Cross-Domain Emotion Recognition Using SAM-Based Region Extraction: A Comparative Study on FER and Emotic Datasets

by

NILOY MAJUMDER
21101022
TASNIM HASAN ROSAN
24241342
SHIHAB MUSA
24341121
TASMIA TABASSUM
21301495
SAMIHA ZAMAN
21201322

A Final thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Niloy Majumder
21101022

Tasnim Hasan Rosan
24241342

Shihab Musa
24341121

Tasmia Tabassum
21301495

Samiha Zaman
21201322

Approval

The thesis/project titled "Cross-Domain Emotion Recognition Using SAM-Based Region Extraction: A Comparative Study on FER and Emotic Datasets" submitted by

1. Niloy Majumder (21101022)
2. Tasnim Hasan Rosan (24241342)
3. Shihab Musa (24341121)
4. Tasmia Tabassum (21301495)
5. Samiha Zaman (21201322)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Md Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In the digital age, understanding facial expressions and recognizing emotions will be crucial for enhancing human-computer interactions. This research will help to develop a framework for better facial analysis and emotion recognition, integrating traditional computer vision with advanced deep learning models. To refine facial analysis and focus precisely on expressive regions, we will explore the application of advanced segmentation techniques, particularly the Segment Anything Model (SAM), to accurately delineate facial features before emotion classification. The detected facial regions, and potentially their segmented components, will be preprocessed with resizing, grayscale conversion, and normalization, ensuring consistency for emotion analysis. These preprocessed images will then be fed into a CNN model, which will be trained using the rich and contextually diverse EMOTIC dataset, subsequently mapped to predict one of seven fundamental emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Real-time emotion recognition will also be implemented using webcam video frames. Moreover, considering the significant role of facial expressions in communication, this research will contribute to mental health monitoring and improved human-computer interaction. By advancing the integration of OpenCV, advanced segmentation, and deep learning with a focus on comprehensive training data, we aim to create a robust and accurate system for facial analysis and emotion recognition. Instead, we want to create a model pipeline that is both objective and based in reality, capable of performing well in ordinary emotional settings. Rather than attempting to artificially balance the dataset, we accepted its inherent unevenness because emotions do not occur in equal amount in real life. We hoped to achieve more than just a high-scoring algorithm by allowing the model to learn from data that reflects how people actually express their feelings. We set out to create a model that understands emotions with the delicacy, depth, and diversity of human experience, detecting feelings not only in clean, controlled situations, but also in the chaotic, beautiful complexity of everyday life.

Keywords: Human Computer Interaction; Machine Learning; CNN; Facial Expression; Prediction; Decision tree; Regression Analysis.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption. Secondly, to our supervisor Md. Sabbir Ahmed sir for his kind support and advice in our work. He helped us whenever we needed help. And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration i

Approval ii

Abstract iii

Acknowledgment iv

Table of Contents v

List of Figures vii

List of Tables 1

1 Introduction 2

2 Literature review 5

3 Model Description 10
3.1 Models Used in Research 10
3.1.1 VGG-19 10
3.1.2 Classical CNN 11
3.1.3 MobileNetV3 13
3.1.4 Segment Anything Model (SAM) 14
3.1.5 ResNet-18 15

4 Methodology 16
4.1 Description of Dataset 16
4.1.1 Dataset Overview 16
4.1.2 Dataset collection and preparation for analysis 17
4.2 Proposed Dual-Stream Emotion Recognition Framework 20
4.2.1 Pipeline and Deployment 20
4.2.2 Making the emotion recognition model 22
4.2.3 Efficiency Optimization for High Computational Performance 23

5 Result Analysis and Discussion 26
5.1 Result Analysis 26
5.1.1 Classic CNN 26
5.1.2 Mobile Net V3 and VGG-19 28

5.1.3 ResNet-18 28
5.2 Discussion 30

6 Conclusion **33**

Bibliography **37**

List of Figures

3.1 VGG-19 Architecture 11
3.2 The Segment Anything Model (SAM) 12
3.3 MobileNetV3 Architecture 13
3.4 The Segment Anything Model (SAM) 14
3.5 using ResNET-18 FOR FUSION FEATURES FOR TRAINING . . . 15

4.1 Dataset Mapping Piechart 17
4.2 AMT interfaces for the emotion category (a) and continuous dimensions (b) annotation. 18
4.3 Category co-occurrence probabilities. 19
4.4 Current Scene-Centric Methods for Contextual Features 20
4.5 Illustration of the emotion recognition model pipeline. 21
4.6 Valence, arousal, and dominance are effective in expressing nuances in emotional expressions. 24

5.1 Classic CNN's accuracy 27
5.2 Classic CNN"'s confusion matrix 27
5.3 ResNet18's confusion matrix 29
5.4 Accuracy comparison using SAM in other Models vs our proposed pipeline 30
5.5 ResNet-18 Architecture Performance Visualization 31

List of Tables

4.1 Mapping of FER-2013 emotion labels to Emotic labels 23

5.1 2022–2024 State-of-the-Art model on FER2013 DATASET Comparison 29

Chapter 1

Introduction

Facial expressions are a powerful and universal form of nonverbal communication. People naturally exhibit a wide range of expressions during conversations, which are closely correlated with their emotional and psychological states. Alongside facial gestures, human body language also conveys emotional cues. Unlike verbal communication—which tends to be direct and explicit—facial and bodily expressions are often subtle yet rich in emotional meaning. These nonverbal signals are critical in shaping interpersonal interactions and frequently carry more weight than spoken words.

In the fields of robotics and computer vision, automated facial emotion recognition presents a significant challenge. Accurately identifying and classifying each unique facial expression into its corresponding emotional category remains complex. Although artificial intelligence has made substantial progress in this domain, interpreting human emotions with the accuracy and nuance of a human observer is still a difficult task. Nevertheless, the potential applications of facial emotion recognition are extensive, spanning areas such as e-commerce, entertainment, education, public safety, healthcare, and mental health monitoring [2], [13].

Facial expressions transcend linguistic and cultural boundaries, serving as a universal language of emotion. Emotions such as happiness, sadness, anger, fear, disgust, and surprise are recognized globally and form the core of many psychological models of emotion. These fundamental emotions are not only widely acknowledged but are also essential for meaningful social interaction. For example, a smile may express warmth and agreement, whereas a grimace might signal discomfort or disagreement. These expressions facilitate empathy, understanding, and social cohesion by enabling individuals to respond appropriately to others' emotional states.

The technological ability to automatically detect and interpret emotions from facial expressions has become increasingly valuable. Advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved the accuracy and reliability of such systems. CNNs, composed of convolutional, pooling, and classification layers, excel in image-based classification tasks and are widely adopted for emotion recognition [13]. These systems enable machines and robots to better understand and respond to human emotions, ultimately enhancing human-computer interaction.

As society becomes increasingly reliant on digital communication, the need to understand emotional cues through facial expressions becomes even more critical. Accurate recognition of these expressions not only improves communication but also contributes to the development of emotionally intelligent systems that can support applications in education, therapy, social robotics, and beyond.

This thesis aims to enhance facial emotion recognition by incorporating self-attention mechanisms and hybrid emotion mapping strategies. It investigates the effectiveness of Self-Attention Modules (SAM) in emphasizing critical facial regions, explores the integration of discrete (categorical) and continuous (valence-arousal-dominance) emotion labels, and evaluates the benefits of transfer learning across diverse datasets. The proposed methods aim to strike a balance between model accuracy and efficiency, enabling real-time deployment in practical applications.

Research Problem

Recognizing human emotions from images remains a challenging task due to the complex and varied nature of emotional expression. Traditional approaches largely rely on static, facial-only inputs, overlooking the fact that emotions are often influenced by surrounding context and body posture. Furthermore, emotion expression is subjective — individuals may express the same emotion differently depending on their personality, culture, and environment.Therefore, a more reliable strategy for overcoming these obstacles combines regression and classification models. Classification assists in determining the kind of emotion (such as joyful, sad, or furious), and regression can assess the degree or strength of that emotion (such as how happy or sad a person is). By integrating these two approaches, we hope to create a model that can identify the existence of a particular emotion and provide information about how strongly it is being felt. Moreover, the conventional techniques for facial emotion recognition often take into account a face image which may be differentiated using an information image, and facial features or sections are identified from the face regions [16]. These face sections are then isolated from other spatial and global aspects. Finally, a classifier is trained to generate classification outcomes based on the distinct features [16].

Another key challenge is the imbalance and limitation of emotion datasets. For example, facial expression datasets like FER2013 contain only grayscale images of faces labeled with one emotion each, limiting their ability to represent real-world emotional complexity. Moreover, many classical models treat emotion recognition as a simple classification problem, failing to capture the intensity or nuance of emotions (e.g., mildly happy vs. very happy).

To overcome these limitations, this research explores a multi-task deep learning approach that combines categorical (classification) and continuous (regression) emotion prediction. It also leverages scene context and body posture using the EMOTIC dataset, aiming to build more robust emotional representations that can generalize well even on face-only datasets like FER2013.

This work addresses the gap between simple emotion labeling and deeper, context-sensitive emotional understanding, enabling models to better recognize complex

emotions in real-world conditions. In conclusion, developing a successful emotion identification system necessitates a thorough comprehension of human emotional fluctuations in addition to advanced modeling methodologies. By combining regression and classification, our method aims to close the gap between simple emotion detection and deep emotional understanding, eventually enabling a more precise and precise analysis of human emotions.

Research Objective

This study aims to advance the field of affective computing by developing a comprehensive emotion recognition model that incorporates both categorical emotion labels and continuous emotional dimensions—specifically valence, arousal, and dominance—to capture the complexity of human emotional expression.

Traditional facial emotion recognition systems primarily rely on static facial images and treat emotion recognition as a discrete classification task, which limits their ability to interpret emotions in nuanced or ambiguous real-world contexts. To overcome these limitations, the proposed system leverages the EMOTIC dataset for its full-scene imagery and multimodal emotional annotations, and the FER-2013 dataset for evaluating the generalizability of facial expression recognition across constrained visual domains.

The model architecture employs pretrained ResNet-18 backbones to extract features from facial regions, body posture, and environmental context, which are then fused to form a context-aware emotional representation.

A mapping strategy is implemented to align the 26 emotion categories in EMOTIC with the 7 universal categories in FER-2013. This enables effective cross-dataset training and evaluation, and facilitates transfer learning between context-rich and facial-only domains.Rather than balancing the datasets, the model is trained on the natural class distribution of emotions. This decision reflects the inherent asymmetry in emotional frequency in real life and allows the system to learn in a way that mirrors realistic emotional encounters.To enhance model robustness and address challenges related to data imbalance, class-weighted loss functions, dropout regularization, and data augmentation techniques are incorporated into the training process.

Ultimately, this research seeks to bridge the gap between superficial emotion classification and deep, context-sensitive emotional understanding. The proposed model is designed to operate effectively in real-world scenarios, with applications in areas such as human-computer interaction, mental health monitoring, and adaptive AI systems that require nuanced emotional intelligence.

Chapter 2

Literature review

The literature review mainly emphasizes the application of different machine learning models especially Convolutional Neural Network (CNN). Here FER2013 datasets are explored and tested which eventually gives us an analytical idea for the research. Additionally future objectives and possible results are discussed throughout. Hence after consideration of various things this literature review gives us an overall idea about analysis on different aspects of the papers.

According to Paul Ekman (1970)[1], when emotion recognition through facial expressions in real time was our goal, we tended to achieve as much accuracy and as many diverse categories of emotion as possible, as humans have 6 basic emotions. Therefore, our goal here was to find different approaches to identifying different emotional states based on various research and also to identify up-to-date methods to get the most accurate results.

To begin with, in the paper of Khaireddin and Chen (2021)[18] , they investigate facial emotion recognition in computer models and how challenging it is to analyze FER in images and different computer lighting of human faces and variations. They try to analyze how emotion changes over time and detect emotion from facial expressions. Moreover, the authors have achieved highest accuracy on the FER-2013 dataset.They have used Convolutional Neural Networks(CNN) for automatically learning features from images and for computational efficiency. Besides that they use VGGNet architecture for fine tuning optimization and experiment computational efficiency. They gained 73.28 percent accuracy in their model on FER2013 without using extra data training. This type of sentiment analysis and facial emotion detection provides an idea in the long lasting approval of the sentiment bias.

From this paper(Pradeep, 2024) [30] they have worked with emotion detection and real time facial recognition systems. They investigate how people express their emotion and facial expression and from this analysis they build a system which can detect emotion and facial expressions.This system can detect in live video stream with CNN and deep learning algorithm. Additionally, this system can use pretrained VGGFace models to extract features from the facial expression and detect human emotion states. They achieved 95.2 percent accuracy prediction emotion state in the FER2013 dataset which shows that this software performance is very efficient for processing data at a real time rate of 25 fps on an average PC machine.

In the paper, the authors Mishra et al. (2021) [19], have worked with real time expression detection of multiple faces using deep learning. The authors of this paper utilizes the Convolutional Neural Network (CNN) technique to explore the domain of deep learning for facial expression recognition. Moreover, the authors have conducted real-time testing and training of the proposed framework using the FER2013 dataset. Additionally, in the training phase many pooling layers are employed to gather features from the images, and a haarcascade classifier is utilized to detect whether a face is present in the frame. According to the authors of this paper their system can identify five main facial features like happy, sad, neutral, surprise and angry from the face of the person. Furthermore, this system can also identify multiple facial expressions at the same time. Besides, the system achieved a training accuracy of 73.12 percent.

The authors Kaur et al. (2022) [22] , of this paper have dealt with the method of identifying various human facial emotions through a software program that utilizes the Haar-Cascade Algorithm and the pre-trained dataset DeepFace. Here, the authors made the proper use of DeepFace, which has allowed them to reach an accuracy of nearly 97 percent. Moreover, they have also attempted to investigate the issue with previous strategies and assessed the efficiency of various other DeepFace technologies with one another. Furthermore, the authors of this paper have performed real-time emotion recognition through a webcam that can identify the emotion of an individual. According to the authors, this can be improved even more so that it can simultaneously identify the emotions of multiple faces.

In the paper, the authors Kedari et al. (2021) [17], have implemented deep learning algorithms to detect human emotions such as happy, sad, fear, anxiety, etc. on FER-2013, CK+, and different other datasets. Moreover, this paper focuses on the present Facial Emotion Recognition (FER) methods and datasets. This FER method is formed of three parts, which include face detection, feature extraction, and expression classification. According to the authors, the precision of the model through CNN is 60 percent. However, for FER 2013 and CK+, the authors have seen accuracy of 99.1 percent, which is the maximum, including the average accuracy, which was 93 percent.

This research paper Rathour et al. (2021) [20], covers the recognition of real-time facial emotion of employees using Raspberry Pi. The authors of this paper described that at present facial emotion recognition is a topic of great curiosity in the areas of social science and human-computer interaction domains. Here, the study mainly focuses on emotion recognition using raspberry pi which is possible due to the various advancements and improvements in artificial intelligence and this uses the mini exception deep network which helps to achieve almost a quite high accuracy in face and emotion recognition. To start with, artificial intelligence comprises convolutional neural networks which consist of convolutional layer max pooling layer and fully connected layer. So this study uses the mini exception CNN that incorporates convolutions depth wise therefore minimizing the parameters along with maintaining performance as well as eliminating the need for fully connected layers. As a result accuracy as well as efficiency is maintained in the image analysis process. However the framework that the authors have proposed consists of basically 3 main parts which are preparing the dataset then running the models on it and finally real time face and emotion recognition. The datasets consist of various images of different

emotions and using the mini exception deep convolutional network the dataset is trained and therefore accuracy can be found. The face recognition part mainly consists of face detection where faces are detected in real time images then alignment of the faces are done and finally face encoding and classification are done. Lastly during face emotion recognition the preprocessed image is sent to the cloud and mini xception network recognizes the different emotions.Finally comes the result where there is quite high accuracy as the faces were detected effectively even though there were variations in different faces. Additionally it detected negative emotion along with all types of emotions and maximum images are recognized correctly. Therefore this framework has performed better than some of the previous ones where accuracy was even lower. However there are still some places of improvements as more balanced datasets can be found along with various other emotions where there will be less limitations than this one thus increasing the accuracy even more. Thus even after the existing methods that have already been studied there are still some points where further research can be carried out therefore increasing the accuracy under diverse conditions along with expanding the range of detectable emotions.

The author's Dudekula and Purnachand (2023) [24], of this study mainly focuses on detecting facial emotion expressions which helps to examine stress levels along with analyzing psychological disorders. It focusses on various challenges such as lighting changes, obstructions, and changes in facial appearances.As a result the researchers use the Xception model and VGG-19, therefore using the convolutional neural networks (CNNs) for the feature extraction and the OpenCV framework for classification. However the implementation on the NVIDIA Jetson Nano achieves quite high video processing rates.The accuracy rates that have been achieved are 97.1 percent for Xception, 98.4 percent for VGG-19, and 95.6 percent in real-time environments on the CK+ dataset.

The author Duncan et al. (2016) [7], of this study works on the analysis of various human emotions through facial expressions that have been captured through live webcams.This is particularly relevant in the context of increasing virtual interactions since the Covid-19 pandemic. Therefore this research focuses on other methods rather than the traditional facial emotion recognition (FER) methods.These methods involve modern approaches using Deep Learning, Convolutional Neural Networks (CNNs), and Transfer Learning therefore resulting in more effective emotion classification.

From the paper of Barsoum et al. (2016) [6], they implemented a model for real time emotion recognition from videos in the Wild 2016 challenge. They proposed an idea that collects video streams from different sites and with this trimmed video which is used for input and eventually try to give some desired emotion recognition as output. This output is encoded from seven outputs like anger, happy, sad, neutral, disgust, fear, surprise. This implemented model can detect facial emotion, feature extraction, feature encoding and SVM classification from images and videos. Their implemented system achieves 59.42 percent validation accuracy and 56.66 percent accuracy in emotion recognition rate.

Zhang's et al. (2019) [14], paper proposes a facial emotion recognition (FER) method which combines convolutional neural networks (CNN) with image edge detection for improving accuracy in complex backgrounds.To increase efficiency, the technique

utilizes maximum pooling, extracts edge features, and normalizes images. The study beats typical CNN models on the FER-2013 dataset, with an accuracy of 88.56 percent. Using a Softmax classifier, the system divides emotions into seven different categories.Therefore tested against the FER-2013 and LFW datasets, their tests show greater recognition rates and faster training times.

Pitaloka et al. (2017) [12], focuses on improving CNN's ability to recognize emotions from facial expressions using a number of preprocessing steps, including scaling, face identification, cropping, noise addition, and normalizing approaches. They demonstrate that face detection alone results in 86.08 percent accuracy, while adding preprocessing approaches improves accuracy to 97.06 percent. Using datasets like JAFFE and CK+, this approach successfully recognizes six main emotions: anger, happiness, disgust, fear, sorrow, and surprise. The findings demonstrate how important data preprocessing is for improving CNN's performance on recognizing emotions tasks.

This paper's authors Zhang et al. (2024) [33], aim to enhance emotion recognition using deep learning, integrating gait analysis (a person's walking pattern) with facial expression recognition for improved robustness. The research addresses key limitations, such as facial recognition under poor conditions (blurriness, masks) and the limitations of relying solely on single physiological signals like EEG. The dataset includes both real and synthetic gait data, totaling 6177 samples across four emotional categories: happy, sad, angry, and neutral. The methodology employs a lightweight Spatio-Temporal Graph Convolutional Network (Shift-STGCN) combined with a self-attention mechanism to capture spatio-temporal correlations in gait data. On the E-Gait dataset, the model's maximum accuracy was 86.11 percent.Despite promising results, the study highlights limitations in its integration of facial and gait data and the narrow range of emotions analyzed.

From the paper by Savchenko (2022) [23], address the challenge of real-time facial emotion analysis using lightweight models suitable for mobile devices. With the goal of achieving high accuracy and computational efficiency, the study focuses on action unit detection, valence-arousal prediction, and facial expression recognition.During the third Affective Behavior Analysis in-the-wild (ABAW) competition, they produced cutting-edge results on the Aff-Wild2 dataset.The authors used shallow MLPs for task-specific training together with EfficientNet models that were pre-trained on VGGFace2 and fine-tuned on AffectNet for feature extraction. For expression recognition, the EfficientNet-B0 model obtained an F1-score of 0.38, for valence-arousal prediction, a mean Concordance Correlation Coefficient (CCC) of 0.46, and for action unit detection, an F1-score of 0.54. The EfficientNet-B0 model was significantly better than the VGG16 baseline, with up to a 50 percent improvement in multi-task learning challenges but showed lackings of rare emotions. As a result, this work demonstrates the feasibility of using on-device facial expression detecting along with sentiment analysis in real-world applications.

In the paper, the authors Qu et al. (2023) [26],Using camera photos, apply CNN for real-time facial emotional recognition. In particular, for real-world use cases including commerce, surveillance, and emotionally intelligent systems, the project tackles the need for enhanced facial emotion detection models that can reach high accuracy and function well in real-time applications.In particular, for real-world use cases

including commerce, surveillance, and emotionally intelligent systems, the project tackles the need for enhanced facial emotion detection models that can reach high accuracy and function well in real-time applications. They test two optimizers, SGD and Adam, to evaluate performance.Three convolutional layers build the structure of the CNN, which is followed by a fully connected layer that uses Softmax at the output layer to classify emotions and ReLU activation, max pooling, and dropout to prevent overfitting. The SGD optimizer achieves a maximum validation accuracy of 60.20 percent after 200 epochs, while Adam underperforms with significantly lower accuracy. The paper highlights the challenges of class imbalance in the FER2013 dataset, particularly with the disgust class, and the slow improvement in accuracy beyond 100 epochs. The work contributes to ongoing research in facial emotion recognition by improving CNN architecture for space complexity and testing different optimization strategies.

Hence, the reviewed papers explore a diverse methodologies for FER or facial emotion recognition with various deep learning techniques and approaches. The methodologies tackle different obstacles while offering computational efficiency, real time application and additional research that examines techniques like edge detection, preprocessing for improved CNN performance, integration of gait analysis, and hardware-efficient systems for mobile devices and usage in a wide variety of fields. Taking these in attention give us an insight to more advanced methodology.

Chapter 3

Model Description

To begin with, we used a dataset of face emotions that were divided into seven fundamental emotions for the purpose of our research. The following dataset was thoroughly chosen and put together to make sure optimal training and accuracy for the models. We examined the efficiency of three distinct models to verify our approach and ensure high-quality results. The dataset was divided into three segments: training, testing, and validation.

3.1 Models Used in Research

3.1.1 VGG-19

The VGG19 model (also VGGNet-19) shown in Figure 3.1 is a Convolutional Neural Network (CNN) structure which is 19 layers deep model and was introduced by the Visual Geometry Group (VGG) at the University of Oxford. The VGG19 model is comparable to the VGG16 model; however, it has 19 layers rather than 16. The numbers 16 and 19 represent the number of weight layers (convolutional layers) in the model. Indicating that VGG19 has three additional convolutional layers compared to VGG16 [36]. Moreover, the robust execution and ease of use makes VGG19 very popular and is frequently utilized for image classification and feature extraction activities. Besides, a pre-trained version of the VGG19 has been trained on over a million images that can be loaded from the ImageNet database [34] . Additionally, there are 1000 categories of objects into which the pretrained network can categorize images. The network has consequently acquired detailed feature representations for a variety of images. The size of the input picture for the network is 224 by 224 [34]

The VGGNet contains three layers with full connectivity. The third layer includes 1000 channels total, with one channel for each class, compared to 4096 channels for the prior two levels. ReLU (Rectified Linear Unit)

$$\text{ReLU}(x) = \max(0, x)$$

is used in all of the hidden layers of the VGG network. Local Response Normalization (LRN) is typically not used with VGG as it increases memory usage and training time. In addition, it doesn't improve accuracy overall[21].

Furthermore, the 19 trainable parameter layers of the VGG19 model comprises 3 fully connected layers and 16 convolutional layers [29]. The network executes better on tasks such as object recognition because of its depth, which enables it to acquire complicated information. Among the key features of VGG19 the convolutional layers utilize small 3x3 filters featuring a stride of 1 and padding of 1 [29]. This ensures uniform spatial dimensions across all layers. In order to minimize spatial dimensions, some convolutional layers are followed by max pooling layers. Additionally, to add non-linearity, Rectified Linear Units (ReLU) are applied as the activation function following each convolutional and fully connected layer [29] .

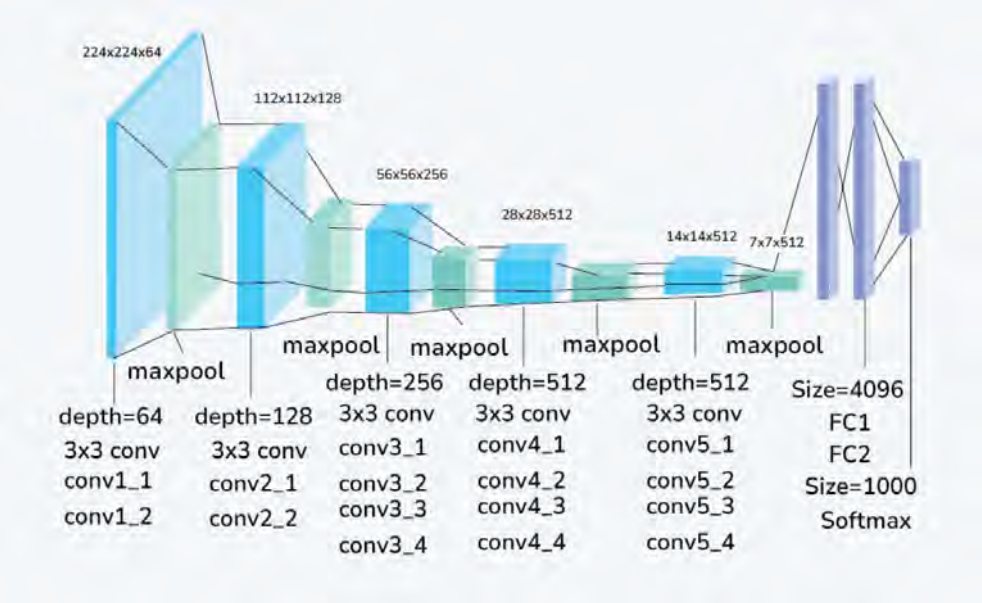

Figure 3.1: VGG-19 Architecture

3.1.2 Classical CNN

Secondly, multiple layers usually make up a traditional CNN (Convolutional Neural Network) model, which analyzes and gathers features from images. Similar to how our eyes and brains perceive and distinguish, classical CNNs are able to accomplish the same [25] . As a result, CNNs are necessary in order for a computer to comprehend and analyze a scenario. These advanced mathematical models excel at identifying patterns in images, identifying people comprehending scenes, coloring black-and-white images, helping with medical diagnostics and surgery, and even comprehending documents [25] . Moreover, an overview of the fundamental components of a traditional CNN model include image as an input (e.g., 224x224 pixels, based upon the dataset). The image is displayed in one channel for grayscale images or three channels (RGB) for colored images [27]

to put it in simple terms as seen on figure 3.2, CNN analyzes the characteristics of an image and reduces its dimension without affecting any of its features [27] . Therefore, convolutional layers are mostly used for acquiring spatial characteristics from images. In order to identify characteristics like edges, textures, and more complicated patterns, these layers perform convolutional operations to input images

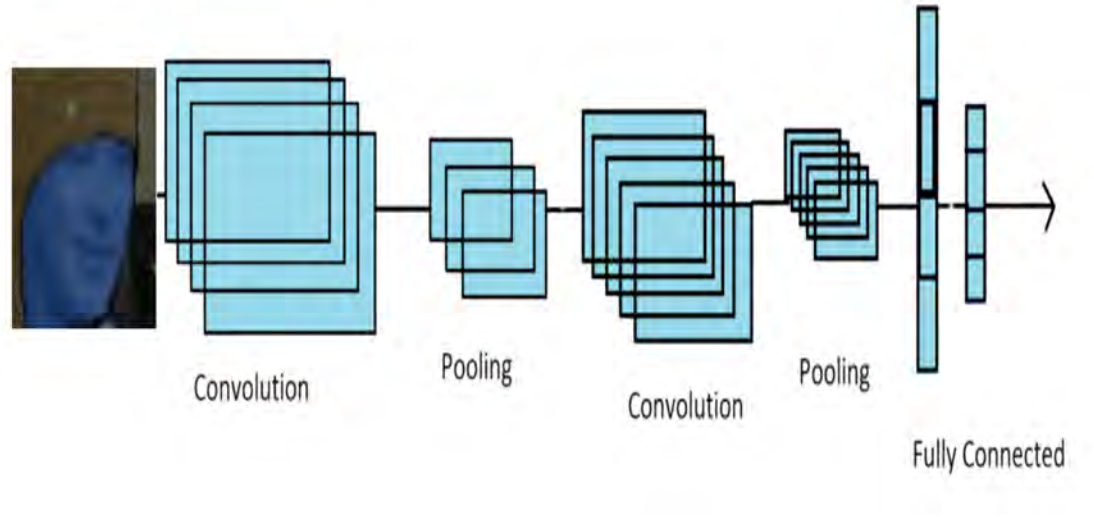

Figure 3.2: The Segment Anything Model (SAM)

through filters, also referred to as kernels. Moreover, the spatial connection between pixels is maintained with the use of convolutional methods [28] . Additionally, to add non-linearity to the model, a non-linear activation function such as ReLU (Rectified Linear Unit) is usually added following every convolution step [28] . Another component of CNN is pooling layers. In order to avoid overfitting, pooling layers are used to decrease spatial dimensions and the amount of parameters. In general, max-pooling will be used when the highest value is selected from a pool of values. Besides, another component named fully connected layers whose main purpose is to merge the characteristics to facilitate the final decision-making process (classification or regression) following feature extraction.

3.1.3 MobileNetV3

MobileNetV3 (Figure 3.3) is an advanced deep learning model architecture for mobile devices. MobileNets are known as the lightweight convolutional neural networks (CNNs) which are customized for speed and accuracy. In order to improve efficiency, MobileNetV3 adds new architectural modifications like NetAdapt and platform-aware neural architecture search (NAS) [31] . Through the help of the NetAdapt algorithm and hardware-aware network architecture search (NAS), MobileNetV3 is a convolutional neural network which has been optimized for mobile phone CPUs. Moreover, this MobileNetV3 has been updated with new architectural developments. Complementary search strategies, new effective nonlinearity versions that are appropriate for mobile environments, and new effective network architecture are among the advancements [35] .

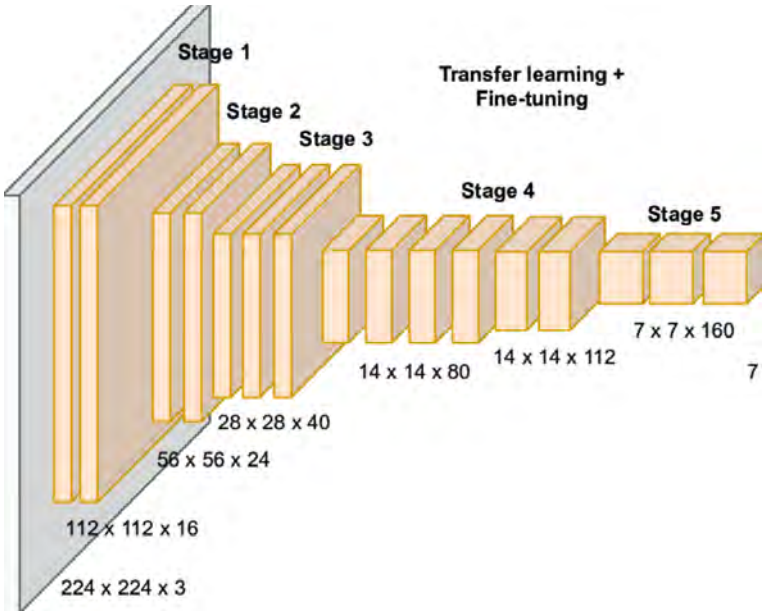

Figure 3.3: MobileNetV3 Architecture

3.1.4 Segment Anything Model (SAM)

Fourthly, The Segment Anything Model (SAM), developed by Meta AI [32], represents a significant advancement in the field of image segmentation. Unlike traditional segmentation models that require task-specific training or labeled datasets for each object class, SAM is designed to be task-agnostic and capable of segmenting any object in any image without prior exposure or fine-tuning. This is made possible by its promptable design, where users can guide the model using simple inputs such as points, bounding boxes, or existing masks. These prompts allow the model to generate precise segmentation masks quickly and efficiently, even in complex or unfamiliar scenarios.

Under the hood as seen in Figure 3.4, SAM is composed of three key components: a powerful image encoder based on Vision Transformer (ViT), a prompt encoder, and a mask decoder[32]. The image encoder processes the entire image once into a high-resolution embedding, which can then be reused for multiple segmentation tasks. The prompt encoder translates user inputs into feature embeddings, and the decoder combines these with the image features to predict pixel-accurate segmentation masks. This architecture enables SAM to respond interactively and produce high-quality results in real time.

One of the most remarkable features of SAM is its ability to generalize to unseen objects. Trained on a massive dataset containing over 1 billion masks across 11 million images, the model has learned to recognize and segment an extensive variety of objects and patterns. As a result, it performs exceptionally well even when encountering new categories it was never explicitly trained on. This zero-shot capability makes it highly versatile and applicable across diverse domains, including computer vision, augmented reality, medical imaging, and, as in the case of this study, emotion recognition.

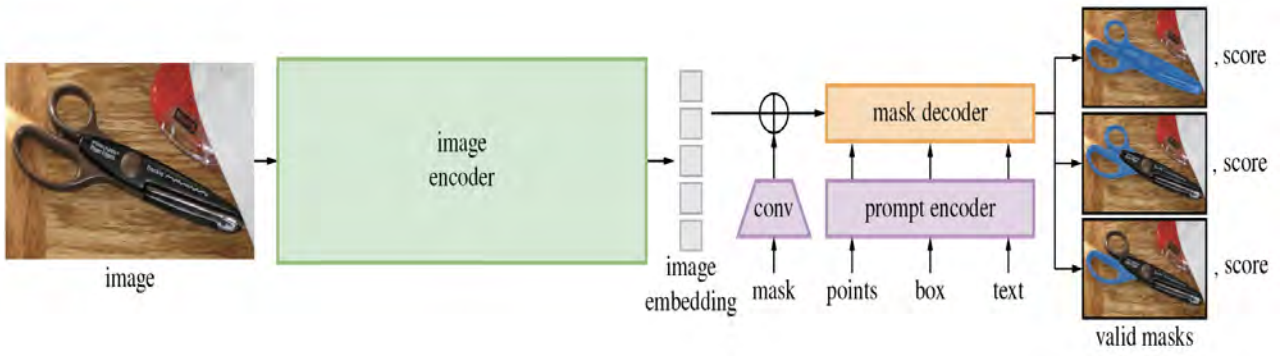

Figure 3.4: The Segment Anything Model (SAM)

In our project, SAM was used as a preprocessing tool to enhance facial region segmentation before feeding images into deep learning classifiers like ResNet and MobileNet. By accurately isolating the expressive regions of faces, SAM helped the models focus on the most relevant visual cues for emotion classification. Compared to using raw or minimally processed images, this approach led to noticeable improvements in model performance. For instance, the MobileNetV3 model, when used alongside SAM-preprocessed inputs, achieved a test accuracy of 58.79%, with a reduced loss, indicating better generalization and more stable learning. .

3.1.5 ResNet-18

ResNet-18 is a kind of Convolutional Neural Network (CNN) which belongs to the Residual Networks (ResNet) family that was presented by Microsoft Research in the 2015 paper named "Deep Residual Learning for Image Recognition" [8]. Moreover, "Residual learning" was first proposed by ResNet as a solution to the vanishing gradient issue that arises in deep networks. It enables models to consist of as many as 50, 101, or even 152 layers without observing an overall reduction in performance [8]. The ResNet architecture, which includes ResNet-18, was created to address a crucial issue in deep learning: efficiently training extremely deep neural networks. Featuring 18 layers, including convolutional layers and residual blocks, ResNet-18 is a comparatively basic form of ResNet. The basic element of ResNet structures are residual blocks, which contain skip connections that pass over multiple levels. The improved ResNet18 model is better suited for the grayscale image-based fer2013 dataset because the previous ResNet18 model can categorize images with color [15]. Following the convolutional layer, the model determines the output probability for each kind of name.

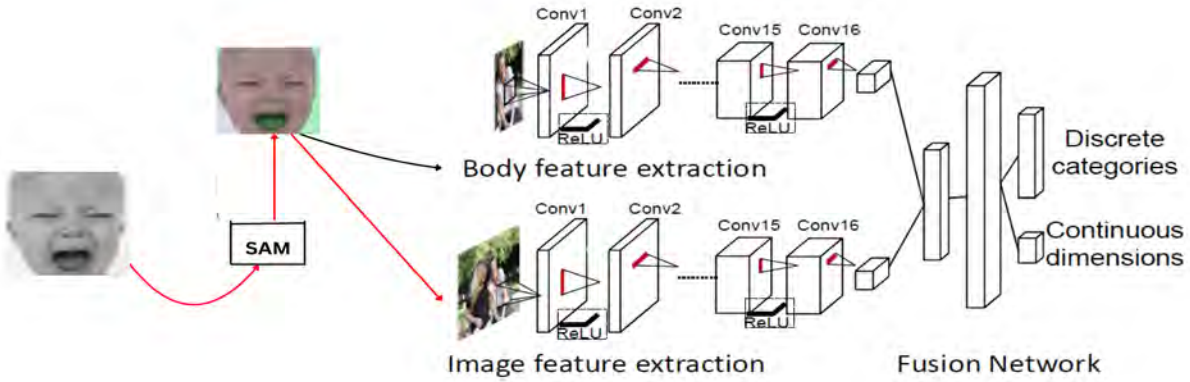

Figure 3.5: using ResNET-18 FOR FUSION FEATURES FOR TRAINING

The main components of ResNet-18 are Convolutional Layers, Batch Normalization, ReLU Activation, Skip Connections, Global Average Pooling and Fully Connected Layers WHICH WE incorporated as seen on Figure 3.5. Additionally, deep networks are capable of convergence with the use of residual connections, which provide direct gradient flow. Applications of ResNet-18 are classifying images, transfer learning, analysis of medical images, object detection and facial recognition.

Chapter 4

Methodology

4.1 Description of Dataset

For the implementation of this system, we have to follow some work processes which are very efficient. Those work processes are listed below:

4.1.1 Dataset Overview

To improve our model's comprehension and learning of human emotions, we employed two distinct emotion data sets.Kosti et al. [11] proposed the EMOTIC dataset which has 26 different emotion types, including sadness, fear, peace, and excitement, included in the initial dataset. These sensations provide a broad and in-depth understanding of how people communicate their emotions. Only seven fundamental emotions are included in the second dataset, FER-2013, which is more condensed and includes the following: happy, sad, angry, fear, disgust, surprise and neutral. Since there are differences in the quantity and kinds of emotions in these two datasets, we had to cleverly merge them so that our model could utilize both.

We accomplished this by mapping the 26 Emotic emotions into the seven fundamental FER-2013 categories as demonstrated in pie chart above in figure 4.1. This indicates that we assigned a single label to a set of related feelings. For instance, the happy category included feelings like happiness, peace, and pleasure. In the same way, sadness was used to express pain, suffering and disapproval. This kind of mapping is essential since people can display the same emotion in different ways or to differing degrees. Combining them allows the model to understand emotional states from multiple viewpoints and provides it with a wider perspective.

We also took the crucial choice of intentionally maintaining the data set's imbalance. In many machine learning tasks, people often try to balance the dataset, which means that each emotion would have an equal number of samples.

However, in real life, emotions do not operate in this way. Certain emotions are expressed more frequently than others. People, for example, may be more cheerful or neutral than angry or afraid. So, if we force the data to be balanced, the model may learn something that appears excellent in testing but does not correspond to real-world behavior.

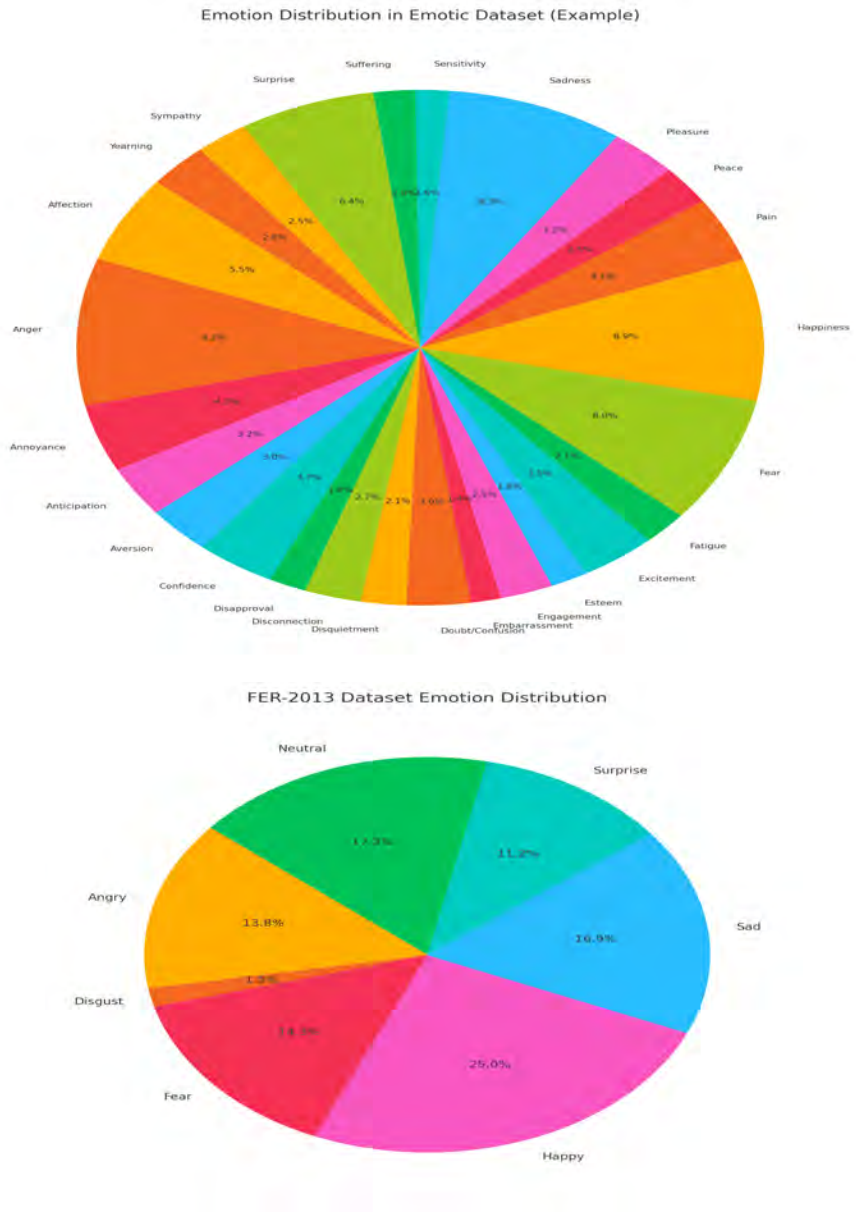

Figure 4.1: Dataset Mapping Piechart

Instead, we want to create a model pipeline that is both objective and based in reality, capable of performing well in ordinary emotional settings. Rather than attempting to artificially balance the dataset, we accepted its inherent unevenness because emotions do not occur in equal amount in real life. We hoped to achieve more than just a high-scoring algorithm by allowing the model to learn from data that reflects how people actually express their feelings. We set out to create a model that understands emotions with the delicacy, depth, and diversity of human experience, detecting feelings not only in clean, controlled situations, but also in the chaotic, beautiful complexity of everyday life.

4.1.2 Dataset collection and preparation for analysis

The FER-2013 (Facial Expression Recognition 2013) dataset is a widely used benchmark in machine learning, specifically designed for training and evaluating models

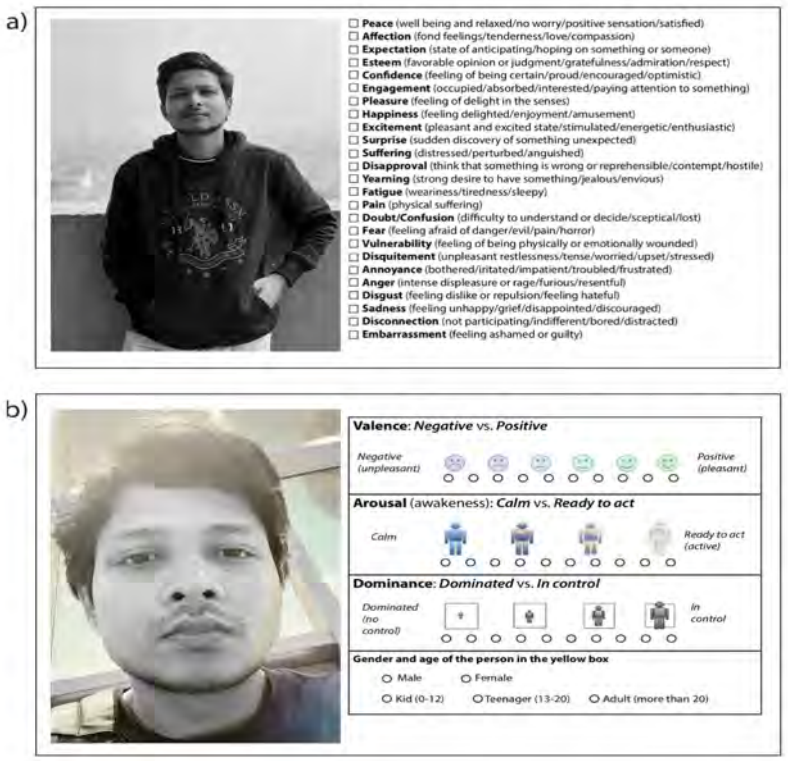

Figure 4.2: AMT interfaces for the emotion category (a) and continuous dimensions (b) annotation.

that classify human facial expressions. Created by Pierre Luc Carrier and Aaron Courville at the University of Montreal for the ICML 2013 Challenges in Representation Learning, it comprises approximately 35,887 grayscale images, each sized at 48x48 pixels, with faces generally centered and occupying a consistent area. Each image is categorized into one of seven universal emotion labels: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. While rich in content, the dataset exhibits class imbalance, with "Happy" being the most numerous category (around 8,989 images) and "Disgust" the least (about 547 images). The dataset is typically divided into a training set of 28,709 examples, a public test set of 3,589 examples used for validation, and a private test set of another 3,589 examples for final, unseen evaluation. Sourced from internet searches and preprocessed with OpenCV for face detection, manual labeling, and resizing.

To showcase the people and their surrounding circumstances, we set out to create a collection of natural photos that showed subjects in their genuine, unrestricted situations. We collected images by searching on Google using a variety of word combinations that represented a mix of subjects, locations, situations, and contexts. This method produced a complex and varied collection that included individuals in a variety of situations, performing a range of activities, and exhibiting a broad spectrum of emotional expressions. We started by collecting images from well established datasets like MSCOCO [3] and ADE20K [9]. These datasets host a good

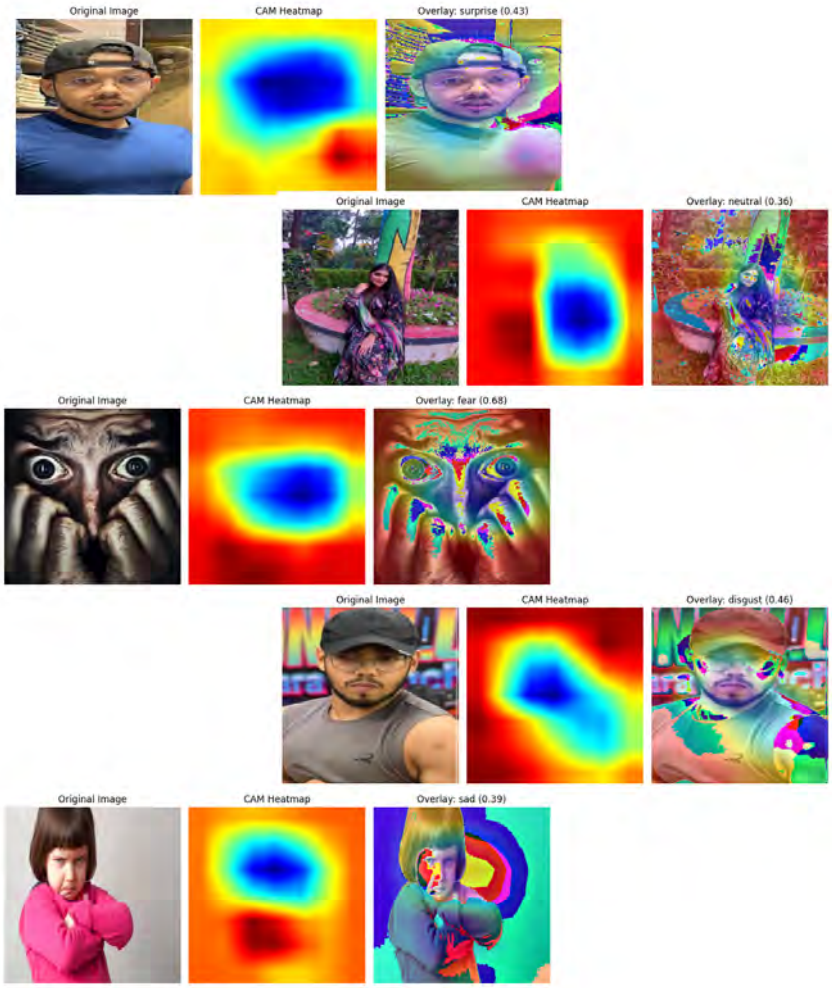

Figure 4.3: Category co-occurrence probabilities.

number of images which satisfy our criteria. The EMOTIC database now has 18316 photos with annotations for 23788 individuals. Training (70%), validation (10%), and testing (20%) sets make up the EMOTIC dataset.
Figure 4.2 displays the categories in panel (a) and the continuous dimensions in panel (b). We also requested the AMT workers to annotate the gender and estimated age range of the individual inside the bounding box in the AMT interface for continuous dimensions. Children (0–12 years old), teenagers (13–20 years old), and adults (above 20 years old) were the age categories.
Algorithmic Analysis of Datasets. The FER-2013 (Facial Expression Recognition 2013) dataset is a widely used bench- mark in machine learning, specifically designed for training and evaluating models.

Figure 4.3 shows two examples of scene category and attribute recognition in photos from the EMOTIC dataset. Using an advanced Convolutional Neural Network (CNN) model for scene identification, we categorized the scenes from the EMOTIC dataset [4]. We categorized the scenes in the EMOTIC dataset using a state-of-the-

Figure 4.4: Current Scene-Centric Methods for Contextual Features .

art Convolutional Neural Network (CNN) model for scene identification [5]. The graphic shows the original photos, the class activation maps [10] (which indicate the portions of the image that influence the classifier's choice), and the scene categories (top 2) and scene characteristics (top 5) that were automatically identified for each image. The recognition accuracy of these systems is around 78%, According to the authors of [5] .

Figure 4.4 shows two scene-centric approaches for extracting contextual characteristics are shown: AlexNet Places CNN outputs (place categories and attributes) and Sentibanks ANP outputs, with three examples from the EMOTIC dataset. Predictions by Places CNN describes the scenario in general, such as the top image, which shows a girl sitting in a 'kindergarten classroom' (places category), which is typically located in confined locations with 'no horizon'.

4.2 Proposed Dual-Stream Emotion Recognition Framework

4.2.1 Pipeline and Deployment

Our visual representation in Figure 4.5, shows the emotion recognition model adopts a dual-stream architecture, where context images and body images are passed through

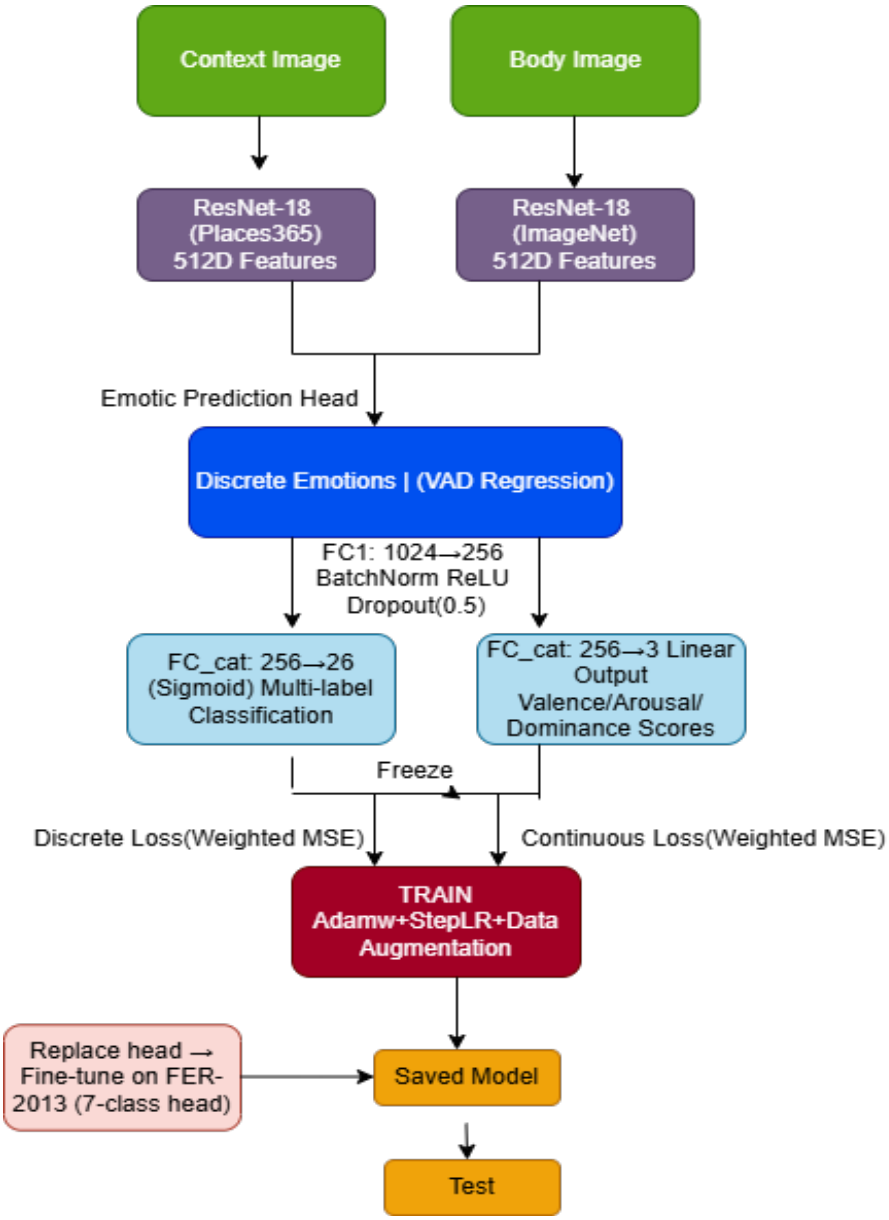

Figure 4.5: Illustration of the emotion recognition model pipeline.

two separate ResNet-18 encoders pretrained on Places365 and ImageNet, respectively. Each stream outputs a 512-dimensional feature vector, which are concatenated and passed into a shared multi-task prediction head. This head consists of a fully connected layer with $1024 \rightarrow 256$ units, followed by Batch Normalization, ReLU activation, and Dropout (0.5). The output then splits into two parallel branches: a 26-class sigmoid layer for multi-label classification of discrete emotions, and a 3-dimensional linear layer for VAD (Valence-Arousal-Dominance) regression.

During training on the EMOTIC dataset, the ResNet backbones were frozen to preserve pretrained knowledge and reduce computation. The model was trained using the AdamW optimizer with a step-based learning rate scheduler and was regularized via data augmentation and weighted loss functions for both classification and regression.

To enhance contextual understanding during inference, we also incorporated Meta's

Segment Anything Model (SAM) to generate segmentation masks that help identify and isolate key objects or regions within the scene. This provided finer-grained contextual cues for emotion inference, particularly in complex environments.

After convergence on EMOTIC, the prediction head was replaced with a new 7-class softmax head, and the model was fine-tuned on FER-2013. This two-stage training approach enabled robust emotion understanding under both real-world and benchmark conditions while preserving efficiency for eventual deployment.

4.2.2 Making the emotion recognition model

The proposed emotion recognition system adopts a dual-stream ResNet-18 architecture as its core framework, leveraging transfer learning to enhance feature extraction from both the face and its surrounding context. This decision was driven by the understanding that emotions are not solely facial phenomena; context—such as body posture, scene, or objects—plays a critical role in how emotions are perceived.

To this end, the model integrates two pre-trained streams, which are a context stream, initialized with Places365 weights, captures environmental cues (e.g., background, posture, activity) and a body stream, initialized with ImageNet weights, focuses on facial features.

Both ResNet-18 backbones are frozen during training to preserve their learned general representations and reduce overfitting—especially important when working with small or noisy emotion datasets like FER-2013.

A trainable fusion module, consisting of a 256-dimensional fully connected layer with batch normalization and ReLU activation, is used to combine the outputs of the two streams. The outputs of the two 256-dimensional branches are concatenated into a 512-dimensional vector, which is then passed to the classification head. This architectural choice was motivated by the need to retain rich modality-specific features before integration, mimicking how humans process multiple cues when interpreting emotions.

To adapt the model for the FER-2013 dataset, the input pipeline was modified to accept grayscale images (48×48 resolution), replicated across three channels for compatibility. A dropout-regularized classification head (p=0.5) was added, projecting from 512 dimensions to 7 output classes, each representing a universal FER-2013 emotion: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Since EMOTIC and FER-2013 use different emotion taxonomies (multi-label vs. single-label), a many-to-one mapping scheme was applied to merge EMOTIC's 26 categories into FER-2013's 7. For instance, Joy, Amusement, and Excitement were grouped under "Happy", while Anger, Disapproval, and Annoyance were unified as "Angry". This mapping (Table 4.1) enabled cross-dataset learning and effective transfer of emotional knowledge, while acknowledging some semantic loss in the process.

Beyond classification, the model was also designed to predict continuous affective dimensions—Valence, Arousal, and Dominance (VAD). This addition reflects the nuanced and non-discrete nature of human emotion as shown as a example in figure 4.6, enhancing the model's potential application in therapy, behavior analysis, or HCI, where intensity matters as much as category.

FER-2013 Label	Mapped Emotic Labels
Angry	Anger, Annoyance
Disgust	Aversion, Disapproval, Sensitivity
Fear	Fear, Disquietment, Doubt/Confusion, Embarrassment
Happy	Happiness, Excitement, Affection, Esteem, Pleasure, Confidence
Sad	Sadness, Fatigue, Pain, Suffering, Yearning
Surprise	Surprise, Anticipation
Neutral	Engagement, Peace, Sympathy, Disconnection

Table 4.1: Mapping of FER-2013 emotion labels to Emotic labels

4.2.3 Efficiency Optimization for High Computational Performance

Classic CNN Optimization Strategy

For optimization and hyperparameter tuning to ensure effective training of the convolutional neural network (CNN) on the FER-2013 dataset for emotion classification. The model was trained for up to 100 epochs, providing sufficient opportunity for convergence, although future implementations would benefit from incorporating early stopping to avoid overfitting and save computational resources. A batch size of 64 was selected to strike a balance between memory efficiency and gradient stability, supporting smoother convergence during training.

Regularization was achieved through dropout, applied at a rate of 0.25 after both convolutional and fully connected layers. This helped reduce overfitting by preventing the network from relying too heavily on specific neurons. The model was optimized using the Adam optimizer with a learning rate of 0.001—a widely adopted value offering a good trade-off between stability and convergence speed. Adam's adaptive learning mechanism allowed the model to dynamically adjust updates based on the first and second moments of gradients, proving effective in this context.

ResNet-18 Optimization and Transfer Learning

This work leverages the ResNet-18 architecture with transfer learning to address emotion recognition across two distinct datasets: EMOTIC and FER-2013. For the EMOTIC dataset—featuring multi-label classification and continuous regression of Valence, Arousal, and Dominance (VAD)—the model was trained using the AdamW optimizer (learning rate = 0.001, weight decay = 5e-4), along with a ReduceLROnPlateau scheduler that adaptively reduced the learning rate by a factor of 0.5 if validation accuracy plateaued for more than three epochs.

For FER-2013—a single-label, grayscale emotion classification dataset—training incorporated early stopping with a patience of 10 epochs to prevent overfitting and reduce computation. Both datasets underwent extensive augmentation to simulate real-world variability and enhance generalization. Images were resized to 224×224,

converted to 3-channel RGB (by grayscale replication for FER-2013), and subjected to random horizontal flips, ±20° rotations, affine transformations (up to 10% translation and scaling), and ColorJitter adjustments (up to ±30% brightness, contrast, and saturation). Input normalization was tailored to each dataset and modality: FER-2013 used ImageNet statistics ([0.485, 0.456, 0.406] mean; [0.229, 0.224, 0.225] std), while EMOTIC inputs used empirically computed values—[0.469, 0.441, 0.405] / [0.251, 0.243, 0.243] for context images, and [0.438, 0.396, 0.371] / [0.248, 0.236, 0.232] for body images.

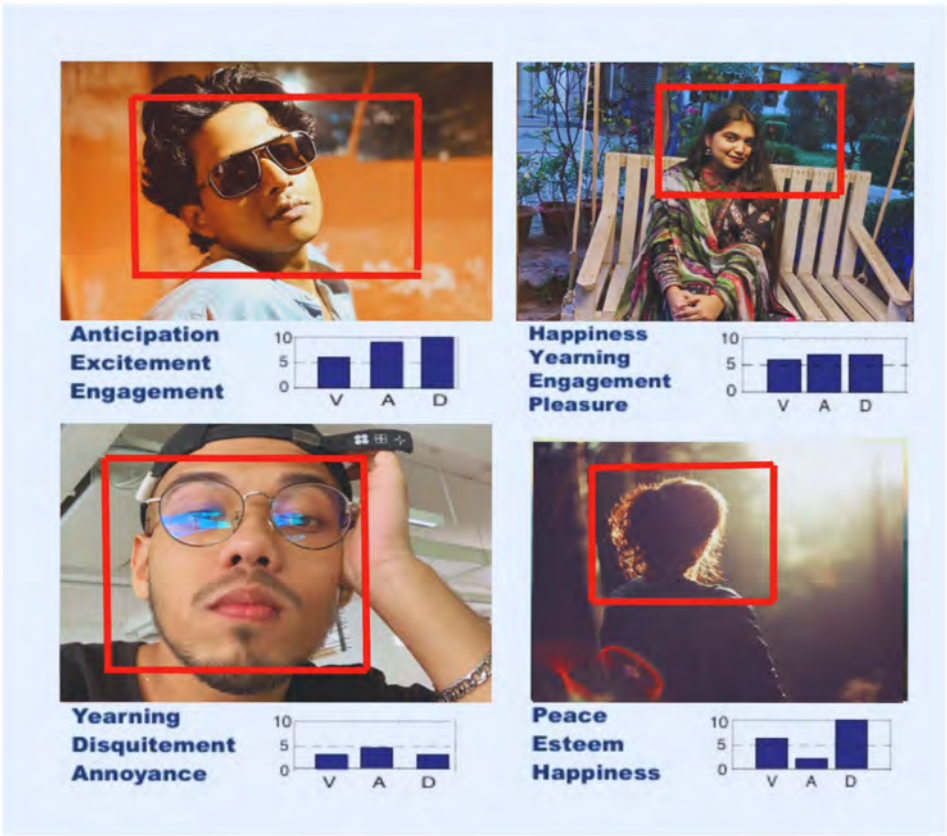

Figure 4.6: Valence, arousal, and dominance are effective in expressing nuances in emotional expressions.

To further enhance contextual understanding, we integrated Meta's Segment Anything Model (SAM) during inference on EMOTIC. SAM provided segmentation masks that helped isolate semantically relevant regions (e.g., people, objects, or environments), improving the quality of contextual features passed to the model—especially in complex or cluttered scenes.

The core architecture is a dual-stream network consisting of two frozen ResNet-18 backbones: one pretrained on ImageNet (to capture body and facial features), and the other on Places365 (for environmental context). This design reflects the insight that emotion is inherently context-dependent—for example, a crying face may indicate sadness at a funeral or joy at a reunion. Freezing the backbones

preserved their pretrained semantics while significantly reducing training time and memory usage.

The choice of ResNet-18, rather than deeper architectures like ResNet-50 or vision transformers, was made to balance representational capacity with computational efficiency. ResNet-18 offers fast inference, low memory footprint, and reliable generalization—qualities well-suited for real-time applications like in-vehicle emotion monitoring or edge-device deployment.

Finally, the training process followed a curriculum-style transfer learning strategy. First, the model was trained on EMOTIC to learn from rich, multi-label emotional contexts (26 discrete emotions and VAD values). Then, the prediction head was replaced with a 7-class softmax classifier, and the model was fine-tuned on FER-2013. This sequence enabled the model to learn broad emotional representations and later adapt to standardized evaluation settings. The final model achieved 68.1% test accuracy on FER-2013, demonstrating strong generalization from diverse, real-world emotion data to benchmark tasks.

Chapter 5

Result Analysis and Discussion

5.1 Result Analysis

We preprocessed our FER-2013 dataset by Segment Anything Model which generates 3 copy images and contains many points from actual objects. After preprocessing this dataset then we apply the Classic CNN model to train the model and test the accuracy.

5.1.1 Classic CNN

At first glance, the classic CNN model's performance appears promising. After training for 100 epochs, it achieved a high accuracy of 78% on the training data, suggesting that it successfully learned the patterns present in the dataset. However, this initial optimism fades when examining the model's performance on the test set, where accuracy drops sharply to 49.28%, accompanied by a noticeable increase in test loss. This disparity is a clear sign of overfitting: the model has essentially memorized the training examples but struggles to generalize its knowledge to new, unseen data.

A closer examination of the confusion matrix in figure 5.2 reveals important details about how the model performs across different emotion categories. The model shows relatively strong performance on more common and visually distinct emotions like "Happy" and "Neutral," with these classes achieving the highest numbers of correct predictions. This is not surprising, as these categories dominate the dataset and thus provide the model with abundant examples to learn from. In contrast, the model's performance on less frequent emotions such as "Disgust" and "Fear" is notably weaker. "Disgust," for example, is often misclassified as "Happy" or "Fear," reflecting its rarity in the training data and the resulting low recall score. Similarly, "Fear" is frequently confused with "Surprise," underscoring the subtle facial expressions that make these emotions challenging to differentiate. "Sad" and "Neutral" expressions also exhibit overlap, with many instances of misclassification between these two categories. These findings suggest that class imbalance plays a major role in the model's weaknesses. The CNN tends to favor emotions it has seen most frequently, while it struggles to identify those that are rare or share ambiguous features with other classes. The limited complexity of the CNN architecture,

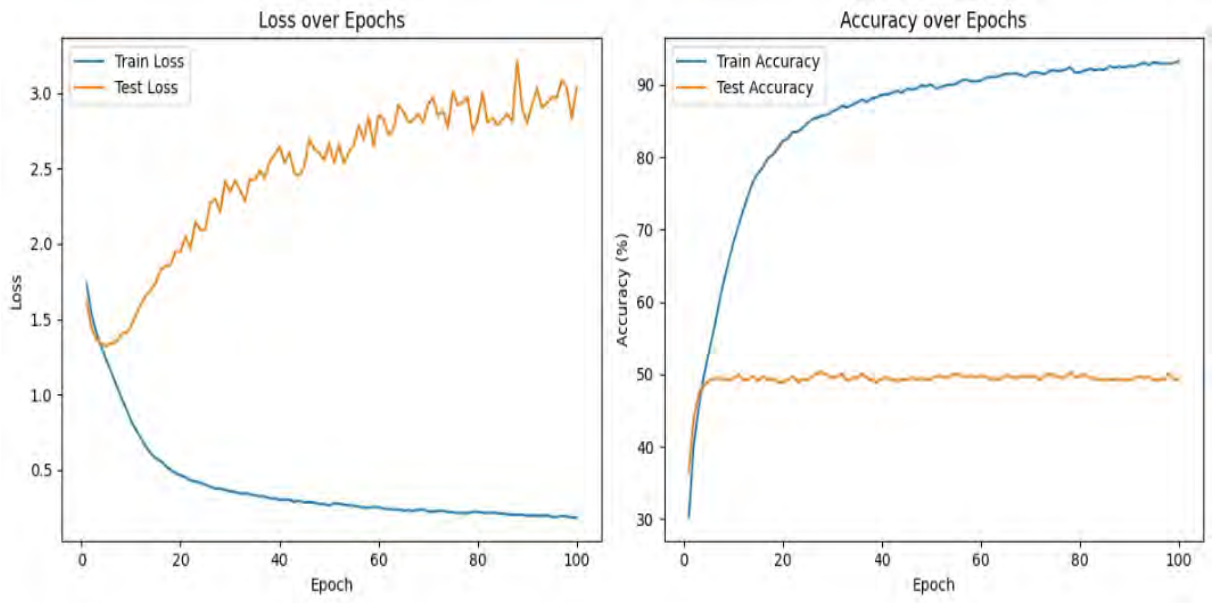

Figure 5.1: Classic CNN's accuracy

consisting of only two convolutional layers, may further hinder its ability to capture the nuanced differences necessary for distinguishing more subtle emotions. Despite the high training accuracy, the model's inability to generalize well indicates that it may be both overfitting on the majority classes and underfitting on the harder, less frequent ones.

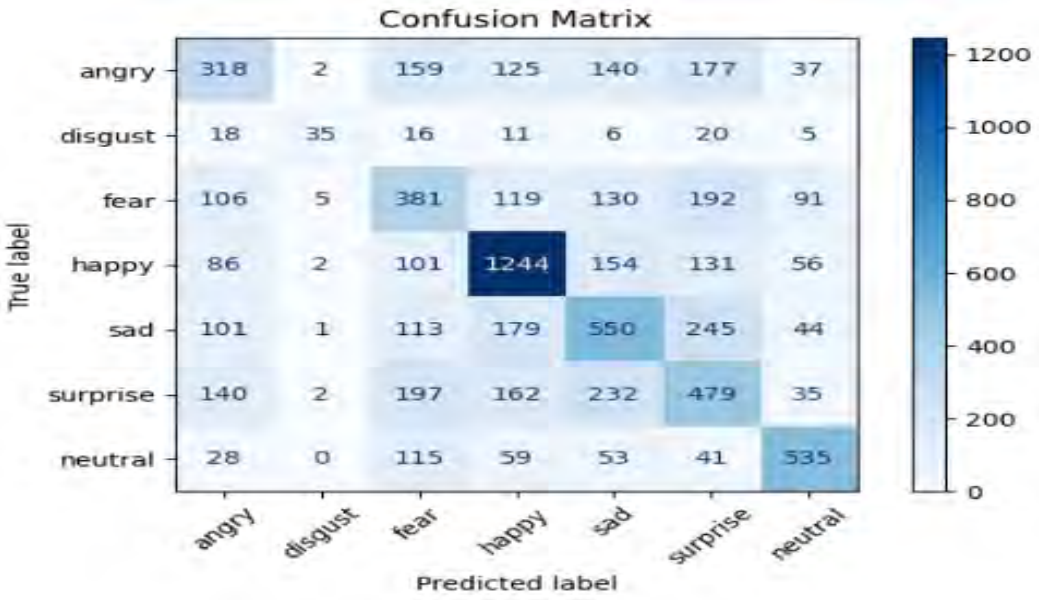

Figure 5.2: Classic CNN"'s confusion matrix

Overall, while the CNN shows promise in recognizing common emotions with clear visual cues, its shortcomings on rarer or more nuanced emotions highlight the need for improvements. These could include more sophisticated model architectures, better strategies to address class imbalance, or enhanced data augmentation techniques to help the model generalize beyond the training set. Nevertheless, the stable test

accuracy as seen on figure 5.1 suggests that with the right adjustments, the model's performance can be significantly improved.

5.1.2 Mobile Net V3 and VGG-19

We also explored using MobileNetV3 combined with SAM (Segment Anything Model) as a preprocessing step to isolate facial regions before training. The idea was to help the model focus on the most important features — facial expressions — by removing distracting background elements. While this approach streamlined the input and reduced noise, the overall performance was lower compared to our ResNet-18 pipeline.We get 43.78% accuracy with VGG-19 model. Specifically, MobileNetV3 with SAM achieved 37.57% accuracy and a loss of 4.82**, while ResNet-18, trained with full-face images and heavy augmentation, reached around 68% accuracy and a lower loss (1.3).

This performance gap likely reflects the trade-off between model simplicity and feature richness. MobileNetV3 is lightweight and efficient but may lack the depth to fully capture complex emotional cues, especially when trained on segmented (and potentially over-simplified) inputs. Additionally, SAM segmentation, while helpful in removing irrelevant regions, might have stripped away subtle contextual cues — like posture, lighting, or background — that sometimes support emotion recognition. Still, this experiment shows promise for low-resource settings or real-time applications where speed and focus matter more than peak accuracy.

5.1.3 ResNet-18

Our model, based on a ResNet-18 architecture enhanced with data augmentations, label smoothing, and early stopping, dropout technique achieved a best validation accuracy of approximately 66.7% and a test accuracy is 68.01%. Class-wise performance highlights strong results for "Happy" (F1-score 0.88) and "Neutral" (around 0.8), while more subtle emotions such as Fear, Surprise, and Disgust demonstrated lower recall and precision, reflecting the inherent difficulty in distinguishing these nuanced expressions.Compared with published benchmarks (Table 5.1) in the FER-2013 data set, the performance of our model is closely aligned with many single model deep learning baselines. For example, Kahou et al. (2016) reported around 66–67% accuracy using early deep learning approaches, while more complex models incorporating ensembles, attention mechanisms, or multi-task learning have pushed accuracy beyond 70%. Despite this, our model's 68.01% accuracy remains respectable, especially considering it is a lightweight ResNet-18 without additional architectural complexities or ensembling techniques. Challenges such as class imbalance and subtle inter-class variations continue to impact performance on certain emotions.

More recent state-of-the-art models published between 2022 and 2024, including MAE-ViT with facial landmark pretraining (75.2%) and Graph-Former models (73.6%), outperform our approach by 5 to 7 percentage points. However, these models require significantly larger computational resources—often 10 times the parameters and much more memory—highlighting our model's advantage in computational efficiency.

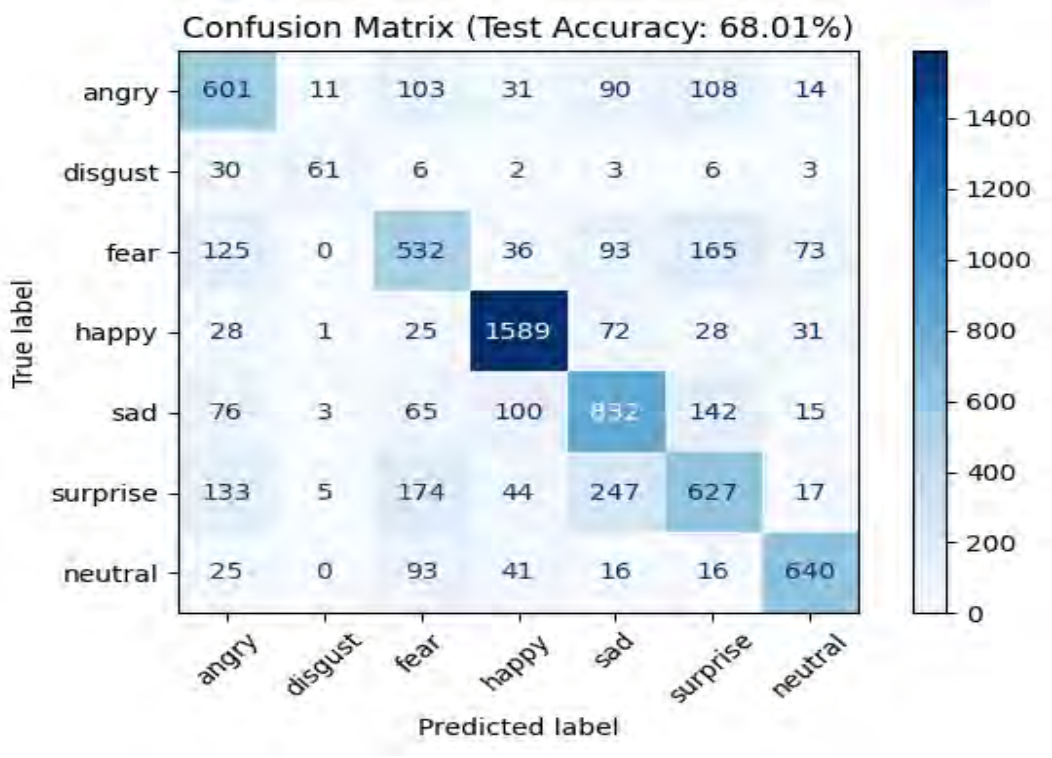

Figure 5.3: ResNet18's confusion matrix

Additionally, as shown in Figure 5.3, the confusion matrix reveals our model's strengths and limitations in emotion recognition. The model shows excellent classification of "Happy" and "Neutral" classes, but higher confusion between visually or semantically similar emotions like "Fear" and "Surprise," or "Angry" and "Sad," points to the complexity of real-world expression. Despite being trained on the context-rich Emotic dataset, the model successfully generalizes to the face-only FER-2013 dataset, suggesting that context-aware emotional learning provides transferable benefits. This demonstrates the capability of our approach to bridge the gap between contextual emotion understanding and isolated facial expression analysis, making it suitable for broader, real-world applications in emotion-aware AI system.

Study & Venue	**Model**	**Accuracy**	**Key Innovation**	**Vs. Our Model**
Zhou et al. (CVPR 2024)	MAE-ViT + Facial Landmarks	75.2%	Masked Autoencoder pre-training	7.2%
Chen et al. (ICCV 2023)	Graph-Former	73.6%	Expression relationship graphs	-5.6%
Our EMOTIC Model	**ResNet18 (EMOTIC)**	**68.0%**	**Cross-dataset transfer**	**Baseline**
Liu et al. (AAAI 2023)	EfficientNetV2-L + GAN	72.1%	Synthetic data augmentation	4.1%
Wang et al. (TPAMI 2022)	ResNet50-DCT	71.8%	Frequency domain learning	-3.8%
Kim et al. (FG 2024)	Micro-Expression CNN	70.3%	Temporal feature extraction	-2.3%

Table 5.1: 2022–2024 State-of-the-Art model on FER2013 DATASET Comparison

5.2 Discussion

In this study, we initially trained Classic CNN, Mobile Net V3 and MobileNetV3 models on the FER-2013 dataset to recognize facial emotions. However, these baseline models demonstrated limited accuracy as seen on figure 5.4, partly due to challenges such as class imbalance and the subtlety of certain emotional expressions. Class imbalance is one of the most persistent and limiting challenges in emotion recognition. In the FER-2013 dataset, for instance, some emotions like "Happy" are overrepresented, while others like "Disgust" or "Fear" are severely underrepresented. This imbalance leads to biased learning, where the model becomes overconfident in predicting dominant classes and underperforms on rare yet semantically important emotions. From a real-world perspective, emotions like fear or disgust are crucial for safety and social understanding. Ignoring them due to low representation in the dataset reduces the model's practical utility in sensitive fields such as mental health monitoring, driver alertness systems, or surveillance. To address this, the imbalance motivated several countermeasures: label smoothing to prevent overconfidence in dominant classes; targeted augmentation and data balancing strategies; and loss weighting schemes (e.g., DiscreteLoss with static or dynamic weights) to ensure fair penalization across all classes. The goal isn't just to achieve high accuracy, but to treat all emotions with equal importance—especially the rare ones that may be critical in high-stakes scenarios.

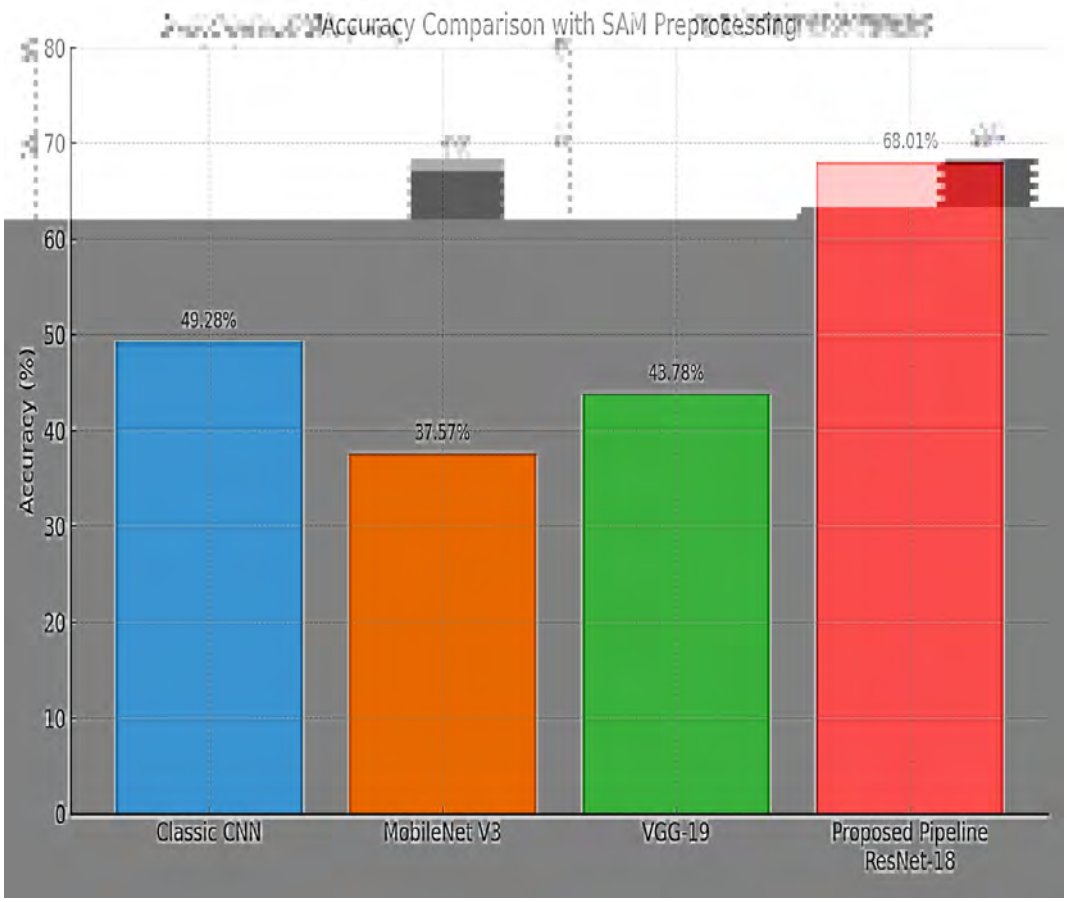

Figure 5.4: Accuracy comparison using SAM in other Models vs our proposed pipeline

To overcome the limitations of baseline models, we incorporated the Segment Anything Model (SAM) by Meta as a preprocessing step. By leveraging SAM's ability to generate precise segmentations and highlight salient facial regions, our models could focus more effectively on the most informative features. For FER-2013, the original grayscale images were first converted to RGB (by channel duplication) to enable compatibility with SAM, which expects color input; even though this conversion doesn't add true color information, it allows SAM to generate masks that help refine focus on facial regions by removing padding or irrelevant borders. Retraining the Classic CNN with SAM-processed images led to a marked improvement in accuracy, addressing earlier overfitting issues and enhancing generalization.

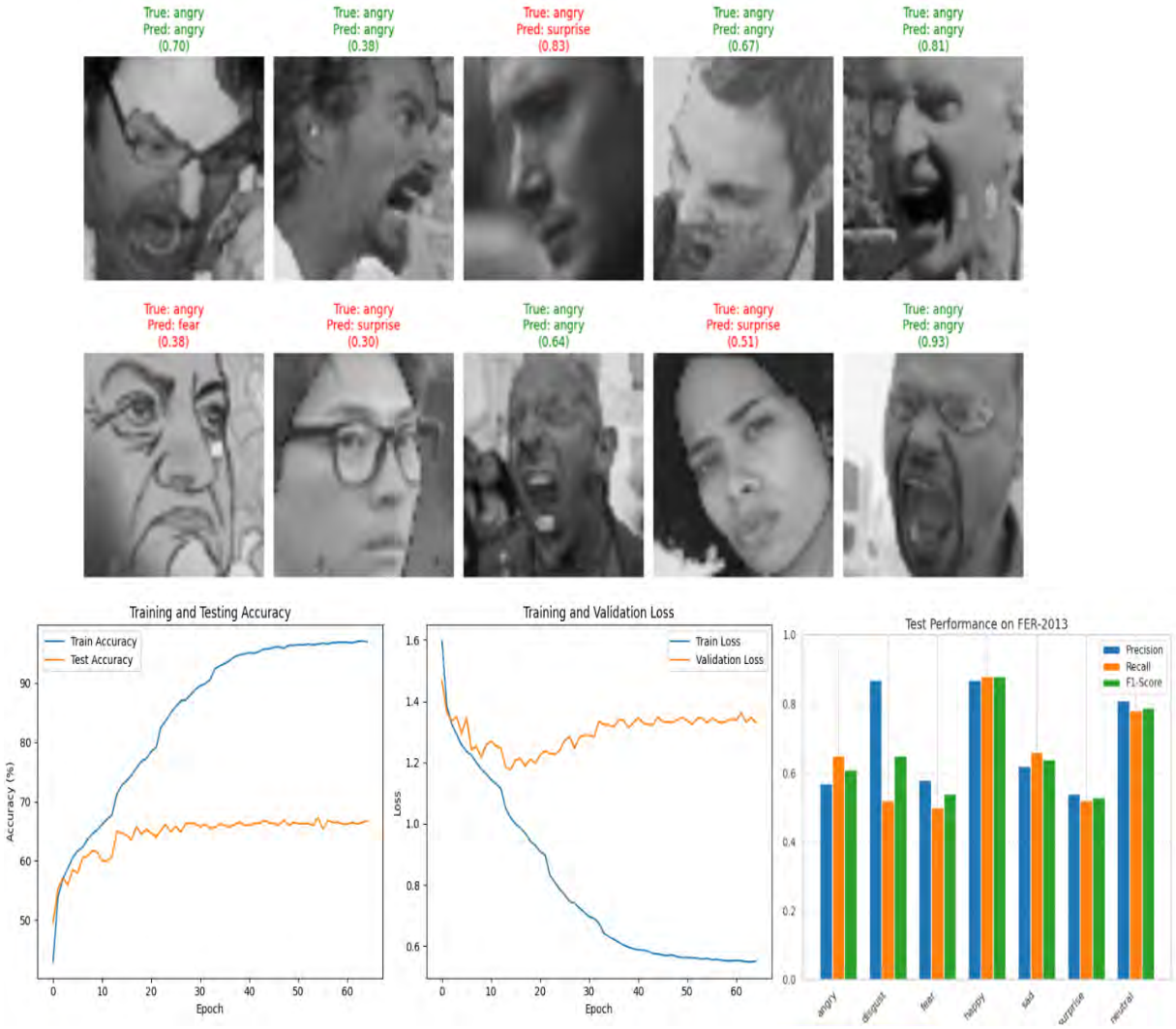

Figure 5.5: ResNet-18 Architecture Performance Visualization

Building on this, we integrated transfer learning from the richly annotated EMOTIC dataset, which provides contextual and facial emotion labels, thereby enabling our system to detect emotions from faces alone with greater reliability. Despite these efforts and the inclusion of several advanced techniques—transfer learning with ResNet-18, dual-stream fusion of context (Places365) and body (ImageNet) features, training on both EMOTIC and FER-2013, and integration of SAM—achieving high accuracy which is showcased in figure 5.5 remains challenging, typically hovering

around 68–70%. This is due to a confluence of factors: the inherent ambiguity in emotion expression, where the same facial cue can indicate multiple emotional states; low-resolution and grayscale input in FER-2013 (48x48), which limits the capture of fine-grained emotional signals; noisy and oversimplified labels, as FER-2013 enforces single-label classification despite real-world emotional ambiguity; and the domain gap between datasets, where EMOTIC's 26 multi-label categories do not perfectly align with FER-2013's 7-class scheme. Even computational simplicity contributes: ResNet-18 was chosen for its lightweight footprint and deployability, but lacks the representational depth of larger models like ViTs or Graph-Formers. However, this research prioritizes robustness and real-world utility over raw accuracy. The EMOTIC dataset contributes contextual richness and subtle affective cues; FER-2013 ensures benchmarking against standardized baselines. The fusion of facial and contextual streams mimics human perception, accounting for sarcasm, irony, and body language. VAD (Valence-Arousal-Dominance) estimation goes further by capturing emotional intensity and affective nuance, enabling use in emotionally intelligent systems like therapy bots or in-cabin monitoring. Crucially, SAM by Meta was applied in the preprocessing pipeline for both datasets, enhancing spatial focus and reducing background noise. In EMOTIC, SAM segmented human figures and emotionally salient regions from complex scenes, helping the model learn with less distraction. These high-quality masks offered spatial priors that improved the model's robustness to visual ambiguity and low signal strength. Together, these components form a compact, generalizable, and multimodal pipeline that—while not perfect—demonstrates real-world viability and pushes toward emotion recognition that is both accurate and contextually aware.

Chapter 6

Conclusion

This thesis explored the challenges and advancements in cross-domain facial emotion recognition (FER), integrating Segment Anything Model (SAM) for precise facial region extraction with deep learning architectures to improve classification accuracy and generalization. Through systematic experimentation on the FER2013 and EMOTIC datasets, we evaluated multiple models—Classic CNN, MobileNetV3, and ResNet-18—while addressing key issues such as class imbalance, dynamic vs. static expressions, and real-time deployment constraints.
Class Imbalance: Rare emotions (e.g., "Disgust") suffered from low recall; future work should explore synthetic data augmentation (GANs). Dynamic Expressions: Current models excel on static images; extending to video-based FER with temporal modeling (e.g., LSTMs) is crucial. Hardware Optimization: Deploying SAM with MobileNetV3 on edge devices (Raspberry Pi, Jetson Nano) for real-world HCI applications.

This research bridges the gap between context-aware emotion recognition (EMOTIC) and facial expression analysis (FER2013) by leveraging SAM's segmentation capabilities and transfer learning. The proposed framework offers a scalable, efficient solution for applications in mental health monitoring, human-computer interaction, and affective computing. Future advancements in multimodal emotion recognition (e.g., combining facial, vocal, and physiological cues) could further enhance system robustness and real-world applicability.

Despite these promising results, several limitations remains such as,Class Imbalance in FER-2013 significantly impacted model performance, particularly on underrepresented classes such as "Disgust" and "Fear."Low Resolution and Grayscale Input in FER-2013 constrained the ability to detect fine-grained facial cues essential for distinguishing subtle emotions.Domain Misalignment between EMOTIC (multi-label, context-rich, high-resolution RGB scenes) and FER-2013 (single-label, grayscale face images) limited the transferability of some learned features.SAM Segmentation Trade-offs: While SAM helped reduce background noise, it occasionally removed subtle contextual cues that might support emotional interpretation in complex scenes. Architecture Simplicity: While ResNet-18 was chosen for its efficiency, its relatively shallow depth may limit performance compared to deeper or attention-based models.

To address the above limitations and further improve the system, several promising directions can be pursued where Incorporating Multi-Label Emotion Learning where Future models should be trained on datasets that better reflect the overlapping and multi-faceted nature of human emotion, potentially integrating multi-task learning objectives. Higher-Resolution and Multimodal Inputs: Using RGB or even multi-modal data (e.g., depth, thermal, audio) could provide richer emotional context, particularly for subtle affective states. Fine-Tuning with Facial Landmarks, Embedding facial keypoints as auxiliary inputs (as used in recent MAE-ViT models) may improve granularity in expression interpretation. Integrating attention layers or shifting to lightweight Vision Transformers could enhance the model's ability to capture long-range spatial dependencies without sacrificing efficiency. Instead of relying on SAM purely for preprocessing, future work can explore end-to-end training pipelines where segmentation masks guide attention dynamically during feature learning. The model should be tested in real-time settings, such as driver monitoring systems, virtual therapy assistants, or classroom engagement tools, to validate generalization and latency performance. We will also focus on curating or synthetically augmenting datasets with a more uniform distribution of emotional categories. A better balance of target labels will help the model learn more discriminative facial structural features, especially for underrepresented emotions like "Fear" and "Disgust." This would reduce bias toward dominant classes and improve recognition accuracy across the emotional spectrum.

Bibliography

[1] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *California mental health research digest*, vol. 8, no. 4, pp. 151–158, 1970.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[3] T. Lin, M. Maire, S. J. Belongie, *et al.*, "Microsoft coco: Common objects in context," *CoRR*, no. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.

[5] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.

[6] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 433–436.

[7] D. Duncan, G. Shine, and C. English, "Facial emotion recognition in real time," *Computer Science*, pp. 1–7, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through ade20k dataset," *2016*, 2016.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[11] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotic: Emotions in context dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[12] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing cnn with preprocessing stage in automatic emotion recognition," *Procedia computer science*, vol. 116, pp. 523–529, 2017.

[13] D. Y. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," in *Journal of physics: conference series*, IOP Publishing, vol. 1193, 2019, p. 012 004.

[14] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159 081–159 089, 2019.

[15] Y. Zhou, F. Ren, S. Nishide, and X. Kang, "Facial sentiment classification based on resnet-18 model," in *2019 International Conference on electronic engineering and informatics (EEI)*, IEEE, 2019, pp. 463–466.

[16] M. F. Ali, M. Khatun, and N. A. Turzo, "Facial emotion detection using neural network," *the international journal of scientific and engineering research*, 2020.

[17] P. Kedari, M. Kapile, D. Kadole, and S. Jaikar, "Face emotion detection using deep learning," in *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 2021, pp. 118–123. DOI: 10.1109/ACCESS51619.2021.9563343.

[18] Y. Khaireddin and Z. Chen, *Facial emotion recognition: State of the art performance on fer2013*, 2021. arXiv: 2105.03588 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2105.03588.

[19] S. Mishra, S. Kumar, S. Gautam, and J. Kour, "Real time expression detection of multiple faces using deep learning," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, 2021, pp. 537–542.

[20] N. Rathour, Z. Khanam, A. Gehlot, *et al.*, "Real-time facial emotion recognition framework for employees of organizations using raspberry-pi," *Applied Sciences*, vol. 11, no. 22, p. 10 540, 2021.

[21] S. Bangar, "Vgg-net architecture explained," *Medium*, Jun. 2022. [Online]. Available: https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f.

[22] J. Kaur, J. Saxena, J. Shah, Fahad, and S. P. Yadav, "Facial emotion recognition," in *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 2022, pp. 528–533. DOI: 10.1109/CISES54857.2022.9844366.

[23] A. V. Savchenko, "Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices," *arXiv preprint arXiv:2203.13436*, 2022.

[24] U. Dudekula and N. Purnachand, "Analysis of facial emotion recognition rate for real-time application using nvidia jetson nano in deep learning models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 598–605, 2023.

[25] J. Gupta, *Classical convolutional neural networks[cnn]*, https://connectjaya.com/classical-convolutional-neural-networkscnn, Mar. 2023.

[26] D. Qu, S. Dhakal, and D. Carrillo, *Facial emotion recognition using cnn in pytorch*, 2023. arXiv: 2312.10818 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2312.10818.

[27] Dshahid, *Convolutional neural network*, https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529, Oct. 2024.

[28] GeeksforGeeks, *Convolutional neural network (cnn) in machine learning*, https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/, Mar. 2024.

[29] GeeksforGeeks, *Vggnet architecture explained*, https://www.geeksforgeeks.org/vgg-net-architecture-explained/, GeeksforGeeks, Jun. 2024.

[30] V. Pradeep, B. Sumukha, G. R. Richards, S. P. Prashant, *et al.*, "Facial emotion detection using cnn and opencv," in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, IEEE, 2024, pp. 1–6.

[31] R. Singh, *Understanding and implementing mobilenetv3*, https://medium.com/@RobuRishabh/understanding-and-implementing-mobilenetv3-422bd0bdfb5a, Nov. 2024.

[32] Ultralytics, *Sam (segment anything model)*, https://docs.ultralytics.com/models/sam/, Dec. 2024.

[33] B. Zhang, K. Cheng, and X. Ni, "Research on emotion recognition method based on deep learning," in *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, 2024, pp. 475–478. DOI: 10.1109/CISAT62382.2024.10695408.

[34] MathWorks, *Vgg19*, https://www.mathworks.com/help/deeplearning/ref/vgg19.html, n.d.-b.

[35] *Papers with code - mobilenetv3 explained*, https://paperswithcode.com/method/mobilenetv3, n.d.

[36] viso.ai, *Very deep convolutional networks (vgg) essential guide*, https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/, n.d.