interpreting the context and subtleties that are often subjective and nuanced. Therefore, the Swin Transformer's current application might require additional layers of complexity, such as more sophisticated data pre-processing, feature engineering, or integration with other modalities that can offer context and depth to the emotional analysis.

In conclusion, while the Swin Transformer remains a powerful and flexible architecture for a range of vision tasks, its adoption for emotion detection and the discernment of high-resolution patterns has not demonstrated the anticipated breakthrough.

4.8. Two Stream Model

The "Two Stream Model" presents an innovative architecture aimed at advancing the analysis of facial features [34, 57]. Departing from traditional convolution-dependent methodologies, this model incorporates a comprehensive attention mechanism across all stages of feature extraction. The emphasis on attention-based processing affords a granular and autonomous analysis of facial features, thereby refining the system's capability to interpret facial expressions with minimal reliance on convolution.

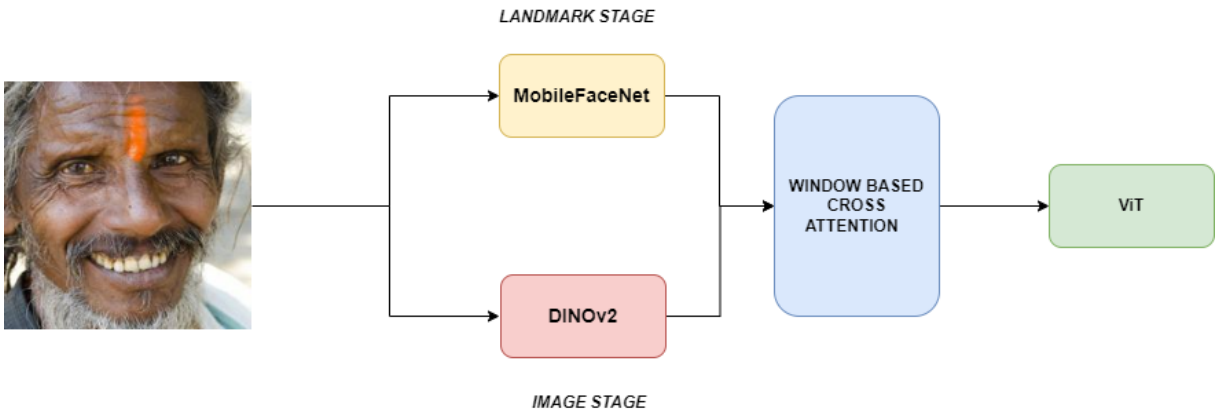

Figure 4.29: Two stream model architecture.

To get a clearer idea of the architecture, we refer to Figure 4.29.

Two distinct yet parallel streams constitute the model's foundation. The primary stream is dedicated to identifying facial landmarks with precision, whereas the secondary stream undertakes a direct examination of the image data. This bifurcation of tasks enables the deployment of attention modules tailored to each stream, optimizing the extraction of features. The landmark-focused stream ensures the accurate detection of critical facial points, paving the way for the image-focused stream to detect subtle pixel-level features

potentially overlooked by conventional approaches.

A multi-scale feature extraction strategy is implemented, capitalizing on the strengths of both the facial landmark detector and the image backbone. This dual-pronged approach ensures that a comprehensive set of features is collected, ranging from macro-level attributes discerned by the image backbone to the nuanced details captured by the facial landmark detector. By extracting features at various scales, the system is aimed at recognizing both overarching patterns and subtle nuances of facial structures.

A synthesis of the insights gathered from both streams is achieved through window-cross-attention modules, amalgamating the knowledge from each pathway. The convergence of these streams progresses towards the Vision Transformer (ViT) phase, which integrates the distinct features into a comprehensive and cohesive output. This fusion results in a framework designed to revolutionize facial recognition and emotion detection paradigms.

4.8.1. Two Stream Model description

Landmark stage

In the landmark detection stage of the proposed architecture, the MobileFaceNet model [7] is employed as a facial landmark detector. The model's parameters are set to a non-trainable state, meaning that the weights are frozen. This approach leverages the robust feature extraction capabilities of MobileFaceNet, trained on comprehensive datasets, ensuring the landmarks detected are both accurate and reliable for subsequent processing stages.

The landmark detection is not just a preliminary step but a crucial component of the model, aiming to guide the subsequent image feature extraction. By accurately pinpointing 39 key facial points (Figure 4.30), the model can focus on salient regions relevant to facial expressions, thus addressing the challenge of intra-class variability inherent in Facial Expression Recognition (FER) tasks. This focus on expressive features helps to reduce the model's sensitivity to non-essential variations within the same class, a significant hurdle in achieving accurate FER.

Initiating with a primary convolutional block, the model rapidly condenses the input data, setting the stage for more granular feature extraction. This is followed by a depth-wise convolutional layer that maintains the integrity of the features while reducing computational complexity, thanks to group convolutions.

As the data progresses through the network, it encounters a series of residual blocks. These blocks employ depth-wise separable convolutions—a sophisticated technique that

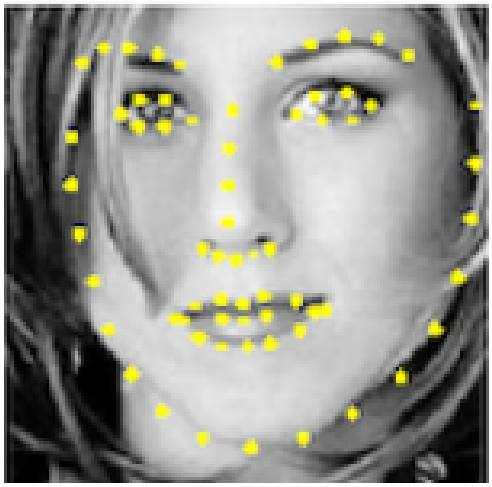

Figure 4.30: Example of facial landmarks extracted from MobileFaceNet. As it can be seen, the extracted points are 39.

decomposes a standard convolution into a depth-wise spatial convolution followed by a point-wise convolution. This not only improves the efficiency of the model, but also enriches feature representation without inflating the parameter count.

Further into the network, the MobileFaceNet introduces additional residual stages, each including multiple blocks that incrementally refine the features. These stages not only capture the complexities of facial structures with varying granularity, but also imbue the model with the ability to learn and retain critical spatial hierarchies within the facial data.

The culmination of this process is through the final convolutional block, which aggregates the diverse array of features into a high-dimensional space. Depending on the configuration, an output layer (either Global Norm-Aware Pooling-GNAP or Global Depthwise Convolution-GDC) is employed to distill these high-dimensional features into a more concise and informative embedding, tailored to the requirements of the task at hand.

Throughout this process, the model's weights are initialized using the Kaiming initialization method. *Kaiming Initialization*, or *He Initialization*, is an approach designed for neural networks that accounts for the non-linearity of activation functions, like ReLU activation. It avoids the issue of reducing or magnifying the input signal's magnitude exponentially. The initialization is performed as follows: weights w_l are drawn from $\mathcal{N}(0, 2/n_l)$, where n_l is the number of input neurons in the layer, resulting in weights that have zero mean and a standard deviation of $\sqrt{2/n_l}$, as derived from the condition $\frac{1}{2}n_l\text{Var}[w_l] = 1$. This choice of initialization ensures that the variances in activation are maintained across layers, promoting a stable gradient flow. Biases are initialized to zero,

further contributing to the stability of the training dynamics.

Moreover, the sparsity of the landmark features allows the model to avoid over-emphasizing facial regions, which can lead to inter-class confusion. As a result, the model is better positioned to distinguish between different facial expressions, even when they share similar attributes. The landmark-to-image branch is an essential part of this process, as it ensures that the image features are informed and enhanced by the landmark detection, as evidenced by the attention mechanisms visualized in the model.

The addition of the image-to-landmark branch, where image features interact with landmark features, offers a compensatory mechanism to enrich the representational capacity of the model. This interaction helps to fill in the gaps where landmark features may be lacking, providing a more detailed understanding of facial expressions. However, this comes at the cost of increased computational resources.

The MobileFaceNet stream processes the input data through various layers, producing outputs in the form of tensors that capture features at different levels of abstraction. The first output tensor, denoted as x_{face1}, has a shape of $[16, 16, 16, 3]$, which corresponds to a batch of 16 feature maps, each with a spatial resolution of 16×16 pixels and 3 channels. This tensor likely represents preliminary feature detection, capturing basic visual elements.

The second tensor, x_{face2}, is reduced to a shape of $[16, 8, 8, 12]$. The halving of spatial dimensions and the increase in channels to 12 suggest a progression to more complex feature extraction, where the network begins to condense and encode higher-level patterns within the images.

Finally, the third tensor, x_{face3}, exhibits dimensions of $[16, 4, 4, 48]$, indicating a further reduction in spatial size coupled with a substantial increase in the depth to 48 channels. This tensor encapsulates the most abstract and detailed aspects of the input, representing the network's most in-depth processing layer, distilling the essential characteristics necessary for subsequent recognition tasks.

The just-described output vectors will then be used as input in the window-based cross attention layer and so analyzed together with the image stage output.

Image stages

The architecture of the original proposed system [34, 57] is underpinned by the IR50 network [13], serving as the image backbone. The "IR" stands for "InsightFace Recognition," and "50" usually denotes the number of layers in the neural network.

The IR50 model is part of the family of ResNet-like architectures (Figure 4.31), where the depth of the network (in this case, 50 layers) is meant to provide a good balance between complexity and performance.

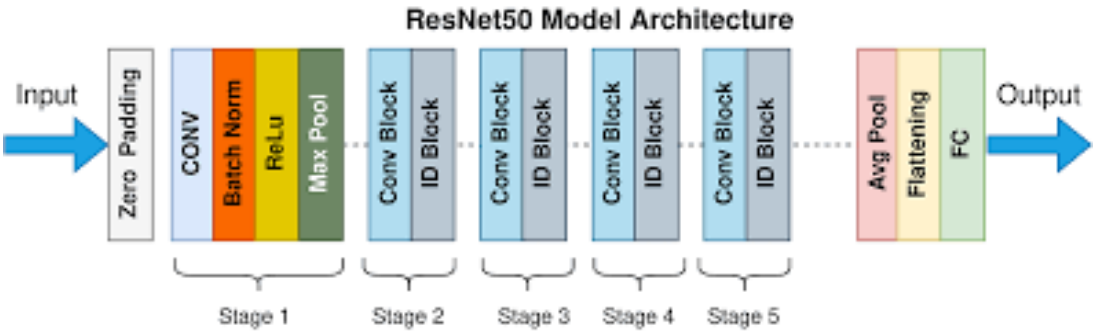

Figure 4.31: ResNet50 architecture.

IR50's feature extraction benefits significantly from pre-training on the comprehensive Ms-Celeb-1M dataset, which includes a wide variety of facial images from numerous celebrities. This pre-training ensures a broad understanding of facial features within the model, which is crucial for complex facial analysis tasks. As such, the IR50 backbone is integral to the pipeline, effectively distilling essential features from raw image data. These features are then leveraged in later stages of the model to enhance facial recognition and analysis capabilities.

Aligning with cutting-edge methodologies that favor attention-based mechanisms over traditional convolutional approaches, the IR50 backbone has been supplanted by the **DINOv2 (DIstillation with NO labels version 2.)** network [6]. The shift to an attention-centric model capitalizes on the attention mechanisms' advanced capability to selectively focus on the most relevant features in an image. DINOv2, thus, serves as the new image backbone, supplanting the convolutional framework with a versatile and dynamic attention-based architecture. This transition significantly augments the model's adeptness in facial recognition and emotion detection tasks.

In brief, by integrating DinoV2 into the proposed system, there is a distinct shift towards a fully attention-based model that promises not only to refine the facial recognition and emotion detection capabilities but also to revolutionize the way visual information is processed and interpreted.

The inception of DinoV2 aligns with the recent paradigm shift in NLP, where models trained on large quantities of data without task-specific supervision have shown exceptional performance on downstream tasks. DinoV2's self-supervised learning framework has demonstrated similar potential in learning robust, all-purpose visual features. By

leveraging larger datasets and eschewing the constraints of text-image alignment, DinoV2 focuses on learning directly from raw pixel data, capturing both image-level and pixel-level information effectively.

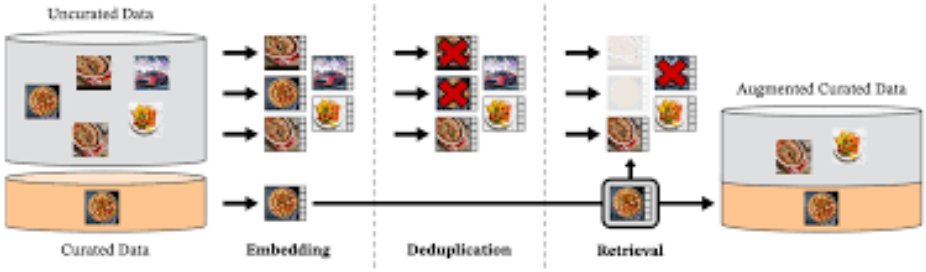

Figure 4.32: DINOv2 data processing pipeline proposed by Oquab et al. in [6].

The technical advancements within DinoV2 have facilitated a more efficient training process, making it both faster and less resource-intensive compared to its predecessors. Moreover, its training pipeline, inspired by NLP methodologies, ensures a balanced and diverse dataset, crucial for high-quality feature production. The end result is a suite of DINOv2 models, trained across various ViT architectures, which have been validated on numerous computer vision benchmarks. These models excel in both image and pixel-level tasks, positioning DinoV2 as a competitive alternative to the best weakly-supervised models available [6].

In the pursuit of enhancing the quality and diversity of datasets for machine learning, a sophisticated data assembly process was undertaken as depicted in Figure 4.32. This process aimed to compile the LVD-142M dataset [6], curated to aid in the training of vision models. LVD-142M refers to a large-scale dataset assembled for the purpose of pretraining and improving the performance of machine learning models, particularly those in the field of computer vision. The acronym "LVD" stands for "Large Visual Dataset," and the number "142M" indicates the number of images it contains, which is 142 million.

The inception of this dataset started with the collection of a vast array of uncurated images sourced from publicly available web data. From this pool, images were extracted based on their presence in <img> tags within web pages, while taking care to exclude any URLs linked to unsafe content or restricted domains. The subsequent post-processing steps included the elimination of near-duplicates through PCA hash deduplication, filtering out non-safe-for-work (NSFW) content, and blurring identifiable features in faces, ultimately resulting in a unique collection of 1.2 billion images.

The core of the process involved a self-supervised image retrieval system. This system

utilized a Vision Transformer (ViT-H/16) pre-trained on ImageNet-22k to generate embeddings for each image, allowing for a cosine-similarity-based comparison. Through this method, images closely matching those from existing curated datasets were selected. The selection process was fine-tuned using k-means clustering, and depending on the size of the query dataset, a specified number of nearest neighbors or a set number of images from each cluster were chosen. This selection was further refined through visual inspection to ensure the quality of the retrieved images.

This step was critical in fostering the dataset's diversity. The deduplication process not only targeted the removal of near-identical images within the uncurated set but also excluded any images that resembled those from the test or validation sets of benchmarks relevant to this research.

DinoV2, as mentioned in the official documentation [6], represents a leap forward in the domain of visual feature extraction. It stands at the forefront of the self-supervised learning revolution, avoiding the need for labeled datasets by learning from the images themselves.

DINOv2 starts with the `PatchEmbed` layer, employing a convolutional operation via `Conv2d` to transmute input images into a sequence of flattened patches. Subsequently, these patches are embedded into a 768-dimensional vector space. Uniquely, the model foregoes additional normalization in the embedding layer, as indicated by the `Identity` layer in the normalization attribute.

At its core, the `DINOv2` includes 12 `NestedTensorBlocks`, each constituting several pivotal elements:

- `LayerNorm` stabilizes the feature distribution, a critical factor for the model's training and convergence.
- `MemEffAttention` is a memory-efficient attention module, inclusive of linear projections for queries, keys, and values, culminating in the attention operation and a subsequent projection to finalize the attention output.
- `LayerScale` introduces a technique that enhances the training dynamics by scaling the weights within the attention and MLP layers.
- `Mlp`, a multilayer perceptron, augments the computational capability with two dense layers interspersed with a GELU activation.

The `Identity` layers, referred to as `drop_path1` and `drop_path2`, imply the absence of stochastic depth, thus maintaining all pathways active during the model's training. This

design choice might be intentional to preserve the full signal flow strength.

Posterior to the block sequence, a `LayerNorm` normalizes the global features, preparing them for the final output phase. Notably, the model's head is replaced with an `Identity` layer, indicating no further transformation is applied to the block outputs. Therefore, the model's output retains the transformer blocks' feature vector sequence, each with a 768 dimensionality.

Similar to the landmark stage, the feature extraction process in the proposed architecture is conducted at three distinct depth levels.

The DINOv2 model, offers a solid foundation for various visual recognition tasks. In the context of facial expression recognition (FER), it is hypothesized that an additional pre-training phase of the DINOv2 model on a specialized facial dataset, such as the Large-scale Celeb Faces Attributes (CelebA) [32], may potentially enhance its effectiveness. Such targeted pre-training is presumed to be beneficial, tailoring the model's inherent strengths specifically towards the nuances of facial feature detection and expression analysis, which are pivotal for accurate FER.

Within the landscape of available facial datasets, CelebA emerges as a particularly detailed repository, featuring over 200,000 celebrity images characterized by a rich diversity of poses, expressions, and backgrounds. Some examples are shown in Figure 4.33. The dataset's breadth is further demonstrated by the presence of 40 binary attribute annotations for each image, alongside precise locations for 5 facial landmarks. These attributes span a wide range of facial features and expressions, providing a granular level of detail suitable for nuanced analysis and model training.

The landmark annotations are pivotal for tasks that require precise facial feature localization, such as facial part recognition and expression synthesis, making CelebA a robust foundation for developing advanced face recognition technologies.

Moreover, the dataset's extensive coverage of various facial attributes makes it an ideal candidate for training models to recognize a spectrum of facial features, which is crucial for applications like biometric authentication, surveillance or FER tasks, i.e., our target. Its utility is exemplified in the realm of deep learning, where it supports the refinement of algorithms for complex tasks like facial attribute detection and landmark localization.

This specialized pre-training is expected to yield a version of DINOv2 that not only possesses the original model's adaptability across various visual tasks but also exhibits enhanced precision in recognizing and interpreting emotions.

The fine tuning process was diligently executed across 50 epochs, culminating in a com-

Figure 4.33: Images from CelebA dataset.

mendable training set accuracy of 91.43%. Such a figure robustly positions DINOv2 as adequately trained for the designated task at hand.

Windows-Based Cross Attention

Once the inputs from the previous two steps have been obtained, we move on to the penultimate layer of the network.

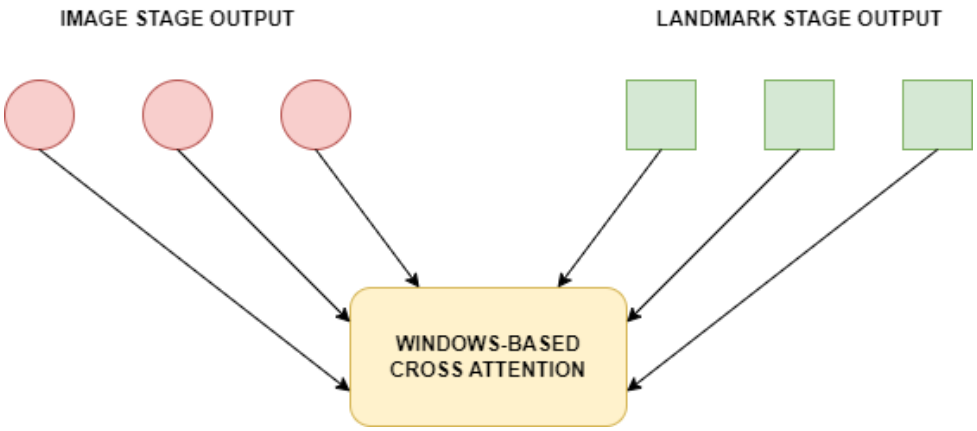

Figure 4.34: Input representation for Window-Based Cross Attention layer.

As seen in Figure 4.34, this layer takes the previous outputs as input. The ultimate goal of this operation is to enhance the model's ability to process images by employing a window-based cross-attention mechanism. This approach has two main advantages:

- **Localized Attention Processing**: The window-based approach restricts the attention mechanism to operate within local windows instead of considering the entire image at once. This focuses the model's processing power on smaller segments of the input, making the computation more efficient and manageable.
- **Reduced Computational Complexity**: By calculating attention within these windows, the model reduces the quadratic complexity typically associated with standard attention mechanisms. This reduction is particularly beneficial for processing large images, where global attention would be computationally expensive.

A Window-Based Multi-Head Cross-Attention (W-MCSA) mechanism not only facilitates efficient processing of image features, but also ensures that the attention is contextually relevant within each local window.

In the context of the proposed architecture, W-MCSA allows focused processing within localized regions of the input image, mitigating the computational expense associated with traditional cross-attention mechanisms that operate on the entire image. This strategic focus on localized regions enhances the model's ability to capture fine-grained details pertinent to the specific tasks of facial recognition and emotion detection.

The mechanism begins by defining the dimensions of the attention windows and the relative coordinates within these windows. The positional bias is calculated to reflect the relative position of each token within a window, which is crucial for the model to understand the spatial relationships between different parts of the image. The attention operation within each window is then defined, utilizing a scaled dot-product attention mechanism with learned query, key, and value projections. This local attention operation is repeated for each window across the image, aggregating local features while significantly reducing computational demands.

Formally, W-MCSA operates in the following way. The window-based attention within each window can be expressed by the equations below, where w_q, w_k, w_v, and w_o are the weight matrices used for projecting the queries, keys, and values, and b is the relative position bias matrix.

$$q = z_{lm} w_q, \quad k = z_{img} w_k, \quad v = z_{img} w_v, \tag{4.7}$$

$$o^{(i)} = \text{softmax}\left(\frac{q^{(i)}(k^{(i)})^T}{\sqrt{d}} + b\right) v^{(i)}, \quad i = 1, \ldots, I, \tag{4.8}$$

$$o = [o^{(1)}, \ldots, o^{(I)}] w_o, \tag{4.9}$$

These equations are invoked for every window, thereby defining the Window-based Multi-head CrosS-Attention (W-MCSA). Consequently, the cross-fusion transformer encoder in the advanced model is formalized as follows:

$$X'_{img} = \text{W-MCSA}(img) + X_{img}, \tag{4.10}$$

$$X^{o}_{img} = \text{MLP}(\text{Norm}(X'_{img})) + X'_{img}, \tag{4.11}$$

where X'_{img} represents the features after W-MCSA, X^{o}_{img} is the final output from the transformer encoder, and $Norm(\cdot)$ denotes the normalization function applied across the combined image features from all windows.

In the architecture of the proposed system, attention mechanisms are tailored to different scales by initializing the `WindowAttentionGlobal` module with a varied configuration of heads and window sizes.

The first window attention layer is configured with a window size of 16×16 and a single attention head, suitable for capturing broader patterns across larger spatial extents. As we progress deeper, the second layer's window size reduces to 8×8 with two attention heads, increasing the focus and allowing for the detection of more nuanced features. The deepest layer, with a window size of 4×4 and four attention heads, is poised to capture the most detailed and fine-grained information.

These configurations are encapsulated in the dimensions array (`dims`), which, in conjunction with the number of heads array (`num_heads`), dictates the dimensionality of the feature space for each corresponding head.

More specifically, the `WindowAttentionGlobal` module was created as follows.

Each attention head's dimensionality (`head_dim`) is calculated by distributing the total dimension size (`dim`) evenly across all heads. The scaling factor (`scale`), derived from a predetermined scale or the square root of the head dimension, adjusts attention scores to stabilize learning and mitigate overly large values from dot products.

A learnable parameter, the `relative_position_bias_table`, is initialized to integrate positional context into the attention framework. This bias table, filled initially with zeros, adapts during training to interpret the relative placements within the attention window.

Grids of coordinates are constructed for both height (`coords_h`) and width (`coords_w`), forming a meshgrid, subsequently flattened and reconfigured to ascertain relative coordinates. These are employed to index into the `relative_position_bias_table`, aiding in

generating position-informed attention scores.

Linear transformation layers (`qkv`) are established to convert input features into the attention mechanism's queries, keys, and values. To combat overfitting, dropout layers (`self.attn_drop` and `proj_drop`) are incorporated, randomly nullifying a portion of the features during training.

The softmax function normalizes attention scores, transforming them into a probabilistic distribution that indicates the input areas to which the model should allocate more focus. This preparation lays the groundwork for a detailed, position-sensitive attention system, capable of detecting intricate patterns with enhanced precision and efficiency.

Visual Transformer

The last layer is totally devoted to the Visual Transformer, an architecture that has already been extensively discussed in Section 3.1.2.

- **Vision Transformer from POSTER++ [34]**:
  - Incorporates *Squeeze-and-Excitation (SE)* and *Efficient Channel Attention (ECA)* blocks to refine the feature maps, indicating an emphasis on feature recalibration.
  - Utilizes a *Conv2d* layer for patch embedding, transforming input images into a sequence of embeddings to capture spatial hierarchies.
  - Features a dedicated *Classification Head* for direct application in classification tasks, indicative of an end-to-end trainable model.
  - Contains additional convolutional and linear layers (CON1, IRLinear1, IRLinear2), suggesting an intricate feature transformation process prior to attention mechanisms.
- **Original Vision Transformer [14] with no projection**:
  - Lacks the *Conv2d* patch embedding layer, implying a dependency on preprocessed inputs or an alternative representation of input data.
  - Exhibits a simplified encoder with a lower feature dimension (256 vs. 768), pointing to a streamlined model with potentially lower computational demands.
  - Adjusts the dropout rates to a slightly higher level (0.2 vs. 0.0), possibly to mitigate overfitting in a smaller or more specialized network.

The **Vision Transformer from POSTER++** is structured to handle a broad range of vision tasks with its advanced features, while the **Original Vision Transformer with no projection** is tailored for specific scenarios where less complexity is warranted.

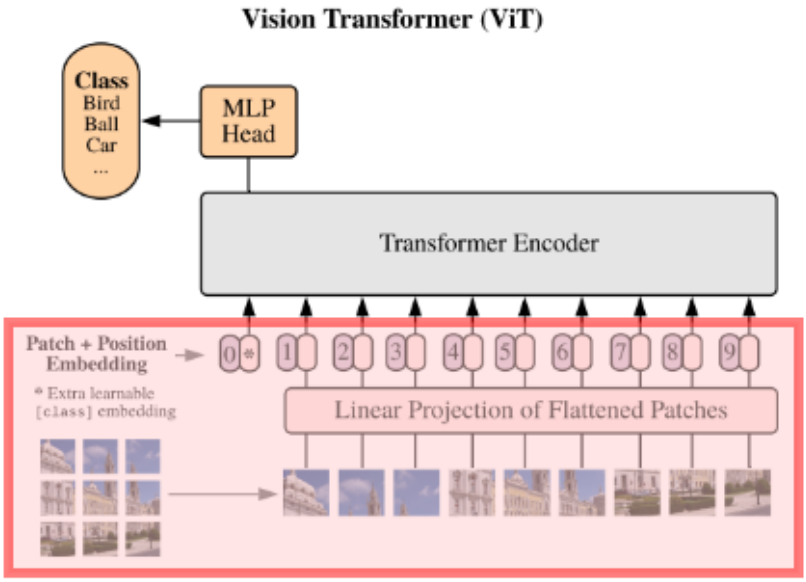

Figure 4.35: The red square shows the part removed from the original network [14].

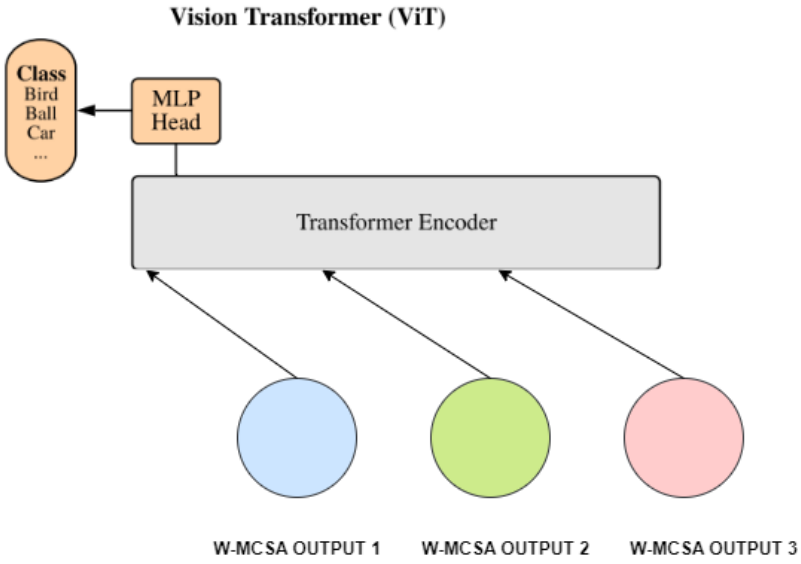

Figure 4.36: Use of ViT in the two streams model.

In testing, ViT was chosen to be used without the projection layer. In Figures 4.35 and 4.36 we see respectively the part removed from the original architecture and how it was used in the two stream model.

4.8.2. Two Stream Model results

This section delves into the comprehensive analysis of the Two Stream model by evaluating its performance across three distinct configurations: the model without the landmark stage, the complete Two Stream model incorporating both stages and the complete Two Stream model incorporating both stages plus pre-training on CelebA, as referred in Section 4.8.1. Each section provides insights into the model's performance, revealing the impact of each component on the overall efficacy of the system. These comparative results try to underscore the contributions of individual and dual-stream approach.

Two Stream without Landmark Stage

As illustrated in Figure 4.37, the validation accuracy presents a highly variable pattern. Despite the training accuracy reaching an impressive high of 0.8686, the validation accuracy is marked by significant fluctuations, only achieving a best score of 0.5501. Such volatility in validation accuracy is indicative of an unstable model that may not reliably interpret the underlying data patterns, thus leading to unpredictable performance when applied to real-world situations.

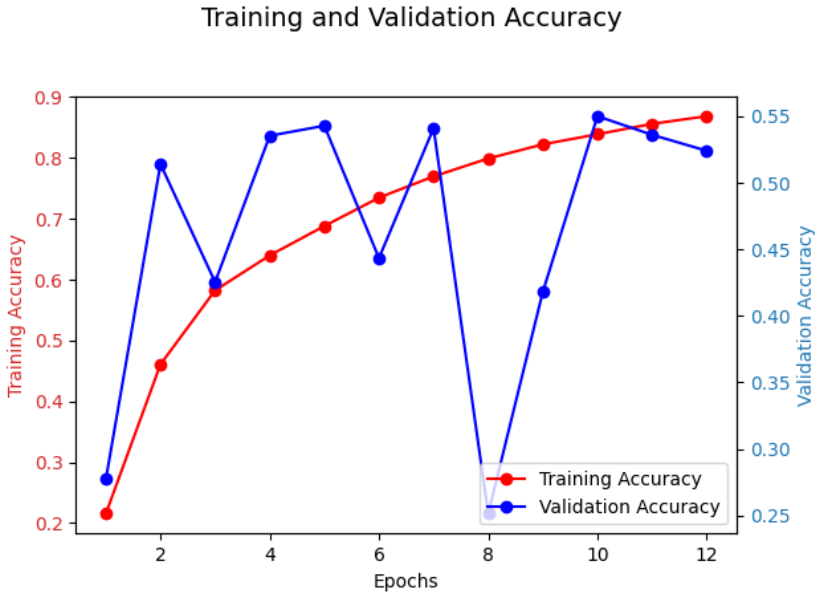

Figure 4.37: Training and validation accuracy for the Two Stream Model without Landmark Stage over 12 epochs.

In Figure 4.38, the Two Stream Model without the Landmark Stage exhibits a downward trend in training loss over 12 epochs, which demonstrates the model's capacity for learning from the training data. The model achieves a best training loss of 1.0112, signifying

effective pattern learning within the training set. However, the validation loss depicted by the blue curve suggests a concerning trend, as it tends to gradually increase and peaks at a validation loss of 1.5720. This pattern raises the specter of overfitting and casts doubt on the model's ability to generalize to unseen data.

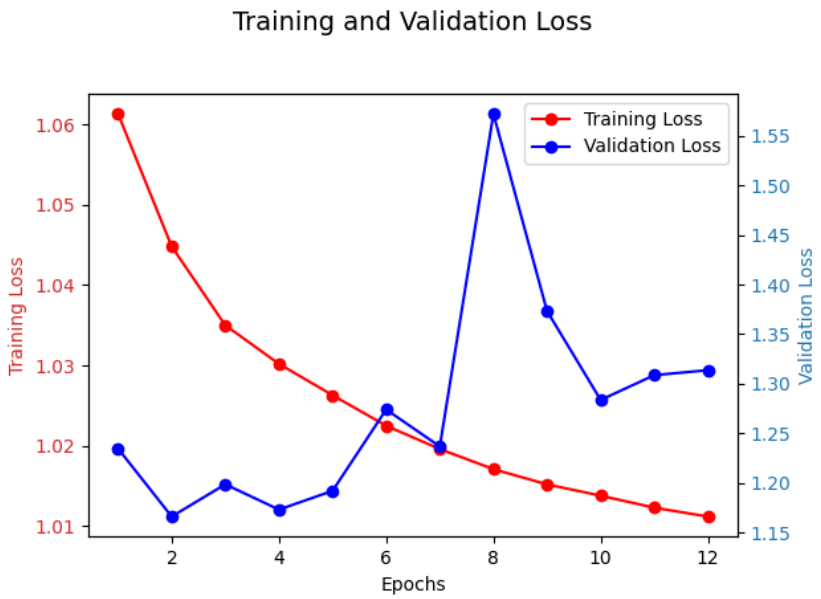

Figure 4.38: Training and validation loss curves for the Two Stream Model without Landmark Stage over 12 epochs.

The confusion matrix depicted in Figure 4.39 provides additional insights into the model's performance. The matrix reveals that while certain classes are predicted with high accuracy, there is notable confusion between others, indicating misclassification issues that need to be addressed. This is especially evident in the significant number of instances where 'Contempt' is mistaken for 'Happiness', suggesting that the model's feature extraction capabilities are lacking without the nuanced information that landmark detection can provide.

These findings underscore the imperative for continued refinement of the Two Stream Model. Enhancements aimed at stabilizing the model's learning and bolstering its ability to generalize will be crucial. The observed discrepancy between the high training accuracy and the much lower validation accuracy underscores the necessity of incorporating techniques to mitigate overfitting and to enhance the model's predictive robustness. The current state of the model, while showing promise, points to a clear need for improvement, particularly when the highest observed validation accuracy still falls short of the state-of-the-art models in facial affect recognition.

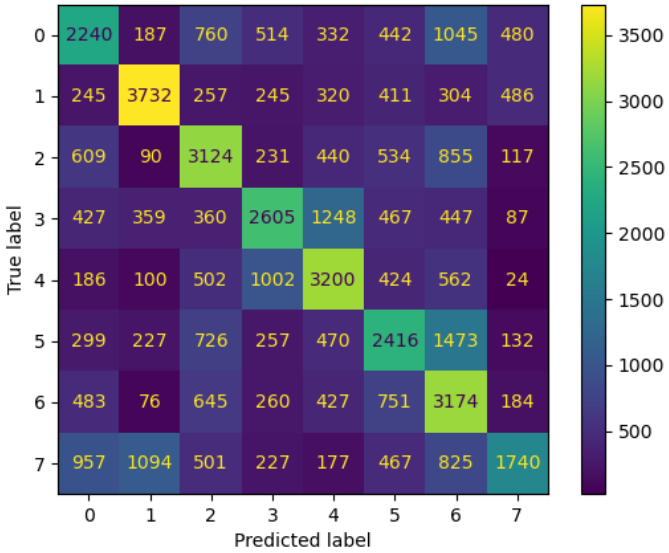

Figure 4.39: Confusion Matrix for the Two Stream Model without Landmark Stage, showing the classification performance for emotions labeled from 0 to 7 corresponding to 'Neutral', 'Happiness', 'Sadness', 'Surprise', 'Fear', 'Disgust', 'Anger', and 'Contempt', respectively.

Complete Two Stream without pre-training

The performance analysis of the Two Stream Model starts with Figure 4.40, illustrating an upward trajectory in training accuracy, reaching as high as 0.7710. This uptrend in training accuracy, however, is not mirrored in the validation accuracy, which peaks at 0.5739. Such a discrepancy underscores a concern that, although the model is refining its predictions on the training set, it does not translate this learning as effectively to the validation set.

Figure 4.41, shows a promising decrease in training loss to its best value of 1.0259, suggesting that the model is successfully learning from the training dataset. Conversely, the validation loss, with its best value at 1.1508, fluctuates, signaling potential difficulties in generalizing to new data.

The confusion matrix presented in Figure 4.42 offers a visual representation of the classification performance of the Two Stream Model without pre-training. The matrix delineates the frequency of each emotion being predicted compared to the true labels. The matrix shows a substantial number of correct predictions for 'Happiness' (label 1), as seen by the high value in its diagonal cell. Conversely, there are noticeable confusions between certain emotions, such as between 'Anger' and 'Disgust' (labels 6 and 5), which may signal

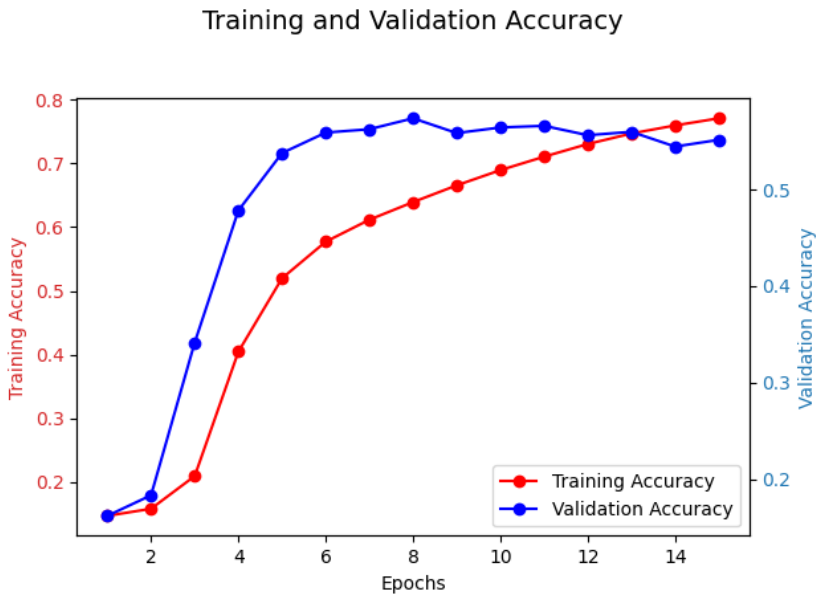

Figure 4.40: Training and validation accuracy trends for the Two Stream Model without pre-training.

Figure 4.41: Training and validation loss trends for the Two Stream Model without pre-training.

similarities in their expression patterns that the model struggles to differentiate.

The high misclassification rates, together with fluctuating validation loss, paint a picture of a model that is not effectively generalizing. It is adept at recognizing patterns it has

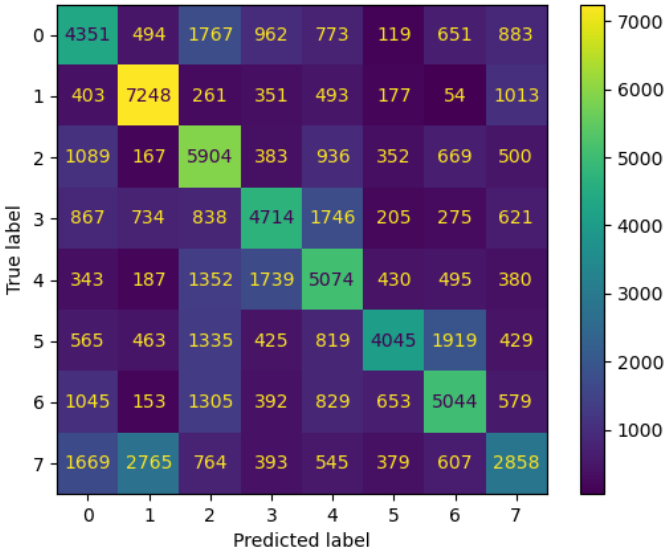

Figure 4.42: Confusion Matrix for the Two Stream Model without pre-training, showing the classification performance for emotions labeled from 0 to 7 corresponding to 'Neutral', 'Happiness', 'Sadness', 'Surprise', 'Fear', 'Disgust', 'Anger', and 'Contempt', respectively..

seen during training but fails to apply this knowledge to new, unseen data. The model's ability to generalize is crucial, especially in real-world scenarios where it will encounter a wide range of expressions and nuances.

In light of these findings, it becomes clear that there is a pressing need to address the overfitting observed in the Two Stream Model. Future directions include the implementation of more sophisticated regularization techniques, exploration of a more diverse training dataset, or the inclusion of additional contextual information such as the Landmark Stage to aid in the differentiation of subtle emotional cues. The ultimate goal is to develop a model that not only performs well on training data, but also maintains this performance when exposed to new, diverse datasets.

Complete Two Stream with pre-training

Figure 4.43 showcases a positive trajectory in the training accuracy, indicating progressive learning, with the best training accuracy reaching 0.8205. However, the validation accuracy, after an initial improvement, shows volatility, peaking at 0.5554, which suggests some challenges in the model's generalization to new data.

Similarly, Figure 4.44 illustrates a decrease in training loss to 1.0157, signifying the model's learning from the training data. In contrast, the validation loss experiences some fluctu-

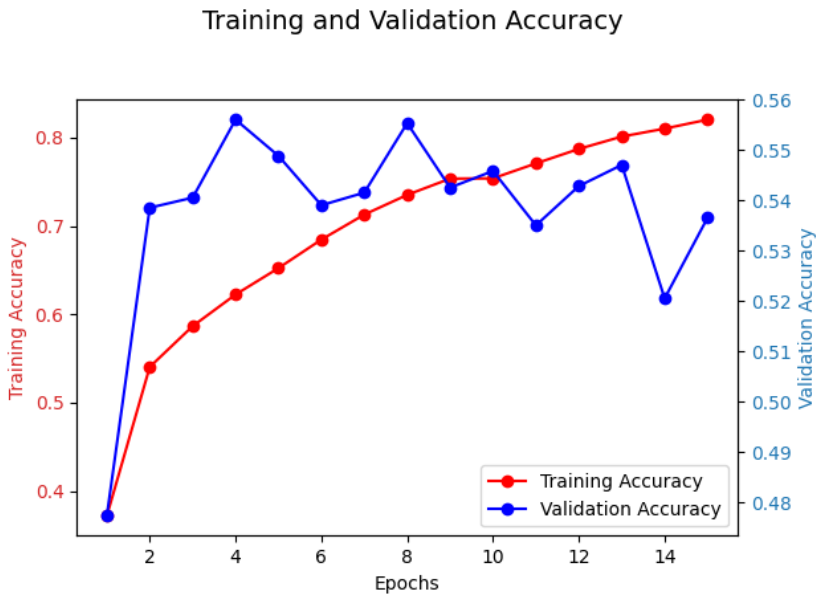

Figure 4.43: Training and validation accuracy trends for the Two Stream Model with pre-training.

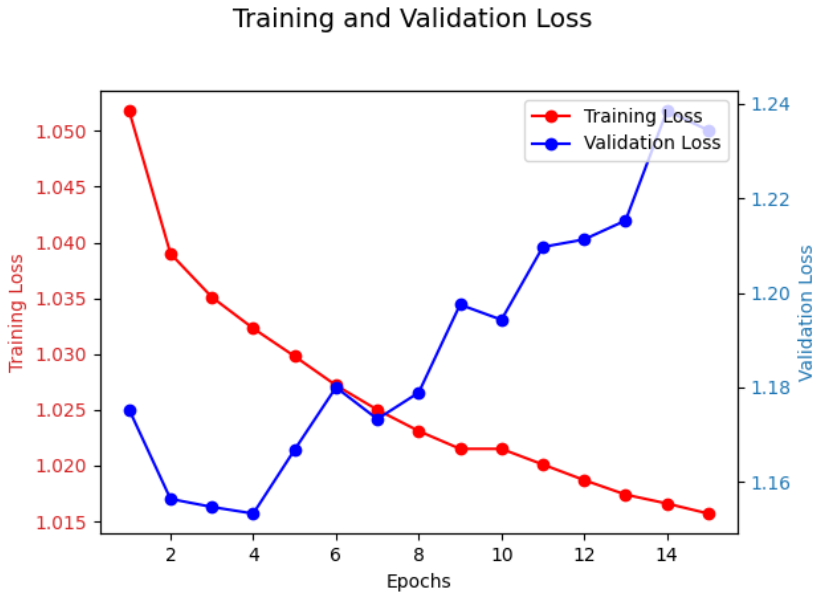

Figure 4.44: Training and validation loss trends for the Two Stream Model with pre-training.

ations, though the lowest point is 1.1533, indicating moments where the model's generalization was more effective.

The confusion matrix in Figure 4.45 further elucidates these points. In each row of the

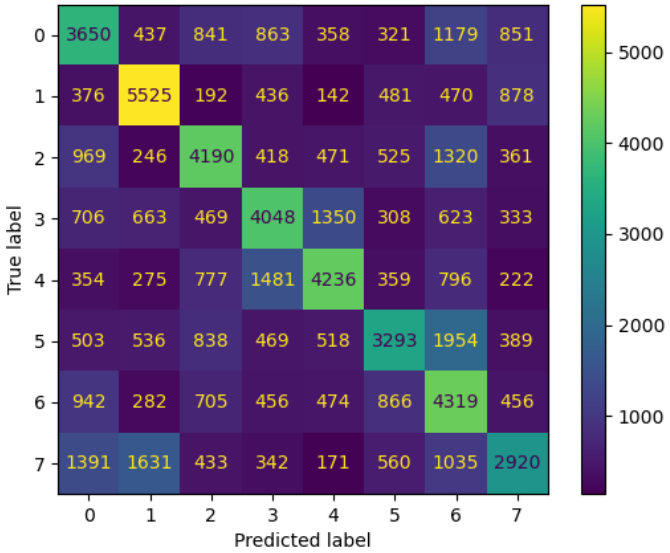

Figure 4.45: Confusion Matrix for the Two Stream Model with pre-training, showing the classification performance for emotions labeled from 0 to 7 corresponding to 'Neutral', 'Happiness', 'Sadness', 'Surprise', 'Fear', 'Disgust', 'Anger', and 'Contempt', respectively..

matrix are reported the the number of instances of an actual class, while in each column are the number of instances in a predicted class. It is a specific table layout that allows visualization of the performance of an algorithm, usually a supervised learning one. The matrix helps in identifying confusion between classes, where one class is mistaken for another by the model. A high value in the diagonal cells of the matrix indicates correct predictions, whereas high values in off-diagonal cells signal misclassification.

The matrix shown in Figure 4.45 for the Two Stream Model without the Landmark Stage shows a significant number of correct predictions, as seen by the high values along the diagonal. However, there are also considerable instances of confusion between certain classes, such as between 'Neutral' and 'Sadness' or 'Anger' and 'Contempt'. This suggests that while the model can learn distinct features for some emotions, it struggles to differentiate between others that may share similar facial expressions.

Overall, the patterns observed across these metrics over 15 epochs indicate that while the Two Stream Model with pre-training is learning effectively, as evidenced by the improvements in training loss and accuracy, its performance on unseen data, as indicated by the validation metrics, remains inconsistent. These insights suggest that further model optimization or data augmentation may be necessary to improve the model's robustness and reliability.

4.9. Comparative Analysis of Two Stream Model Performance

The performance of the Two Stream Model has been evaluated, as depicted in the various Figures of Section 4.8.2. These visual analyses delve into the model's training and validation dynamics, clarifying both its strengths and its limitations.

Interestingly, pre-training on an alternative facial dataset with DINO has not led to the expected performance gains. This unexpected result challenges the common presumption that pre-training on similar tasks universally leads to better model performance. It implies that for facial expression recognition (FER) tasks, the transferability of features may not be as straightforward as in other domains, it may require more carefully curated pre-training regimens or more simply, the dataset acquired for such pre-training is not suitable for FER tasks.

The Landmark Stage's incorporation significantly improves the model's stability. Without it, as Figures 4.38 and 4.37 demonstrate, the model learns efficiently but fails to generalize well, evidenced by the erratic validation loss and accuracy. This finding suggests that spatial information about facial landmarks is crucial for FER tasks, possibly because it supports the model to better contextualize facial expressions.

Moreover, Figures 4.41 and 4.40 reveal that omitting the pre-training causes a notable decline in training loss but an increase in validation loss variability. This points to a potential overfitting issue, where the model is unable to generalize its learning to new, unseen data, a critical requirement for effective FER systems. However, the stable training and overall performance achieved by this model put it at the top of all the tests carried out in this work.

In contrast, pre-training appears to solidify the model's initial learning, as seen in Figures 4.43 and 4.44. However, it does not translate into a consistent improvement in generalization, which is essential for robust real-world application.

State-of-the-art models like POSTER++ have achieved an average accuracy of 63.76% on the AffectNet dataset, as shown in [34]. Although commendable, these accuracy levels are relatively low, indicating that even the best models struggle with the FER task's complexity. This reveals the inherent challenges in FER tasks, which stem from the subtle and subjective nature of human expressions and the vast diversity of expression manifestations across different individuals and cultures.

In summary, while advancements like the Two Stream Model represent significant strides

in FER research, the relatively low ceiling of accuracy achieved by current state-of-the-art models calls for novel approaches. These may include more sophisticated feature extraction techniques, advanced pre-training strategies, or innovative model architectures that could better capture the nuances of human emotional expressions.

5 | Conclusions

5.1. Contributions of the Thesis

This thesis has contributed to the evolution of the neural network landscape toward the emerging potential of attention patterns, particularly in the field of facial emotion recognition (FER). The investigation was driven by the objective to test and evaluate leading approaches that harness the capabilities of attention mechanisms, culminating in the development of a specialized Two Stream model.

A notable contribution of this work is the introduction of a pre-trained DINOv2 model, adapted for FER tasks. This model was pre-trained on a dataset with similarities to our target domain, aiming to leverage the transfer learning paradigm to enhance performance. The empirical analysis provided within this thesis demonstrates that such pre-training can be considered as a double-edged sword; while it holds the promise of improved model performance, it also poses the risk of overfitting or misalignment between pre-learned representations and the target tasks.

The development of the Two Stream model within this thesis stands as a testament to the viability of a more generalized approach to image input handling. By incorporating a robust feature extraction process, the Two Stream model aims to encapsulate the nuances of emotional expressions more effectively. The results suggest that this generalization leads to performance enhancements, indicating the model's ability to capture the intricacies of facial emotions.

Personal reflections on this journey underscore the need for continued exploration into novel architectures and training methodologies. The field of FER is ripe for innovation, particularly through the integration of attention models that can holistically interpret the subtleties of human expressions.

5.2. Limitations and Future Perspectives

In this thesis, while notable progress has been made, certain limitations have surfaced, crucial to the understanding and future development of neural network models for facial emotion recognition. The modest levels of accuracy achieved by state-of-the-art models such as POSTER++ [34] indicate a significant challenge in bridging the gap between current AI capabilities and the intricate nature of human emotional expression, while being mindful that the datasets being worked on in this area are highly complex and varied.

The prevalence of overfitting is a concern that cannot be overlooked. It suggests that, while models can become highly tuned to the training data, they struggle to maintain performance when confronted with novel, unseen data. This limitation points to the necessity for improved regularization techniques and a more judicious approach to model complexity.

Moreover, the quality and diversity of datasets are of paramount importance. The models may benefit from exposure to richer datasets that encompass a wider array of expressions, scenarios, and cultural contexts, thereby enhancing their generalization capabilities.

Exploring additional avenues for pre-training presents another potential area for advancement. By pre-training on more diverse and extensive datasets, especially those closely aligned with the target application, models may develop a more nuanced understanding of emotional cues.

The consideration of emotions from multiple perspectives, such as incorporating contextual information like body language and environmental cues, may significantly enhance model performance. For instance, incorporating a Landmark Stage may prove beneficial, as it would allow the model to anchor its learning on key facial features that are pivotal in emotion recognition.

The development of multi-modal models also stands out as an exciting frontier. The proposed model could form a part of a larger, more sophisticated architecture that integrates various types of inputs, such as facial data, gestures, and sensory information, to achieve a finer understanding of emotions.

Lastly, the ongoing trend towards attention mechanisms holds great promise. These mechanisms enable models to prioritize and weigh information differentially, focusing on the most salient features relevant to the task at hand. Embracing and refining these mechanisms may lead to significant improvements in the accuracy and reliability of emotion

recognition models.

In summary, the journey toward creating AI systems that can empathize with and understand human emotions is underway. In the future we will most likely see development of datasets, in model architecture and learning paradigms, which will require a concerted effort by the research community.

Bibliography

[1] L. F. Barrett et al. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[2] C. Benitez-Quiroz, R. Srinivasan, and A. Martinez. Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, pages 5562–5570. IEEE, 2016. doi: 10.1109/CVPR.2016.600.

[3] L. Berkowitz. On the formation and regulation of anger and aggression: A cognitive-neoassociationistic analysis. *American Psychologist*, 45(4):494–503, 1990. doi: 10.1037/0003-066X.45.4.494. PMID: 2186678.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[5] N. Carion et al. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comput. Vis.*, pages 213–229, 2020.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[7] C. Chen. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021. URL `https://github.com/cunjian/pytorch_face_landmark`. Open-source software available at *https* : *//github.com/cunjia/pytorch*$_{f}$*ace*$_{l}$*andmark*.

[8] M. Chen et al. Generative pretraining from pixels. In *Proc. Int. Conf. Mach. Learn.*, pages 1691–1703, 2020.

[9] X. Chu, B. Zhang, Z. Tian, X. Wei, and H. Xia. Do we really need explicit position encodings for vision transformers? *arXiv preprint arXiv:2102.10882*, 2021.

[10] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion

bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[11] H. Dabas, C. Sethi, C. Dua, M. Dalawat, and D. Sethia. Emotion classification using eeg signals. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 380–384. ACM, 2018. doi: 10.1145/3297156.3297177.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR09*, 2009.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, volume 19, pages 207–283, 1971.

[16] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992. ISSN 0269-9931. doi: 10.1080/02699939208411068. URL `https://www.tandfonline.com/doi/full/10.1080/02699939208411068`.

[17] E. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[18] N. H. Frijda. *The emotions*. Cambridge University Press, 1986.

[19] I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: a report on three machine learning contests. In *Neural Inf. Process.*, pages 117–124. Springer, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-42051-1_16.

[20] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: https://doi.org/10.1109/CVPR.2016.90.

[22] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfew: a large-scale database for recognizing dynamic facial expressions in the wild. In *Proc. 28th ACM Int. Conf. Multimed.* ACM, 2020.

[23] X. Jiao et al. Distilling bert for natural language understanding. In *Findings Proc. Empir. Methods Natural Lang. Process.*, pages 4163–4174, 2020.

[24] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[25] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *Proc. Int. Conf. Neural Inf. Process. Syst.*, pages 1097–1105, 2012.

[26] J. Larsen and A. McGraw. Further evidence for mixed emotions. *Journal of Personality and Social Psychology*, 100(6):1095–1110, 2011. doi: 10.1037/a0021846.

[27] Y. LeCun et al. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[28] H. Li, N. Wang, Y. Yu, X. Yang, and X. Gao. Lban-il: A novel method of high discriminative representation for facial expression recognition. *Neurocomputing*, 432: 159–169, 2021. doi: https://doi.org/10.1016/j.neucom.2020.12.076.

[29] H. Li et al. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021.

[30] R. Li, T. Yuizono, and X. Li. Affective computing of multi-type urban public spaces to analyze emotional quality using ensemble learning-based classification of multi-sensor data. *PLOS ONE*, 2022. doi: https://doi.org/10.1371/journal.pone.0269176.

[31] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, pages 2584–2593. IEEE, 2017. doi: 10.1109/CVPR.2017.277.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin

transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[34] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang. Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023.

[35] A. Mollahosseini, B. Hasani, and M. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10:18–31, 2019. doi: 10.1109/TAFFC.2017.2740923. URL `https://doi.org/10.1109/TAFFC.2017.2740923`.

[36] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings British Machine Vision Conference 2015*, pages 41.1–41.12. British Machine Vision Association, 2015. doi: https://doi.org/10.5244/C.29.41.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. URL `https://doi.org/10.48550/arXiv.1912.01703`.

[38] R. W. Picard. Affective computing for hci. In *Proceedings of the HCI International 1999 - 8th International Conference on Human-Computer Interaction*, Munich, Germany, 1999.

[39] R. Plutchik. Chapter 1 - a general psychoevolutionary theory of emotion. In R. Plutchik and H. Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: 10.1016/B978-0-12-558701-3.50007-7.

[40] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340.

[41] A. Radford et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[42] F. Rosenblatt. The perceptron, a perceiving and recognizing automaton project para. Technical report, Buffalo, New York, USA: Cornell Aeronautical Lab., 1957.

[43] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT press, 1986.

[44] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980. ISSN 0022-3514. doi: 10.1037/h0077714.

[45] A. V. Savchenko. Granular computing and sequential analysis of deep embeddings in fast still-to-video face recognition. In *Proceedings of 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 000 515–000 520. IEEE, 2018.

[46] K. Schaaff. *EEG-based Emotion Recognition.* PhD thesis, Universitat Karlsruhe, 2008.

[47] K. Scherer. *Emotions, psychological structure of.* The Publisher's Name, 2001.

[48] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2015.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. doi: https://doi.org/10.1109/CVPR.2015.7298594.

[51] Unknown. Affective computing in marketing: Practical implications and research opportunities afforded by emotionally intelligent machines. *Idea Corner*, 33:163–169, 2022. Published: 04 January 2022.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, l. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[53] K. Wang et al. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020.

[54] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[55] Z. Wang, S.-B. Ho, and E. Cambria. Multi-level fine-scaled sentiment sensing with ambivalence handling. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 28:683–

697, 2020. doi: 10.1142/S0218488520500294. URL https://doi.org/10.1142/S0218488520500294.

[56] Z. Wang, F. Zeng, S. Liu, and B. Zeng. Oaenet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognit.*, 112:107694, 2021. doi: https://doi.org/10.1016/j.patcog.2020.107694.

[57] A. T. Wasi, K. Serbetar, R. Islam, T. H. Rafi, and D.-K. Chae. Arbex: Attentive feature extraction with reliability balancing for robust facial expression learning. *arXiv preprint arXiv:2305.01486*, May 2023. [cs.CV].

[58] Wikipedia. Emotion, 2021. URL https://en.wikipedia.org/wiki/Emotion.

[59] Z. Witkower and J. L. Tracy. Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems. *Emotion Review*, 11 (2):184–193, 2019.

[60] W. Yan et al. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS One*, 9(1):e86041, 2014.

[61] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[62] S. Zheng et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 6881–6890, 2021.

A | Appendix A

If you need to include an appendix to support the research in your thesis, you can place it at the end of the manuscript. An appendix contains supplementary material (figures, tables, data, codes, mathematical proofs, surveys, ...) which supplement the main results contained in the previous chapters.

List of Figures

List of Tables

Ringraziamenti

Quanto è fondamentale il respiro? Mi sento un atleta che, a fine corsa, inizia ad accusare la stanchezza dello sforzo fisico. Ogni boccata d'aria diventa quindi fondamentale, quasi vitale. Mi serve ossigeno per essere lucido e guardare al mio percorso con la serenità che questo merita, pur sapendo che non è facile trovare il tempo per prendersi tempo. Tantomeno lo è riassumere, in questi piccoli ritagli di giornata che uno si concede, 7 anni di vita che 7 anni di vita non sono. Purtroppo, e per fortuna, sono molti di più.

DA DOVE PROVENGO?

Tornando indietro al Settembre 2017, ricordo di un ragazzo un po' perso appena uscito dal liceo scientifico. Idee poco chiare, ma ben consapevole del suo valore. L'avventura al Politecnico non è per tutti e, per capire dove si sta andando, è necessario sapere da dove provengo. Nessun eroe infatti viene dal nulla.

Non posso non iniziare dai miei genitori, da mio padre Daniele e mia mamma Luara. Nessun percorso sarebbe iniziato e tantomeno nessuna meta si sarebbe raggiunta senza il vostro aiuto e la stima che mi avete sempre dimostrato, seppur tante volte io non ci abbia creduto. La mappa per interpretare il mondo è un vostro regalo e per ora non ha mai fallito. Vi dedico dei versi de *la Ballata dell'amore vero* di Claudio Chieffo:

Io ti voglio bene e ne ringrazio Dio,
che mi dà la tenerezza, che mi dà la forza,
che mi dà la libertà che non ho io.

L'affetto tra di voi l'ho sempre visto nè più nè meno di questo. E non potrò mai scordarmene. Vi sono immensamente grato, mi avete dato tutto.

Insieme sulla linea di partenza trovo poi mio fratello Pietro, sempre in cerca del suo posto nel mondo e costantemente proiettato ad altro. Ecco, ricordati che a qualcosa appartieni: sicuramente a queste pagine, e quindi alla mia e nostra storia. Non so bene dove le nostre strade si incroceranno di nuovo, ma sapremo di provenire dalla stessa storia.

In these bodies we will live, in these bodies we will die
And where you invest your love, you invest your life
Awake my soul
Awake my soul
Awake my soul
For you were made to meet your maker.

A differenza dei Mumford and Sons (la canzone è Awake my soul) non so bene chi ci abbia fatto (un'idea la potrei anche avere), ma se investirai la tua passione in qualcosa o qualcuno che ami, la felicità che hai sempre cercato ti investirà senza neanche accorgertene. In bocca al lupo

Mi imbatto poi in una torcia accesa, una luce improvvisa che ti investe totalmente e ti rende cieco per qualche secondo: ecco quindi Benedetta. Penso che non avrei potuto trovare metafora migliore. Non sembri neanche sorella dei tuoi fratelli per come ti poni sulle cose che ti capitano e sulle persone che incontri. Ad avere avuto la tua maturità a 18 anni, probabilmente ora avrei anche il dottorato. Ti auguro di non perdere mai quest'aria gioiosa che investe tutto quello che ti appartiene. Non ti cito il testo che renderebbe meno poetico il tutto, ma riascoltati Tranqui Funky degli Articolo e ricordati di quando te la cantavo in macchina da piccolina: vivitela tranquilla come tu sai fare.

Tra un arnese e l'altro trovo finalmente Antonio, nascosto in qualche stanza dell'azienda che ancora tratti come un figlio. Non avrei potuto chiedere nonno migliore: tenero, generoso e attento. Fermati un secondo e pensa che tutta la fatica che hai fatto è servita a farmi arrivare fino a qua. Un pensiero va anche alla nonna Annunciata, vista per poco, ma mai dimenticata. Se avete avuto una vostra canzone ecco, te la dedico insieme a un semplice Grazie.

Ovviamente, non posso dimenticarmi di mia nonna Graziella, mio zio Marco, mio cugino Riccardo, mia cugina Martina e mia zia Federica.

I PREPARATIVI

Come per ogni viaggio, è importante partire con l'attrezzatura giusta: lo zaino deve reggere, le scarpe non devono essere bucate e la bussola deve indicare la giusta direzione. Iniziamo i preparativi ringraziando quindi chi già c'era e partiva al mio fianco. Si parte sul facile: ringrazio Andrea, Fabio e Matteo per l'amicizia fraterna che ci lega. Dalle delusioni d'amore, alle serate da sfigati alla play, alle nottate a Ovada, alla mia pizza con l'uovo, tutto quanto ha arricchito la mia vita e continuerà a farlo. Sapere di avere amici

a cui puoi dire tutto, qualcuno che da qualsiasi parte del mondo condividerà gli stessi pensieri e ricordi indelebili è come sentirsi sempre a casa.

Hmm Insta Lova, Insta Lova
Insta Lova, Insta Lova
Insta Lova
La G, la U, la E...

In particolare il buon Burato, che oltre ad essere mio parente lontano, è anche un fratello che non ho mai avuto. Le nostre strade si divideranno, ma difficilmente non si reincontreranno. Saremo pure pessimi nel sentirci, ma ogni volta è come non essersi mai lasciati.

Lui chi è?
È un altro uomo che è impazzito per te...

Arrivando ad anni più recenti, si incontra il Marie Curie, "Il Lager". In ogni prigione è fondamentale affidarsi ai compagni di cella più sgamati per riuscire a sopravvivere, e in voi ho trovato quelli perfetti. Ringrazio quindi tutto il fantatroye per questi anni di amicizia e di presenza costante, partendo da Benedetta, Giacomo, Simone, Nicoló, Daniele, Alessandro, Marco fino ad arrivare a Pietro (il mio indimenticabile compagno di Interrail) e Marco il napoletano. Grazie a tutti e, seppur a modo mio, ci sarete sempre.

Big city life,
Me try fi get by,
Pressure nah ease up no matter how hard me try.
Big city life,
Here my heart have no base,
And right now Babylon de pon me case.

Parlavo di compagni di cella sgamati, giusto? Sicuramente (non me ne voglia nessuno) Lele Vaghi ne è il primo e migliore esempio, seguito a ruota dal buon Burgio. Grazie Samuele per le chiacchierate, le gite in montagna e l'affetto sincero che mi hai sempre mostrato nella nostra diversità. Grazie Lorenzo per le mille sigarette, i caffè al poli, le estate a Ossuccio e gli insulti sul calcio. Ovviamente c'è tanto altro, ma ho solo voluto elencare i pregi.

Oltre a tutto, i miei due tossici preferiti, Mirco e Pietro. Le giornate in stazione col primo, i ritorni in after col secondo. Le sgamate e i compiti copiati a scuola, eppure studiando un cazzo siamo riusciti a diplomarci e (quasi) a laurearci tutti e tre. Vi auguro il meglio: troverete il vostro posto nel mondo, la vostra arte verrà compresa e sarete contenti, ne sono certo.

Ricordo mamma che mi chiamava
Quando il fine settimana
Al tramonto tornavo a casa
Che puzzavo di marijuana (Scusa mamma)
Quattordici anni li hai una volta sola (Eh)
Quindi pensa bene se devi calpestare l'aiuola
I miei coetanei (Eh) pensavano alla scuola (Sì)
Io a dosare bene il bibitone, rum e coca-cola...

L'INIZIO

Si entra finalmente nel pieno, nella foresta più oscura. Ad accogliermi istantaneamente ci pensa Andrea: non potrò scordare il primo anno buttato tra ping pong e il 18 in fisica che ha ucciso definitivamente la nostra media. Ovviamente c'è anche Ale, con cui ho condiviso le tante chiacchierate fuori dall'interfacoltà, e Roi, eterno deluso da Milano. Vi auguro il meglio in tutto e per tutto.

Ecco che poi, tra i meandri del trifoglio e degli spazi che all'inizio sembrano immensi, c'è bisogno di tornare a qualcosa. Solo, frastornato e triste, mi rendo conto della necessità di avere amici che possa chiamare compagnia, amici. Decido di incontrare il CLU, e iniziare la storia più divisiva (fino ad ora!) della mia vita. In quell'esperienza mi son giocato tutto, fino in fondo. A volte risparmiandomi, a volte no, ma sempre come Giacomo (e Chris). Non è ancora chiaro, non so ancora che ruolo abbia avuto e avrà nella mia vita ma, tra tanti alti e bassi, sono comunque grato a tante persone che mi hanno accompagnato per questo lungo viaggio. Mi tornano in mente Deme (grazie per le mille sigarette), Bob, Emma (grazie per Marra), Fra, Elisa, Matilde, Marta (grazie per quell'aperitivo e la tua unicità), il Doc, il Bomber, Ciano, il Vaz (senza di te questa tesi non sarebbe venuta alla luce), Samsung, Ilaria, la Manu, il Gobbo, Landri, Franco, Rebe, Paolo, il Chad, Alex (Sandro, Eibom, grazie per il bot del McDonald), Verde (indimenticabile, con te ho rappato per la prima volta), Alo (per 5 mesi siamo sembrati sposati), MdV, Termi, Martina, Beppe.

Dicitencello
a 'sta cumpagna vosta
ch'aggio perduto 'o suonno
e 'a fantasia...

Tra tutti spiccano quelli che posso definire ormai compagni di vita, persone che sono certo mi accompagneranno sempre e per sempre. La forma è ancora da scoprire, ma non sarà

lasciata al caso.

In primis Sara, eterna indecisa e sempre in cerca del qualcosa in più, dell'emozione ancora non scoperta. Abbiamo condiviso di tutto: letture di SDC (a pensarci ora...), viaggi, sbronze, chiacchierate infinite, supporto psicologico a ogni dramma della vita. Ti ringrazio, non è facile prendermi, eppure la tua tenerezza mi è servita in tanti momenti.

Jacopo, non so come ma mi son laureato prima io. Per i primi 3 anni pensavo sarei stato brutalmente superato, e invece... sei stato il mio primo assurdo compagno di studi. Non dimenticherò mai il periodo nel tuo appartamento durante l'estate del Covid, uno dei momenti che ricordo più felicemente nella mia vita. Sei la persona più limpida che conosca. La tua semplicità mi è sempre stata d'esempio. Ti ringrazio.

Luigi, compagno vero di tanti pensieri, fatiche e gioie, ospite sempre gradito a casa con cui condivido un'amicizia a tratti fraterna. Ti ho visto sbocciare in questi anni, da un principio di incellaggio fino al trovarti in pari con gli esami e accompagnato da Luci. Sei sempre stato accompagnato dalla fede: non quella posticcia del "ce la farò", ma "sarò contento e realizzato". Ecco, la tua chiarezza me la porterò per sempre dietro.

Gigino, la bontà, l'umiltà, la gentilezza fatta persona. Ricordo quelle vacanze di natale in cui siamo stati i più ambiti in ogni singola cena, le confessioni amorose in stile 2a superiore, tutti i tuoi amati giochi matematici, la tua festa di laurea che mi ha fatto perdere la patente, il viaggio a Marsiglia. Grazie per avermi msotrato il gusto delle cose un po' più semplici e il tuo sguardo così puro.

BONUS: Siete tutti venuti a Varsavia ragazzi. Che giornata quella: scudetto del Milan, io quasi arrestato per le pizze, la messa alle 11 dopo essere andati a letto alle 9 post Cubano...

Filippo, fortunatamente ci sei stato: la tua componente marcia e tifosa è qualcosa che ho sempre avuto e grazie al tuo contributo l'ho coltivata e fatta diventare grande. Milan-Tottenham 1-0 2023, non ricordo nulla e tu nemmeno, so solo che stavi aspettando la mail da Matteucci di conferma per avere la tesi da 7. Questo è il riassunto di chi sei, grazie così.

ANDANDO AVANTI, TORNANDO INDIETRO

Siete sempre stati davanti ai miei occhi e ci siamo sfiorati per tanto tempo. Ci siamo visti, ma non ci siamo subito guardati. Questo capitolo è totalmente dedicato ai miei so, i miei migliori amici, persone che hanno visto il vero Giacomo e mi hanno costruito una

casa. Chiamare Limbiate la mia città mi ha sempre lasciato una sensazione straniante, come se la parola non corrispondesse appieno alla realtà dei fatti. Da straniero nella mia terra sono passato a orgoglioso maranza (e gatto) di zona.

Vi ringrazio di qualsiasi cosa possibile, per la vostra amcizia così pura, così semplice, per avermi sempre accettato e riaccettato anche nei momenti di stacco. Saprò sempre dove tornare quando passerò di qua. Mi basterà farmi trovare alle 11.30 alla Brianza, sapremo già cosa prendere e tutto sembrerà mai cambiato.

Ora elencherò tutte le cose che mi vengono in mente, senza filtri, ve le dedico una per una: le serate al deja, i 3 drink a 10 euro, il Polaris, il Giovi, Azzolini, Gessaga, Tilotta, Sergio, Lomu, la Maris (G.), la pizzata a Capriano con il Giuse che offre le birre, la pasta marcia di Sorrento, Sorrento, io che stavo per morire e ho pianto in quel posto di merda innominabile sulla costiera, l'Interrail con alcuni, Pietra Ferrazzana con altri, la Puglia e il viaggio in flixbus, Huge che ruba la tipa a mimmo, la mia festa di laurea abusiva in cantina, gli spritz campari da 20 gradi della Brianza, le cene a fare schifo, i metri di Medo, i tornei di fifa, la tombotana, l'Europeo incredibile tutti assieme, i calcetti, io che collasso in spiaggia in Puglia, quel cesso di Luchè, Selene, le bolle alla Snai, lo gnomo, don Gianluca (non dico che fine deve fare), Pioli (non dico che fine deve fare), la sparatoria a don Davide, la Cristina Storaci e il tradimento di Giuda, le grigliate di Mimmo, i ritardi di Tana, lo sporco di quel ratto di Taja, le canzoni dello chief manager Huge (Steven), i gossip di Ange, il Mons sempre in oratorio, le punizioni di Ga, le sbronze e le infinite chiacchierate con Rick, il petto enorme (come la sua fede) di Tax, le sigarette di Mestro, Vlahovich Lelli.

Grazie a tutti, vi voglio bene.

Cerco un po' di te nei testi di De André
Ci saranno lividi di cui andare fiero
Altri meno...

E fino a qua, tutto bene. Manca però un pezzetto: i giorni all'avventura sono a volte pesanti, altre volte molto faticosi ed è impossibile resistere senza fede, senza quella speranza che ti guida. Impossibile non avere una promessa da compiere, un destino che ti insegue, quel qualcosa di cui a volte ti dimentichi, che a volte odi pure, ma che ti sussurra sempre nell'orecchio. Una voce, una flebile voce, che intona qualcosa di simile: *"Dicono che siamo tutti dei delinquenti... solo perchè ci piace fare incidenti..."*. Mio caro Milan, sei un amore viscerale. Ovunque sarò nel mondo saprò di sentirmi a casa quando, ogni volta che vedrò il rosso, penserò automaticamente al nero. Vinceremo di nuovo la Champions League io sarò pronto a sbronzarmi di gioia come facevo a 20 anni. Ti prego, dopo lo

scudetto a Varsavia, questa me la devi.

Dedico una piccola postilla a un posto fondamentale per la riuscita dei miei studi. Mi hai accompagnato in triennale, in magistrale, mi hai coccolato nei momenti peggiori (ricordi elettrotecnica?). Quindi grazie Tilane, senza di te, Samuel e quei schifosi caffè al letterario non sarei arrivato fino a questo punto.

VIAGGIO INTIMO

In preda a una crisi umana, spiritica, d'amore, di senso, ogni uomo che si rispetti prende la strada della fede. Trovare il senso in altro, così rassicurante a pensarci ma così scomodo da seguire. Nel bel mezzo del mio avventuroso viaggio, intravedo tra le fronde fitte della foresta un qualcosa di simile a un campanile. Una punta, una campana, qualcosa che spunta sopra la mia testa. Decido di seguire e mi imbatto in una chiesa. A quel punto la curiosità è troppa: apro il portone e... sorpresa.

Catapultiamoci ad Assisi. Dedico un piccolo capitolo della mia avventura a quella città cosi bella e a quell'esperienza così meravigliosa. Quel silenzio, quell'immersione totale, il ripartire completamente da capo. Dopo il primo classico impatto da maranza, ho scoperto un modo di vivere la fede che mi ha lasciato il sapore di autentico. Con le fatiche che ogni uomo si trascina, ma autentico.

Oltre al mio pianto davanti al crocifisso e al momento epifanico in cui ho deciso di togliermi gli orecchini per apparire nudo agli occhi di Dio, ringrazio le persone incredibili che ho conosciuto in quei 4 giorni. Mi tornano alla mente Adele (il tuo fratellone cos' tenero, i tuoi serpenti), Emma (mi devi ancora far consocere quella tua amica...), Kristina (che ricordi quella chiacchierata durante il viaggio di ritorno), Alberto (i tuoi ritiri sulla figura dell'uomo) e Matteo (cosa non ti appassionava?).

Grazie, a distanza di 3 anni vi porto ancora nel cuore. I ricordi son sempre fatti di persone, e voi avrete sempre un posto nei miei.

ESPATRIO

Come in ogni storia che si rispetti, a un certo punto c'è bisogno della svolta. Troppo tempo nello stesso posto può inaridire i rapporti, può banalizzare gli avvenimenti e le abitudini che si pensa possano essere intoccabili. Dopo la fatica della triennale, non esiste miglior vacanza possibile se non un liberante semestre in Erasmus. Mai scelta fu più giusta nella mia vita: con tutta la lungimiranza del mondo, intuendo la voglia di scoprire

e di cambiare che avrei avuto negli anni a venire, scelgo di partire e di farlo verso uno dei miei più grandi amori a prima vista, Varsavia.

Non so cosa mi abbia così tanto affascinato di quella città lontana. Parte come ultima tappa sfigata di un interrail tra amici e si rivela essere quella che posso tutt'ora chiamare casa. Mi sentirò sempre accolto dal freddo gelido del clima, dal freddo gelido dei suoi abitanti, dalle grigliate sulla Vistula, dai mercoledì al Park, dai lunedì al Cubano, dal Milan al TuttiAmici, dai viaggi che mi sono regalato in quel periodo, dal Politechnika e il suo carico di studio ridicolo, dalle partitelle a basket, dalle pizze (in realtà birre) rubate al Carrefour, da quella chiacchierata infinita con Elisa all'angolo di quella strada verso casa, da Grojecka70A, da Audino e quel benedetto one pot all you can eat, dallo spagnolo che scopava di fianco alla mia stanza svegliandomi la notte, dallo sporco del mio appartamento, dalle paste di Franek, dal mio vagabondaggio continuo, dai pierogi (che ho tatuato).

E, nel mio infinito elenco di nomi e cose, non possono non ringraziare Caterina (best coinquilina e badante), Chiara, Gabriele, Carolina, Riccardo, Filippo, Fabietto (chissà dove sei, personaggio assurdo), Olek, Ylenia (con il tatuaggio delle coordinate del park), Elisa (amica del cuore), Jasmina (grazie per Berlino, persona splendida), Vanessa, Franek (amico del cuore, persona d'oro), i 5 pierogi a Burgas Sebastian, Luchetto, Lukas e l'olandese di cui non ricordo il nome, Nadja, Viktoria, Gabriela, Benedetta (la prima persona con cui ho mai parlato appena arrivato e best coinquilina), Janek, Gabriele, Viola, Mohamed (con la birra alle 6 di mattina al park), Don Giuseppe (indimenticabile aver salvato quella famiglia ucraina dalla guerra) e i mille altri volti che, a distanza di due anni, non posso più ricordare. Sarete sempre nel mio cuore (e tatuati).

Hej suczki, ra-ra-ra-ra
Znamy wasze sztuczki, ra-ra-ra-ra-ra...

Tra tutti, il rapporto più intenso e che tutt'ora mi accompagna è quello con Leonardo. Tra i mille gossip, i giudizi, le malattie, gli sbam, sei stata una grandissima scoperta e sei una certezza. Siamo destinati a dividerci per la nostra natura che ci porta sempre lontano e sempre a scoprire altro, ma non ci perderemo.

Un'altra notte in down
E bevo finché poi
Dimentico di noi...

VERSO LA FINE

L'ultimo anno prima di laurearmi è stato il periodo più rilassato della mia vita. Son finalmente riuscito a godermi il tempo e le amicizie costruite negli anni. Ho fatto milioni di grandi/piccole cose: viaggio a Roma 1, viaggio a Roma 2, La Coruna, Salerno, Lisbona, Marsiglia.

Per un periodo così tranquillo, mi sono circondato di persone tranquille. Ringrazio quindi Francesco, Stefano, Francesca, Alice, Sara, Riccardo, Martina, Vittorio. Vi auguro il meglio.

Senza Andrea Bonarini non starei scrivendo nessun ringraziamento e, pur avendo avuto vari momenti di disagio e di poca chiarezza sullo scopo del lavoro, quando ho avuto la vera necessità di concretizzare lei si è rivelato un grandissimo professore (oltre che una grandissima persona). La ringrazio profondamente per la fiducia e la possibilità che mi ha dato.

MARZIA

Devo metterti in fondo perchè sei arrivata per ultima, ma è evidente come queste pagine non sarebbero state scritte senza di te. Questi sono i ringraziamenti a tutte le persone che mi sono state accanto durante il percorso di studi e, giustamente, il tutto è condito da quel magico pizzico di nostalgia, da quel sapore di cose già vissute e che non riaccadranno più; ecco, la veritá è che con te non esiste. È tutto ancora da scrivere. Non c'è prima e non c'è dopo: c'è solo il tempo assieme. Sei ció che veramente mi ha fatto laureare da uomo, ció che ha compiuto il percorso, ciò che mi ha reso consapevole di essere Giacomo, spogliandomi da tutti i veli costruiti negli anni.

Quindi, cara Marzia, grazie per avermi supportato (e sopportato!!) in questi ultimi incredibili mesi.

Quando sei qui con me
Questa stanza non ha più pareti.

E ORA? DOVE VADO?

Voglio rileggere questi capitoli tra vent'anni. Nella mia ottica sognatrice, a questa domanda fondamentale avrò già risposto da un pezzo. Purtroppo dubito sarà così, ma sicuramente tanti altri pezzetti si saranno aggiunti e tante altre spille saranno state at-

taccate al mio zaino. Ora, senza affogare nella paura del futuro, è giusto fermarsi un altro secondo e pensare che "ho scritto pure i ringraziamenti della mia tesi (dopo quasi due mesi dalla laurea). Chi lo avrebbe mai detto che ce l'avrei davvero fatta?". L'obiettivo è chiaro, essere contento: la strada non la conosco, ma conosco me stesso e so di valere. Anche quando te lo dimentichi, guardati attorno: sei circondato da gente che ti vuole bene. Impara a volertene e a fartene volere.
In bocca al lupo al Giacomo che verrà

Will you come home and stop this pain tonight
Stop this pain tonight ?