

DEVELOPMENT OF FRAMEWORK FOR FACIAL EMOTION RECOGNITION

*A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of*

**DOCTOR OF PHILOSOPHY
in
INFORMATION TECHNOLOGY**

By

**NIDHI
(2K21/PHDIT/04)**

Under the Supervision of

DR. BINDU VERMA



**Department of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)**

Shahbad Daulatpur, Main Bawana Road, Delhi-110042. INDIA

NOVEMBER-2024

Acknowledgements

I would like to express my heartfelt gratitude to all the individuals who have supported me throughout my PhD journey. Their encouragement, advice, and assistance have played a pivotal role in the successful completion of this work. First of all, I would like to express my heartfelt gratitude to my supervisor, **Dr. Bindu Verma**, for her unwavering guidance, support, and encouragement throughout the course of my research. Her expertise and unwavering belief in my abilities have been a constant source of motivation. I am truly fortunate to have had the privilege of working under her supervision.


I would also like to extend my sincere thanks to **Prof. Dinesh Kumar Vishwakarma** (Head of the Department (IT) & DRC Chairman) for his valuable guidance and support during my academic journey. I am also grateful to my SRC committee members: **Dr. Priyanka Meel**, **Dr. Chaavi Dhiman** for their thoughtful reviews, suggestions, and valuable time. Their expertise and recommendations have significantly contributed to enhancing the quality of my research. A special thanks to **Prof. Kapil Sharma** for providing the necessary resources and guidance for my research. His support has been crucial in facilitating the successful completion of my work. My heartfelt thanks extends to the esteemed faculty of the IT Department at DTU, including **Prof. Seba Susan**, **Dr. Virender Ranga**, **Dr. Ritu Agarwal**, **Dr. Anamika Chauhan**, **Dr. Rahul Gupta**, **Dr. Varsha Sisaudia** and **Ms. Geetanjali Bhola**.

I am deeply thankful to my family, whose love, sacrifices, and constant support have been the foundation of my success. To my father, **Mr. Dinesh Singh**, his unwavering belief in me and constant motivation have always pushed me to do my best. To my mother, **Mrs. Indra**, her selflessness, care, and boundless encouragement have been my greatest strength. I am equally grateful to my husband, **Mr. Bhupendra Singh**, for his patience, understanding, and unwavering support. His love and encouragement have been my anchor through this entire journey.

I would also like to extend my heartfelt thanks to my in-laws, **Mr. Kanhaiya Lal Singh** and **Mrs. Malti Singh**, for their warmth, support, and kindness. Their belief in me has been an immense source of strength. I am also grateful to my siblings, **Mrs. Janvi Bhadauria**, **Mr. Aditya Kanwar**, **Mr. Pushpendra Singh**, and **Mrs. Mansi Singh**, for always being there for me with their love, care, and moral support. Their presence in my life has been a continuous source of motivation and joy.

Lastly, I would like to express my gratitude to my dear friends, **Reena Tripathi** and **Lakshita**, for their unwavering friendship and encouragement. Their support, both academically and emotionally, has been invaluable to me, and I am truly fortunate to have them by my side throughout this journey.

To all those who have contributed to my growth and success, whether directly or indirectly, I express my deepest thanks. Your kindness and support have made this accomplishment possible, and I will always be grateful.


(Nidhi)



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultapur, Bawana Road-Delhi-42

CANDIDATE'S DECLARATION

I **Nidhi** hereby declare that the work which is being presented in the thesis entitled “**Development of Framework for Facial Emotion Recognition**” in partial fulfillment of the requirements for the award of the Degree of **Doctor of Philosophy**, submitted in the **Department of Information Technology, Delhi Technological University** is an authentic record of my own work carried out during the period from 02/08/2021 to 2/12/2024 under the supervision of **Dr. Bindu Verma**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Nidhi
Kaniz
17/12/24

Nidhi

(Ph.D. Student)

**Department of Information Technology,
Delhi Technological University, New Delhi-110042 India**



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultapur, Bawana Road-Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **Nidhi** (2K21/PHDIT/04) has carried out her research work presented in this thesis entitled “**Development of Framework for Facial Emotion Recognition**” for the award of the Degree of **Doctor of Philosophy** from **Department of Information Technology, Delhi Technological University, Delhi**, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

(Supervisor)

Bindu Verma
77/12/24

Dr. Bindu Verma

(Assistant Professor)

Department of Information Technology

Delhi Technological University, Delhi-110042

Date: 17/12/2024

Dedicated to

*My Parents: for their love, endless support
and encouragement*

*My Husband: who has been a considerate, supportive
and cooperative throughout the entire journey...*

Abstract

Human ideas and sentiments are mirrored in facial expressions. Facial emotion recognition (FER) is a crucial type of visual data that can be utilized to deduce a person’s emotional state. It gives the spectator a plethora of social cues, such as the viewer’s focus of attention, emotion, motivation, and intention. It is considered a powerful instrument for silent communication. AI-based facial recognition systems can be deployed at different areas like bus stations, railway stations, airports, or stadiums to help security forces identify potential threats. In this thesis, different aspects of facial emotion recognition are explored and addressed. A facial emotion recognition (FER) has significantly advanced over the past few decades, but issues related to occlusion robustness and pose invariance, are relatively less encountered, specifically in uncontrolled environments. Therefore, a novel framework is presented in this thesis that can capture geometric facial features from occluded and pose-variant images and classify the facial emotions accurately. Facial emotion recognition has garnered significant interest in recent years due to its wide-ranging applications in human-computer interaction (HCI), affective computing, and psychological research. In this thesis, a capsule neural network-based model is introduced which incorporates an attention mechanism to improve the classification performance. More specifically, an empirical study is performed on the effectiveness of various attention mechanisms in the context of FER. Considering this aspect, four attention techniques are explored and compared: channel attention, spatial attention, CBAM attention, self-attention, and multi-head attention (MHA). The experimental results show the distinct impacts of each attention mechanism on improving the recognition performance of different facial expressions.

Along with the basic emotions addressed above, there are complex expressions which are made up of two basic emotions like “Happily Disgusted”, “Happily Surprised”, “Sadly Surprised”, etc. Compound emotion provides a richer understanding of the human emotional state by capturing the nuanced combinations of the basic seven emotions. To recognize compound emotion, this thesis contributes by present-

ing a lightweight Swin Transformer model (LSwin-CBAM) for compound emotion recognition. Facial expression recognition systems can have advanced applications like public safety, information management and retrieval, social media, psychological studies, and patient care. Addressing its application in driver safety by recognizing the driver's emotions can enhance driving comfort, safety, and the adoption of intelligent vehicles. This thesis presents a Vision Transformer-based framework capable of recognizing driver emotions across various resolutions. It offers an improved solution for emotion recognition, even with small datasets. In this thesis, comprehensive experiments have been carried out on several FER datasets, including RAF-DB, AffectNet, EmotioNet, CK+, JAFFE, D3S, and KMU-FED to show the effectiveness of the proposed frameworks.

List of Publications

Journals

- **Nidhi**, and Bindu Verma. “From methods to datasets: a detailed study on facial emotion recognition.” *Applied Intelligence* 53, no. 24 (2023): 30219-30249. (SCIE Indexed, IF: 3.4) DOI: 10.1007/s10489-023-05052-y (Published)
- **Nidhi**, and Bindu Verma. “A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition.” *Applied Intelligence* (2024): 1-17. (SCIE Indexed, IF: 3.4) DOI: 10.1007/s10489-024-05598-5 (Published)
- **Nidhi**, and Bindu Verma. “ViT-SLS: Vision Transformer with Stochastic Depth for Efficient Driver’s Emotion Recognition System” is communicated in *IEEE Transactions on Human-Machine Systems* (SCIE Indexed, IF: 3.6) (Communicated)
- **Nidhi**, and Bindu Verma. “In-the-Wild Facial Emotion Recognition using Relation-aware Geometric Features and CapsNet” is communicated in *Computers and Electrical Engineering* (SCIE Indexed, IF: 4.0) (Communicated)

Conferences

- **Nidhi**, Bindu Verma. “A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition”. In *2023 Seventh International Conference on Image Information Processing (ICIIP)* (pp. 591-596) (2023, November). IEEE. (Published)
- **Nidhi**, Bindu Verma. “Empirical Insights: Unraveling the Impact of Various Attention Mechanisms on Facial Emotion Recognition” presented in *5th International Conference on Data Analytics & Management (ICDAM-2024)*, Springer. (Accepted & Presented)

Contents

Acknowledgements	ii
Declaration	iv
Certificate	v
Dedication	vi
Abstract	vii
List of Publications	ix
List of Tables	xv
List of Figures	xix
List of Abbreviations	xx
1 Introduction	1
1.1 Applications of Facial Emotion Recognition	3
1.2 Research gaps and challenges in Facial Emotion Recognition	5
1.3 Problem Definition	5
1.3.1 Research Objectives	6
1.4 Contributions in the Thesis	6
1.5 Outlines of the Thesis	10

2	Literature Survey	13
2.1	Facial Emotion Recognition using Traditional Methods	14
2.2	Facial Emotion Recognition using Deep Learning Methods	15
2.2.1	Basic Emotions Recognition	16
2.2.2	Facial Emotion Recognition In-the-wild	18
2.2.3	Compound Emotion Recognition	19
2.2.4	Driver’s Emotion Recognition	20
2.3	In Thesis Prospective:	21
3	Investigation of Different Mechanisms on Facial Emotion Recognition	23
3.1	Introduction	23
3.2	Literature Survey	25
3.3	Proposed work	26
3.3.1	Capsule Neural Network	27
3.3.2	Different Attention Mechanisms	29
3.4	Experimental Results	33
3.4.1	Experimental results on RAF-DB dataset	33
3.4.2	Analysis of different Attention Mechanisms	35
3.5	Conclusion	36
4	Design and development of a framework for a Facial Emotion Recognition system in-the-wild	38
4.1	Introduction	38
4.2	Literature Survey	40
4.3	Proposed Work	42
4.3.1	Feature Extraction	43
4.3.2	Capsule Neural Network	47
4.4	Experimental Results	50
4.4.1	Experimental results on RAF-DB dataset	53
4.4.2	Experimental results on AffectNet dataset	53

4.4.3	Ablation Study	55
4.4.4	Comparative study of FMR-CapsNet with state-of-the-art methods	57
4.5	Conclusion	59
5	Design and development of robust and generic framework for Com- pound Emotion Recognition	60
5.1	Introduction	60
5.2	Literature Survey	63
5.3	Proposed Work	64
5.3.1	Input Data and Augmentation	66
5.3.2	CBAM Attention mechanism	67
5.3.3	Swin Transformer	69
5.3.4	Compound Emotion Classification	71
5.4	Experimental Results	72
5.4.1	Datasets	75
5.4.2	Class Weights	76
5.4.3	Evaluation Metrics	76
5.4.4	Experimental results on RAF-DB(Compound) dataset	77
5.4.5	Experimental results on EmotioNet dataset	78
5.4.6	Ablation Study	81
5.4.7	Comparative study of LSwIn-CBAM with state-of-the-art	83
5.5	Conclusion	86
6	Design and development of a deep learning framework for driver’s emotion recognition	87
6.1	Introduction	87
6.2	Literature Survey	89
6.3	Proposed Method	91
6.3.1	Formulation of General ViT	93
6.3.2	Shifted Patch Tokenization	95

6.3.3	Locality Self-Attention Method	96
6.3.4	Stochastic Depth	97
6.4	Experimental Results	98
6.4.1	Datasets	100
6.4.2	Experimental results on D3S dataset	103
6.4.3	Experimental results on KMU-FED dataset	103
6.4.4	Experimental results on CK+ and JAFFE datasets	104
6.4.5	Ablation Study	105
6.4.6	Comparative study of ViT-SLS with state-of-the-art	107
6.5	Conclusion	108
7	Conclusion & Future Directions	109
7.1	Summary and Contribution of the Thesis	109
7.2	Future Directions	112
	Bibliography	114
	Appendix A Capsule Neural Network	134
A.1	Capsule Neural Network	134
	Appendix B Vision Transformer	137
B.1	Vision Transformer	137
	Appendix C Swin Transformer	140
C.1	Swin Transformer	140
	List of Publication and their Proofs	143
	Plagiarism Report	150
	Curriculum Vitae	153

List of Tables

2.1	Comparative study of different algorithms proposed in the literature.	16
3.1	Computational analysis of different attention mechanisms *I= Number of input channels, C= number of channels, H= height of the feature map, and W= width of the feature map.	34
3.2	Comparative analysis of different attention mechanisms on RAF-DB dataset.	34
4.1	Predicted blendshapes from LandMarker solution.	44
4.2	Summary of hyperparameters used in FMR-CapsNet.	52
4.3	Distribution of samples in training and testing subset of RAF-DB and AffectNet dataset.	52
4.4	Class weights assigned to each class of RAF-DB and AffectNet dataset.	54
4.5	Performance comparison of different SOTA methods with FMR-CapsNet on RAF-DB and AffectNet datasets.	56
4.6	Detailed comparison of FMR-CapsNet with state-of-the-art methods with respect to class-wise accuracy obtained on RAF-DB and AffectNet datasets; * Highest class-wise accuracies achieved for RAF-DB and AffectNet datasets are highlighted in red and blue color respectively. .	57
4.7	Evaluation of the effect of different modules used in FMR-CapsNet on RAF-DB and AffectNet datasets.	57
5.1	Summary of hyperparameters used in LSwin-CBAM.	74

5.2	Experimental results of LSwin-CBAM on RAF-DB dataset, * indicates results of model trained with class weights.	78
5.3	Experimental results of LSwin-CBAM on EmotioNet dataset, * indicates results of model trained with class weights.	80
5.4	Significance of adding each module to LSwin-CBAM while experimenting on RAF-DB dataset.	82
5.5	Significance of adding each module to LSwin-CBAM while experimenting on EmotioNet dataset.	82
5.6	Parametric and architectural influence of Swin Transformer stages on the performance of model on RAF-DB dataset.	82
5.7	Parametric and architectural influence of Swin Transformer stages on the performance of model on Emotionet dataset.	83
5.8	Comparative analysis of state-of-the-art methods and proposed method on RAF-DB dataset.	84
5.9	Comparative analysis of state-of-the-art methods and proposed method on EmotioNet dataset.	85
6.1	Comparative analysis of state-of-the-art methods and proposed method on KMU-FED dataset.	106
6.2	Comparative analysis of state-of-the-art methods and proposed method on D3S dataset.	106
6.3	Comparative analysis of state-of-the-art methods and proposed method on CK+ dataset.	107
6.4	Comparative analysis of state-of-the-art methods and proposed method on JAFFE dataset.	107

List of Figures

2-1	Frequency of models used in the literature.	17
3-1	Attentive Capsule Neural Network	28
3-2	Validation loss of CapsNet with different attention mechanisms.	36
3-3	ROC_AUC Curves of applied attention mechanisms with CapsNet *(a) CapsNet (b) CapsNet-C, (c) CapsNet-S, (d) CapsNet-CBAM, (e) CapsNet-Self, (f) CapsNet-MHA.	36
4-1	The proposed architecture of FMR-CapsNet incorporating Facemesh mediapipe for geometric feature extraction, ResNet50 for refinement of features, and CapsNet for emotion classification for facial emotion recognition in-the-wild.	42
4-2	478 Facial landmarks extracted using FaceMesh Mediapipe.	44
4-3	Facial blendshapes scores extracted using FaceMesh Mediapipe.	45
4-4	Visualization of blendshape weights of 100 random images of RAF-DB dataset.	45
4-5	Heatmap of distance matrix generated using euclidean distance on (a) RAF-DB (b) AffectNet datasets.	47
4-6	Architecture of ResNet50 pretrained model	48
4-7	Samples of facial expressions from AffectNet and RAF-DB dataset	51
4-8	Confusion metric evaluated on (a) RAF-DB (b) AffectNet Dataset using FMR-CapsNet.	54
4-9	ROC curve evaluated on (a) RAF-DB (b) AffectNet Dataset using FMR-CapsNet.	55

4-10	ROC_AUC score of different classes of AffectNet and RAF-DB dataset.	55
4-11	(a) Influence of number of dynamic routings in FMR-CapsNet on AffectNet and RAF-DB dataset (b) Variation in number of trainable parameters on adding different modules to FMR-CapsNet *CapsNet: Capsule neural network, Face-Caps: FaceMesh + CapsNet, Res-CapsNet: ResNet50 + CapsNet, FMR-CapsNet: FaceMesh + ResNet50 + CapsNet.	58
5-1	Proposed model architecture integrating CutMix augmentation, CBAM attention, and lightweight Swin Transformer module for Compound Emotion Recognition.	66
5-2	Samples of Images Post-CutMix Augmentation.	68
5-3	Architecture of Convolutional Block Attention Module (CBAM).	68
5-4	Visualization of learned attention weights of CBAM attention on input images.	69
5-5	Streamlined Swin Transformer Architecture used in the proposed model.	70
5-6	Number of samples in a) RAF-DB dataset b) EmotioNet Dataset.	75
5-7	Evaluated results on RAF-DB dataset a) ROC_Curve b) Confusion metric.	79
5-8	Evaluated results on RAF-DB dataset (with class weights) a) ROC_Curve b) Confusion metric.	79
5-9	Evaluated results on Emotionet dataset a) ROC_Curve b) Confusion metric.	80
5-10	Evaluated results on Emotionet dataset (with class weights) a) ROC_Curve b) Confusion metric.	81
5-11	Class-wise average accuracy obtained on RAF-DB without and with class weights.	83
5-12	Class-wise average accuracy obtained on EmotioNet without and with class weights.	83

5-13	Performance Evaluation: LSwin-CBAM vs. Transformer Variants on RAF-DB dataset.	85
6-1	Proposed architecture of ViT-SLS (Vision Transformer integrated with Shifted patch tokenization and modified transformer encoder (including locality self-attention and stochastic depth)) for driver’s emotion recognition.	92
6-2	Modified transformer encoder module with locality self-attention and stochastic depth layer *Locality self-attention is incorporated in “MHSA with diagonal attention mask layer”.	93
6-3	Model overview of general Vision Transformer [140].	94
6-4	Distribution of classes for (a) D3S, (b) KMU-FED, (c) CK+, and (d) JAFFE datasets.	101
6-5	Performance of the general Vision Transformer on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.	102
6-6	Performance of ViT-SL on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.	102
6-7	Performance of the proposed model (ViT-SLS) on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.	102
6-8	Performance of the General Vision Transformer on KMU-FED dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.	104
6-9	Performance of the proposed model (ViT-SLS) on KMU-FED dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.	104
6-10	Class-wise accuracy of ViT-SLS on (a) D3S, (b) KMU-FED, (c) CK+, and (d) JAFFE.	105
6-11	Ablation study on D3S and KMU-FED datasets.	106
A-1	Capsule neural network	135
B-1	Model overview of general Vision Transformer	138

C-1 Swin Transformer architecture 141

List of Abbreviations

1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
ACNN	Attentional Convolutional Network
Adam	Adaptive Moment Estimation
AdamW	Adaptive Moment Estimation with Weight Decay
ADAS	Advanced Driver Assistance Systems
AI	Artificial Intelligence
AMP-Net	Adaptive Multilayer Perceptual Attention Network
ARM	Amending Representation Module
AU	Action Units
AU-D	Action Unit Detection
Bi-LSTM	Bidirectional Long Short Term Memory
CapsNet	Capsule Neural Network
CapsNet-C	Capsule Neural Network with Channel Attention
CapsNet-CBAM	Capsule Neural Network with CBAM Attention
CapsNet-MHA	Capsule Neural Network with Multi Head Attention
CapsNet-S	Capsule Neural Network with Spatial Attention
CBAM	Convolutional Block Attention Module
CBLNN	Convolution Bidirectional Long Short-term Memory Neural Network
CCE	Categorical Cross-Entropy
CE	Cross-Entropy
CER	Compound Expression Recognition
C-EXPR-DB -	Compound-Expression-Network
CFC	Coarse-to-fine Cascaded Network
CF-DAN	Cross-Fusion Dual-attention Network
CK+	Extended Cohn Kanade

CNN	Convolutional Neural Network
Conv2D	2D Convolution
ConvLSTM	Convolutional Long- Short Term Memory Networks
D3S	Driver Drowsiness Dataset
DAKL	Deep Attentive Center Loss,
DCNN	Deep Convolutional Neural Network
DDRGCN	Double Dynamic Graph Convolutional Network
DL	Deep Learning
DLP-CNN	Deep Locality-Preserving CNN
ECAM	Enhanced Capsule Attention Module
EmNet	Emotion Network
ESLM	Expression Soft Label Mining
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
FER+	Face Expression Recognition Plus
FG2017	Face and Gesture Recognition 2017
FG-Emotions	Fine Grain Emotions
FMR-CapsNet	Facemesh Mediapipe features with ResNet50 and Capsule Neural Network
FPR	False Positive Rate
FRR- CNN	Feature Redundancy-reduced Convolutional Neural Network
GACN	Geometry-Aware Conditional Network
GAF2	Group Affective Database 2.0
GAF3	Group Affective Database 3.0
GELU	Gaussian Error Linear Unit
GER	Generalized Emotion Recognition
GFE2N	Gaussian-based Facial Expression Feature Extraction Network
GSDNet	Gradual Self Distillation Network with Adaptive Channel Attention
HPFS	High-Purity Feature Separation
ICL	Intra-dataset Continual Learning

iCV-MEFED	iCV Multi-Emotion Facial Expression Dataset
IE-DBN	Identity-expression Dual Branch Network
ISLM	Iterated Soft Label Mining
JAFFE	The Japanese Female Facial Expression
KDEF	Karolinska Directed Emotional Faces
KMU-FED	Keimyung University Facial Expression of Drivers
KNN	K Nearest Neighbor
LD	Linear Discriminant
LFW	Labeled Faces in the wild
LSA	Locality Self-attention
LSGB	Local Relation Module, an Self- and Global-attention Module, and Batch Normalization
LSR	Label Smoothing Regularization
LSwin-CBAM	Lightweight Swin Transformer with CBAM Attention Mechanism
MAD	Multi-head Attention Dropping
MAE	Masked Autoencoder
MHA	Multi-head Attention
MLP	Multi-layer Perceptron
MSAD	Multi-head Self-attention Dropping
MSAU-Net	Two-stream Multi-scale AU-based Network
MTCNN	Multi-task Cascaded Convolutional Networks
MTL	Multi-task Learning
NCA	Neighborhood component
NCA	Neighborhood Component Analysis
NCRB	National Crime Research Bureau
NIR	Near Infrared
NLP	Natural Language Processing
OBU	Onboard unit
PSR	Pyramid with Super-resolution
QA-CNN	Quaternion CNN with attention mechanism

QA-CNN	Quaternion CNN with attention mechanism
RAF-D	Radboud Faces Database
RAF-DB	Real-world Affective Faces Database
RAM	Random Access Memory
RAN	Residual Attention Network
R-CNN	Regions with Convolutional Neural Network
ReCNN	Relation Convolutional Neural Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
RGB	Red, Green, and Blue
RNN	Recurrent Neural Network
ROC	Receiver operating characteristic curve
ROC AUC	Area under the Receiver Operating Characteristics Curve
RUL	Relative Uncertainty Learning
SAM	Segment Anything Model
SCN	Self-Cure Network
SD	Stochastic Depth
SFEW	Static Facial Expressions in the Wild
SJMT	Selective Joint Multitask Approach
SOTA	State-of-the-art
SP	Smooth Predicting
SPT	Shifted Patch Tokenization
SSA	Spectral and Spatial Attention
ST	Swin Transformer
SVM	Support Vector Machine
SW-MSA	Shifted Windows-based Multi-head Self-attention
TFE	Tandem Facial Expression
TPR	True Positive Rate
VGG	Visual Geometry Group
ViT	Vision Transformer

ViT-SL	Vision Transformer with Shifted Patch Tokenization and Locality Self Attention
ViT-SLS	Vision Transformer with Shifted Patch Tokenization, Locality Self Attention, and Stochastic Depth
WHO	World Health Organization

Chapter 1

Introduction

Emotion plays a vital role in human behavior analysis. Genuine emotions are often displayed unconsciously through facial expressions, revealing what a person is feeling or thinking. Facial expressions are unlearned and innate to human nature, therefore, important for non-verbal, interpersonal communication as well as intention recognition. Body movements, facial expressions, and postures can be used to predetermine rich information about a human's status, awareness, intention, and emotional state [1]. Among these, facial expressions are an outward display of a person's emotions. There are seven basic emotions namely, sadness, happiness, surprise, anger, fear, disgust, and contempt. Automated facial expression recognition plays an important role in recognizing a person's emotions. As humans can read the facial expressions of others, how computers can be trained to recognize emotions by observing the facial expressions. This question motivates us to make a contribution to facial emotion recognition. Socially intelligent machines that can sense, adapt, and respond to non-verbal behavioral cues can improve the quality of human life by understanding and appropriately responding to the user's needs. Therefore, there is a growing need to develop emotionally aware machines that can recognize human emotion. These systems find applications in areas such as surveillance, robotics, human-computer interaction, advanced driver assistance systems (ADAS), psychological studies, etc. Camera-based emotion recognition systems are better suited for monitoring a person's emotions since it is non-intrusive and does not require the active participation

of the person. The main challenges of a camera-based emotion recognition system are that the person's face should be correctly detected in spite of varied illumination conditions in naturalistic settings, occlusions and other people in the scene. Facial movements including head movements and changes in emotions across time make it difficult to recognize emotions in real time. Moreover, the emotion recognition system should be generic enough to accommodate the expressions of different people. Below we have discussed the types of emotions that we have targetted in this thesis.

Basic Emotions

According to Ekman [2], there are seven basic expressions: Joy, Anger, Sad, Surprise, Disgust, Fear, Contempt. Basic emotions can be captured posed or spontaneous. Posed expressions are captured in a controlled environment where subjects are asked to pose the expression and can disguise their inner feelings. Spontaneous expressions are captured when subjects produce expressions naturally and freely in the lab environment, these expressions can expose one's real feelings. Both posed and spontaneous emotions are recognized using the proposed model in the Chapter 6, where JAFFE dataset is used for posed and CK+ dataset for spontaneous.

In-the-wild Facial Emotions

In-the-wild refers to the emotions captured in the natural scenario where complexity arises from various factors such as variations in lighting conditions, head positions, illumination, and occlusion. These challenges are usually found in unconstrained environments like images collected from the internet where background noise is complimentary included to pose the challenge for facial emotion recognition systems. In-the-wild emotion recognition accuracy is very low in the literature. so, a primary solution is to propose a novel robust facial emotion recognition for in-the-wild images. A novel robust facial emotion recognition system is proposed which extracts the relation-aware geometric features of the facial images which improves the classification accuracy of in-the-wild images. Geometric features help model to better understand the emotion of occluded and pose variant images as well.

Compound Emotions

Other than the basic seven emotions, there are compound emotions which are the combination of basic emotions to form complex ones like “Happily Surprised”, “Happily Disgusted”, “Sadly Fearful”, “Sadly Angry”, “Sadly Surprised”, etc. Authors usually focus on seven basic emotions, but there is a need to examine more detailed and precise facial expressions. To examine a person’s emotional state in greater detail using facial emotion expression analysis, compound emotion categories have been proposed [3]. Compound emotions offer a deeper and more detailed insight into human feelings, reflecting the complexity of real-life emotional experiences that involve blends of basic emotions. So, there is a necessity for contribution in the context of compound emotion recognition which comprehends and identifies fine-grained facial expressions.

1.1 Applications of Facial Emotion Recognition

Facial expression recognition systems can have advanced applications like public safety, information management and retrieval, social media, psychological studies, and patient care. Such systems can be installed at different public areas like bus stations, railway stations, airports, or stadiums to help security forces identify potential threats. Some of the applications are detailed below:

- *Public Safety*

Facial expression recognition can be used for public safety by detecting lies, can be incorporated into smart border control to detect irregular activity, can be deployed in public areas like bus stands and railways station to detect possible terrorism threats, and can be used to prevent accidents by detecting driver’s behavior like exhaustion, drowsiness, and drunkenness.

- *Social Media*

People use social media to convey their feelings for various reasons, including expressing their identity, exchanging support, and seeking community. Academics and technologists have recently employed machine learning and facial

expression recognition to determine emotions and mental health status based on less apparent signals (for example, visual markers of depression or stress on Instagram) [4] [5]. Building agents that provide emotional support to users [6] or assisting individuals on the autistic spectrum with communication [7] [8] are two examples of possible valuable applications based on emotion recognition.

- *Health Care*

Human's expressions changes with their different state of health and mind. E-health is helping healthcare centers to use "Information and Communication Technology" in health surveillance, healthcare services, knowledge and research, health literature, and health education. FER can help doctors to counsel and examine the patient's current health state to determine the patient's comfort and feeling about the given medication and treatment.

- *Information Management and Retrieval*

Sometimes in daily life, non-verbal communication is required where people communicate via expression, hand gestures, eye blinks, etc. Relevant information can be retrieved from facial expressions to identify the unusual activity, alerting the respective concerned group.

- *Educational Institutes*

Facial expression recognition can be used in educational institutes to analyze the teaching states of the students and improve the teaching quality accordingly. Along with that, it can be used to read the expressions of students to analyze their interests, understanding, and student's state of mind. This can help teachers modify the teaching style according to the student's feedback and eventually increase their learning.

- *Customer Behavior Analysis*

It is difficult to measure customer satisfaction with repetitive surveys and review forms. So companies are more interested in measuring satisfaction levels

through a deep neural network-based expression recognizer instead of having a one-to-one interview or questionnaire session.

1.2 Research gaps and challenges in Facial Emotion Recognition

After extensively reviewing the work published in the field of facial emotion recognition, we feel that there are still a few research gaps exist, some of which are addressed in our research objectives.

- One of the major challenges of the FER system is recognizing emotions in-the-wild.
- There is very less work done on compound emotion recognition and as per the literature, the proposed methods have not achieved impressive accuracy.
- Natural occlusion and non-frontal face images can affect the visual representation of actual facial emotion, posing a hurdle in deploying a FER system in the driver's emotion recognition field.
- Insufficient and imbalanced datasets are major barriers when applying deep learning to FER tasks.

1.3 Problem Definition

The facial emotion recognition system is a very effective for having notable human-computer interaction. Face recognition in a real environment is a challenging task as head orientation, pose variant, occlusion, and illumination conditions can significantly affect the value of facial action unit (AU) coefficients (Facial Action Coding System (FACS) classifies distinct muscle movements as action units which are related to one or more facial muscle movements). Thus, a pose-invariant face expression recognition system is proposed to overcome this challenge, focusing on accurately

recognizing facial emotions in unconstrained environments, including occluded and pose-variant faces. Attention mechanisms have gained prominence in enhancing FER systems by selectively focusing on relevant facial regions. Therefore, we also investigated the impact of different attention mechanisms on FER. Compound emotions, are characterized by combining two or more emotions with one being dominant and the other complementary. This composition makes the recognition difficult. So, there is comparatively less work done on compound emotion recognition, and comparatively less accuracy achieved in state-of-the-art methods. So, we worked on developing a lightweight and generic framework for compound emotion recognition. The lightweight model enhances its applicability in real-time due to fast performance, less storage, and can work with few training samples as well. As a conclusive part of this thesis, we worked on one of the applications of FER by developing a driver’s emotion recognition system. This system leverages FER model to monitor the emotional state of the driver, aiming to enhance road safety and driving experiences by detecting emotions such as “eye-close”, “happy”, “yawn”, etc.

1.3.1 Research Objectives

OBJECTIVE 1: To study and investigate different attention mechanisms and to design a framework for a facial emotion recognition system in-the-wild.

OBJECTIVE 2: To design a robust and generic framework for compound emotion recognition.

OBJECTIVE 3: To design a deep learning framework for driver’s emotion recognition.

1.4 Contributions in the Thesis

In this thesis, we focus on facial emotion recognition and briefly address the few research gaps in the field of facial emotion recognition. In this thesis three research

objective were defined as mention in Section 1.3.1 and we address all these objectives one by one discussed below as contribution of the thesis:

- (I) **OBJECTIVE 1:** *To study and investigate different attention mechanisms and to design a framework for a facial emotion recognition system in-the-wild.*

In this objective we have worked on two models. Firstly we presented a comprehensive study of attention mechanisms on facial emotion recognition. Various types of attention mechanisms have been explored including channel attention, spatial attention, CBAM (Convolutional Block Attention Module), self-attention, and multi-head attention. In the second model, we developed a capsule network-based framework for facial emotion recognition in-the-wild.

Model 1: Attention Mechanisms for Facial Emotion Recognition:

Facial emotion recognition has garnered significant interest in recent years due to its wide-ranging applications in human-computer interaction, affective computing, and psychological research. A capsule neural network-based model is proposed which incorporates an attention mechanism to improve the classification performance. Specifically, an empirical investigation into the effectiveness of various attention mechanisms in enhancing facial emotion recognition systems is presented. In recent years, attention mechanisms have gained prominence in enhancing the performance of FER systems by selectively focusing on relevant facial regions. While looking in that direction, four attention mechanisms are explored with their comparison: channel attention, spatial attention, CBAM attention, self-attention, and multi-head attention (MHA) attention. An experiment has been conducted using capsule network (CapsNet) architecture on an in-the-wild facial emotion recognition dataset. The results demonstrate the distinct impacts of each attention mechanism on improving the recognition performance for different facial expressions. Furthermore, the computational complexity and interpretability of these attention mechanisms are analyzed to provide insights into their practical feasibility.

This research contributes to the understanding of attention mechanisms in the context of facial emotion recognition. It offers valuable guidance for the design and optimization of emotion recognition systems in real-world applications.

Model 2: Facial Emotion Recognition In-the-wild: Although automatic FER has significantly advanced over the past few decades, issues encountered in in-the-wild images such as occlusion robustness and pose variance are relatively less encountered, specifically in uncontrolled environments. Therefore, this model presents a robust FER method designed to alleviate the challenges of in-the-wild emotions posed by pose variations and occlusions. This model improves the classification accuracy of in-the-wild emotions by exploiting geometric features. The proposed method FMR-CapsNet employs the FaceMesh model for geometric feature extraction, utilizing facial blendshape scores that adeptly capture features from pose-variant and occluded images. The Euclidean distance metric is used to construct a distance matrix for relation-aware blendshapes, providing relative information between blendshapes scores. To further refine these extracted features, transfer learning is applied with a pretrained ResNet50 on the evaluated distance matrix. Additionally, a capsule neural network is employed to capture both directional and spatial information, improving the accuracy of inter-class feature differentiation. The proposed method is evaluated on two very popular in-the-wild datasets, namely RAF-DB and AffectNet. Experimental results show that the FMR-CapsNet greatly improves the performance of FER for in-the-wild images (including occluded and pose-variant images) by achieving classification accuracy 97.01% on RAF-DB and 71.12% on AffectNet dataset.

- (II) OBJECTIVE 2: *To design a robust and generic framework for compound emotion recognition*

Compound emotion recognition is challenging due to very less publicly available compound emotion datasets which are imbalanced too. In this objective,

we have proposed an LSwin-CBAM for the classification of compound emotions. To address the problem of the imbalanced dataset, the proposed model exploits the CutMix augmentation technique for data augmentation. It also incorporates the CBAM attention mechanism to emphasize the relevant features in an image and Swin Transformer with fewer Swin Transformer blocks which leads to less computational complexity in terms of trainable parameters and improves the overall classification accuracy as well. The experimental results of LSwin-CBAM on RAF-DB and EmotioNet datasets show that the proposed transformer-based network can well recognize compound emotions by achieving an accuracy of 51.81% on RAF-DB and 39.67% on EmotioNet dataset.

(III) OBJECTIVE 3: *To design a deep learning framework for driver’s emotion recognition.*

There are various applications of facial emotion recognition such as information management and retrieval, public safety, health care, educational institutes, etc. According to the World Health Organization (WHO), road accidents are majorly caused by driver distractions. So, recognizing a driver’s emotions is vital to enhance driving comfort, safety, and the adoption of intelligent vehicles. In this thesis, we worked on the application of facial emotion recognition for driver’s safety while driving the vehicle. If a system can detect the driver’s emotions like yawn, or eye-close, then it can generate an alarm or alert to keep the driver aware which can eventually prevent possible road accidents.

In this thesis, we proposed an effective Vision Transformer-based framework termed as “ViT-SLS”, which employs a Vision Transformer with shifted patch tokenization where input images are shifted and broken into patches which are further propagated through the locality-self attention module (to enhance the performance of small-size datasets) and followed with stochastic depth for regularization. The proposed fused method performs invariably well for images with different resolutions and provides a better solution for recognition of driver’s emotions for small datasets. The proposed model was evaluated on

four datasets: D3S, KMU-FED, CK+, and JAFFE datasets, and achieved an accuracy of 99.7%, 99.9%, 99.6%, and 93.02% respectively which depicts the efficacy of the proposed model.

1.5 Outlines of the Thesis

The thesis comprises seven chapters and each chapter addresses the key aspects of facial emotion recognition.

1. Chapter 1 Introduction

This chapter gives an overview of the thesis, including the problem statement, motivation, and challenges present in facial emotion recognition. We also discuss our contributions and provide an outline of the thesis in this chapter.

2. Chapter 2 Literature Survey

In this chapter, we present an overview of the state-of-the-art methods in facial emotion recognition. First, we have discussed the survey on traditional methods and deep learning based emotion recognition. Further, we have discussed the facial emotion recognition in-the-wild followed by compound emotion recognition. We have also done literature review in the field of driver's emotion recognition. We have also done detailed analysis of the state-of-the-art methods and represented statistically in the form of pie-chart and table.

3. Chapter 3 Investigation of different attention mechanisms on Facial Emotion Recognition

This chapter presents an empirical study on the effectiveness of various attention mechanisms in facial emotion recognition (FER). Considering this aspect, four attention techniques are explored and compared: channel attention, spatial attention, CBAM attention, self-attention, and multi-head attention (MHA). This chapter provides an insight to relevance of each attention mechanism for facial emotion recognition.

4. **Chapter 4 Design and development of a framework for a facial emotion recognition system in-the-wild.**

This chapter introduces a novel framework for a facial emotion recognition system in-the-wild. A robust Facial Emotion Recognition (FER) method is designed to overcome the challenges posed by pose variations and occlusions. The proposed method FMR-CapsNet employs the FaceMesh model for geometric feature extraction, utilizing facial blendshape scores that adeptly capture features from side-facing and occluded images. Additionally, a capsule neural network is employed to capture both directional and spatial information, improving the accuracy of inter-class feature differentiation.

5. **Chapter 5 Design and development of a robust and generic framework for compound emotion recognition**

This chapter presents an LSwin-CBAM model for the classification of compound emotions. To address the problem of the imbalanced dataset, the proposed model exploits the CutMix augmentation technique for data augmentation. It also incorporates the CBAM attention mechanism to emphasize the relevant features in an image and Swin Transformer with fewer Swin Transformer blocks which leads to less computational complexity in terms of trainable parameters and improves the overall classification accuracy as well.

6. **Chapter 6 Design and development of a deep learning framework for driver's emotion recognition**

This chapter presents an effective Vision Transformer-based framework termed as 'ViT-SLS', which employs a Vision Transformer with Shifted Patch Tokenization where input images are shifted and broken into patches which are further propagated through the locality-self attention module (to enhance the performance of small-size datasets) and followed with stochastic depth for regularization.

7. **Chapter 7 Conclusion & Future Directions**

This chapter summarizes the key findings and main contributions of the thesis.

Along with that, future research directions are also discussed in this chapter.

8. **Appendix A Capsule Neural Network**

In this appendix, we explain the capsule neural network in detail.

9. **Appendix B Vision Transformer** In this appendix, we provide the detailing of Vision Transformer model along with its formulation and architecture.

10. **Appendix C Swin Transformer**

In this appendix, we present the Swin Transformer in detail along with its architecture and formulation.

Chapter 2

Literature Survey

Nidhi, and Bindu Verma. “From methods to datasets: a detailed study on facial emotion recognition.” *Applied Intelligence* 53, no. 24 (2023): 30219-30249. (SCIE Indexed, IF: 3.4) DOI: 10.1007/s10489-023-05052-y (Published)

In the previous chapter, facial emotion recognition, its applications, and challenges are introduced in detail. This chapter discusses a literature review on facial expression recognition, including work on basic, in-the-wild, and compound emotions. Humans can easily recognize emotions accurately but making a fully automated machine to recognize facial expressions is still a challenging task. If this capability of emotion recognition can be empowered in robots or computers then robotic applications can be developed to understand the emotions of a human with a limited understanding of the relationship between emotional expressions and body movements [9]. According to Mostafa et al. [10], emotion recognition system generally comprises three steps, first is face detection in image or video, second is feature extraction and last is the classification process.

2.1 Facial Emotion Recognition using Traditional Methods

Research done in social psychology explains that facial emotion helps in coordinating the conversation between speaker and listener that can perceive the interest of the listener [11]. Many researchers have developed different models for emotion recognition which are based on a kNN classifier, decision tree, random forest classifier, convolutional neural network, multilayer perceptron, etc. Regression and classification are two tasks that traditional machine learning can handle. They can be used to recognize textures or detect disorders in medical photos. Its key advantage is speed and relative simplicity. Furthermore, several of these algorithms can be interpreted by humans, making them useful for failure analysis, model refinement, and the finding of insights and statistical regularities. Ruiz-Garcia et al. [12] combined CNN for the extraction of features and SVM for the classification of emotions. Kulkarni et al. [13] in 2021 proposed a methodology that can distinguish genuine facial expression and unfold facial expression and improve results on CK+ and OULU-CASIA datasets. The proposed methodology extracts the features from the given video sequence and localized the facial landmarks. VGG-face deep network is fine-tuned to recognize facial expressions and to train the Emotion Network (EMNet). Ab et al. [14] developed a hybrid model based on CNN and kNN for Raspberry Pi and achieved 75.26% when tested on FER2013.

Tarnowski et al. [15] performed facial classification using k-NN (3-NN) classifier and multi-layer perceptron neural network. They tested two different methods for emotion recognition i.e. subject dependent and subject independent. They identified two emotions that were difficult to recognize i.e. sadness and fear as they were a little mixed up with neutral and surprise emotions. Salmam et al. [16] proposed a new feature extraction method that is based on a geometric approach. They used viola-jones algorithm for the face detection process. In their work, the CK+ and JAFFE database is used where six distances were calculated (using Euclidean, Manhattan, or Minkowski) for each face in the database. Calculated distance forms a

matrix that can be fed into a decision tree as input. Bailly et al. [17] used Random forest for dynamic pose-robust FER system in videos by incorporating conditional random forest to extract low-level expression variation patterns. They used pairwise conditional random forest to maintain head-pose variations in the FER system and achieved better performance on popular datasets as compared with state-of-the-art methods.

Traditional machine learning techniques such as Bayesian classifier and Support Vector Machine (SVM), k-NN, and decision tree do not provide the required performance in an unsupervised environment for face recognition in images. According to Nguyen et al. [18], it has been observed that the combination of classifiers can influence the accuracy of classification positively. Due to advancement in computing resources deep learning-based models gives extraordinary results for emotion recognition [9], [19], [20], [21].

2.2 Facial Emotion Recognition using Deep Learning Methods

Conventional or traditional machine learning algorithms tend to struggle with the large and highly complex model because the complexity increases with the dimensionality of the input data. To overcome these problems, deep learning has emerged. Deep learning incorporates many “efficient algorithms for extracting multiple levels of feature abstractions” and uses low-cost computation [22]. The most important concept in deep learning is representation [23] which refers to the automatic extraction of hierarchical features from the given input without having any prior knowledge [24]. They are more concerned with creating far larger and more complicated neural networks, and many of the methods are concerned with very huge datasets of tagged analog data, such as text, image, video, and audio, as mentioned previously. The comparative study of different algorithms evaluated on different datasets is shown

in Table 2.1 with column report: paper, an algorithm used, the database (used for training/testing), expression category, accuracy achieved, and limitation. The frequency of different models used in the literature is shown in Figure. 2-1. Among all the proposed methods, CNN is considered an effective algorithm for feature extraction and for developing different representations in data [25]. CNN is used as a feature extractor and classifier in state-of-the-art methods. It can be depicted from Fig. 2-1 that CNN is majorly used in the published work over the literature. But the incorporation of transformer-based models is also increasing due to its superior performance, handling long-range dependencies, and leverage attention mechanisms for better contextual understanding.

Table 2.1: Comparative study of different algorithms proposed in the literature.

Paper	Algorithm	Dataset	Expression	Accuracy	Limitations
Khattak et al. [26]	CNN	CK+, JAFFE	Micro-expression	95.65% (JAFFE)	Couldn't perform well on CK+ dataset
Bentoumi et al. [27]	VGG16, ResNet50 and MLP	CK+, JAFFE and KDEF	Macro-expression	100%, 100% and 98.78%	Model not investigated for dynamic images
Zhi et al. [28]	ResNet	CK+ and eINTERFACE'05	Macro-expression	97.2%, 94.5% and 98.6%	Not-performed well on major occlusion faces
Poux et al. [29]	Autoencoder	CK+ dataset	Macro-expression	91.1% (Eye occlusion), 85.9% (Mouth occlusion) and 75.2%	Not applied on the occluded faces due to variation in head pose.
Kaminska et al. [30]	ResNet	iCV-MEFED	Compound Emotion	21.83%	Dataset includes noisy data which affects training process.
Zhang et al. [31]	CNN	CK+, Oulu-CASIA and RAF-DB	Macro-expression	96.02%, 85.21% and 84.75% (RAF-DB)	Achieved less accuracy for disgust and fear expression for RaF-DB and Oulu-Casia Dataset
Fei et al. [32]	AlexNet and Linear Discriminant Analysis Classifier	KDEF, CK+, JAFFE, AffectNet, FER2013	Macro-expression	87.7%, 95.0%, 94.7%, 60.1% and 56.4%	Model not performed well on in-the-wild dataset
Sen et al. [33]	SVM	CK+ and MUG	Macro-expression	91.85% and 82.94%	Not validated on in-the-wild dataset
Sun et al. [34]	CNN and InceptionNet	CK+, MMI and RAFD	Macro-expression	98.38%, 99.17% and 99.59%	Not achieved highest accuracy on CK+ dataset
Aghamaleki et al. [35]	Multi stream CNN model	Image data (CK+ and MUG)	Macro-expression	99.61%	Accuracy achieved for negative emotions is comparatively less
Xu et al. [36]	SVM	SMIC, CASMEI and CASMEII	Micro-expression	71.43%, 42.02% and 41.96%	Need to design a framework for micro-expression recognition in long videos for real-time applications
Perveen et al. [37]	SVM	BP4D & Micro-expression	AFEW dataset	81.3%, 74.1%	Not well predicted expressions in temporally untrimmed clips

2.2.1 Basic Emotions Recognition

Jabboore et al. [38] proposed Fuse-CNN which detected the face using viola-jones haar cascade algorithm. Researchers used hybrid features utilizing a β -skeleton undirected graph along with an ellipse shaped by parameters trained through a 1D-CNN.

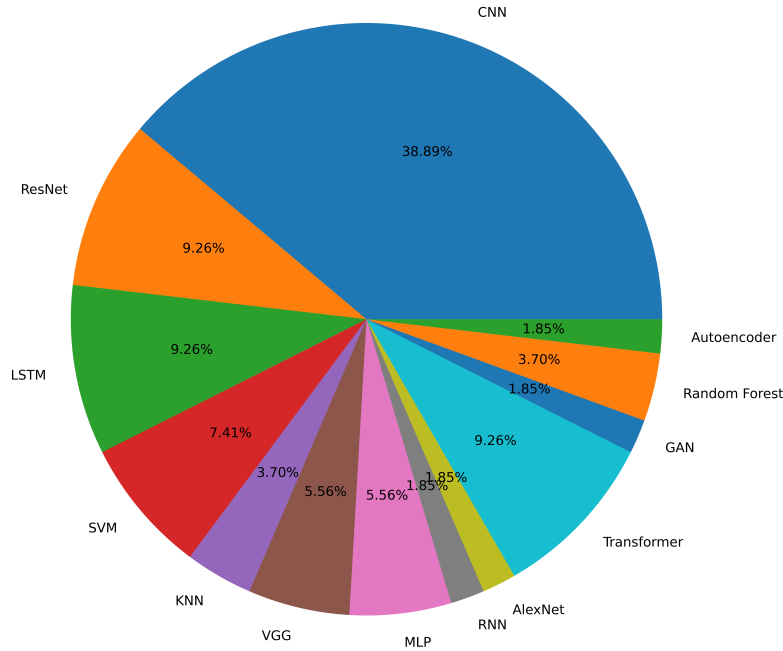


Figure 2-1: Frequency of models used in the literature.

Jin et al. [39], developed a double dynamic relationships graph convolutional network (DDRGCN) in which a facial graph is constructed using 20 regions of facial interest and enables the network to learn the relationships between the vertices during training. The performance of the proposed model is evaluated on four different datasets CK+ (94.32%), RaFD(94.48%), Oulu-CASIA(73.28%) and RAF-DB(58.25%) [39]. Wen et al. [40] used prior knowledge of facial recognition systems to develop domain information loss functions to improve the recognition of similar expression categories. Researchers applied dynamic objective learning in which different objectives are decided by different loss functions at each stage. Nguyen et al. [41] in 2022 developed a two-stage method and observed that besides high-level features, mid-level features also play a vital role in the classification process.

Kim et al. [42] proposed a facial expression recognition model where appearance and geometrical features are fused in a hierarchical manner. They proposed a technique to produce facial images with the help of neutral emotion using an autoencoder scheme and compared the results obtained for CK+ and JAFFE datasets which are 96.5% accuracy (1.3% higher than obtained from other algorithms) and 91.3% ac-

curacy (1.5% higher than obtained from other algorithms) respectively [42]. Li et al. [43] proposed a CNN-based model which includes joint learning using identity and emotion features which are further concatenated as a deep learned Tandem Facial Expression (TFE) feature for a fully connected layer. The proposed model [43] outperforms the state-of-the-art methods and achieves 99.31% and 84.29% accuracy for CK+ and FER+ databases. Fujii et al. [44] developed a model to recognize group-level emotions using a hierarchical classification method. Firstly, binary classification is performed to differentiate positive and non-positive emotions using scene features. Secondly, object-wise information is used for the classification of image into one of the following labels: positive, neutral or negative. Pre-trained VGG16 model is used for the classification whose parameters are optimized and comparatively less than the original model [44]. Group Affective Database 2.0 (GAF2) and Group Affective Database 3.0 (GAF3) dataset is used for the training of the model.

Zhang et al. [45] proposed a facial emotion recognition method using a convolutional neural network (CNN) and an image edge detection method. The proposed method is able to learn pattern features which can lower the inadequacy caused by artificially designed features. The model is taking less training and testing time with a high recognition rate as compared to FRR-CNN and R-CNN proposed in the respective literatures [46] [47]. Kim et al. [48] proposed a lightweight CNN architecture (86.58% accuracy) to improve the memory usage and operations performed for the embedded systems.

2.2.2 Facial Emotion Recognition In-the-wild

Liu et al. [49] proposed a patch attention unit that can perceive the occluded regions by learning local features for facial expression recognition, whereas self-attention of Visual Transformer is used to highlight the relevant patches with discriminative features to neglect the presence of occlusion in the facial images. Mu et al. [50] proposed a model based on Visual Transformers with feature fusion which is tested on in-the-wild expression datasets like RAF-DB, AffectNet, and FERPlus. Xue et al. [51] used the Vision Transformer model to produce relation-aware local representation

and the relation between them whereas Multi-head Attention Dropping (MAD) is also incorporated in the proposed model to drop the attention maps. Multi-head self-attention joins the relevant features using information subspaces at various locations. But different self-attention may produce identical relations which can be alleviated using Multi-head Self-attention Dropping (MSAD). Huang et al. [52] developed a FER system that exploited two attention mechanisms named grid-wise attention for low-level feature learning and Visual Transformers attention mechanism for high-level feature learning. The FER-2013 facial expression database and LFW (Labeled Faces in the wild) [53] data set are scientifically merged to design a simulation experiment which verified the robustness of the proposed method [45] and proved it under the complex background as well. Zhou et al. [54] developed a FER system for facial expression recognition of color images without converting them to gray-scale images. They used quaternion CNN to represent the RGB color in the form of quaternion metrics. QA-CNN (Quaternion CNN with attention mechanism) reduces the overhead by decreasing the number of network parameters by 75% as compared with real CNN structures [54]. As analyzed by the results, QA-CNN outperforms the quaternion CNN. QA-CNN model is evaluated on different datasets like Oulu-CASIA, MMI, SFEW and achieved 99.48%, 99.12%, 44.12% respectively.

2.2.3 Compound Emotion Recognition

Apart from the seven basic emotions, there are more complex emotions that enrich the understanding of human emotion. These complex emotions are made up of two basic emotions: “Happily Surprised”, “Happily Disgusted”, “Sadly Surprised”, “Angrily Surprised”, etc. Kollias et al. [55] introduced C-EXPR-DB, an in-the-wild audio-visual database containing 400 videos with 200,000 frames, annotated for 13 compound expressions, valence-arousal descriptors, action units, speech, facial landmarks, and attributes. They also propose C-EXPR-NET, a multi-task learning (MTL) model designed for compound expression recognition (CER) and action unit detection (AU-D), where AU-D is included to improve CER accuracy. For AU-D, they incorporate both AU semantic descriptions and visual information. Jarraya et al. [56] created a

system to recognize compound emotions in autistic children during meltdown crises. The study focused on analyzing deep spatio-temporal geometric features of involuntary facial expressions in these crisis moments. The researchers compared various feature selection techniques, including filter methods, wrapper methods, and the Neighborhood Component Analysis (NCA) method. They achieved the highest accuracy, reaching 85.5%, by using an RNN classifier combined with the Information Gain feature selection method [56].

Shaila et al. [57] employed the MobileNet model, utilizing depthwise convolution operations, to recognize compound emotions. Liang et al. [58] developed a model called the “two-stream multi-scale AU-based Network (MSAU-Net)” to predict fine-grained emotions. Kaminska et al. [30] proposed a two-stage model for recognizing compound emotions. In the first stage, they performed a coarse recognition by combining appearance-based features extracted with a DCNN and facial-point features. The second stage involved fine recognition using a binary classifier. Additionally, the authors [30] experimented with an ensemble approach that combined one-stage and two-stage models, ultimately enhancing the recognition rate from 19.28% to 21.83%.

2.2.4 Driver’s Emotion Recognition

Research [59] indicates that co-passenger alerts can reduce the risk of mishaps or collisions. In future vehicles, incorporating driver emotion detection systems could enhance safety by raising an alarm when unfavorable emotions are detected, helping keep the driver alert. Monitoring driver emotions is thus a crucial component of Advanced Driver Assistance Systems (ADAS), which are being developed to protect drivers and passengers by alerting drivers to potential risks. However, even the most advanced autonomous vehicles still require the driver to remain vigilant and ready to take control during emergencies. Tavakoli et al. [60] examined the impact of external factors on the driver’s state by analyzing facial expression data, gaze variability, stress levels, and workload. Yang et al. [61] proposed a robust driver emotion recognition method that addresses the challenges of individual differences and lighting variations by separating relevant features from irrelevant ones. To achieve this, they developed

a high-purity feature separation (HPFS) framework, which utilizes partial feature exchange and multiple loss function constraints.

Zaman et al. [62] employed various models, including convolutional neural networks (CNN), recurrent neural networks (RNN), and multi-layer perceptron classification models, to build an ensemble CNN-based model for enhanced driver facial expression recognition. They applied a feature-fusion technique to combine the features extracted from three CNN models, which were then used to train the proposed ensemble classification model. To improve face detection accuracy and efficiency, they replaced the enhanced Faster R-CNN feature-learning block with a new CNN block, InceptionV3. Mou et al. [63] proposed a novel approach for recognizing driver emotions by using a multimodal fusion framework that combines convolutional long-short term memory networks (ConvLSTM) and a hybrid attention mechanism. This framework integrates non-invasive multimodal data from the eyes, vehicle, and environment.

2.3 In Thesis Prospective:

In this thesis, four frameworks have been proposed for facial emotion recognition. In the context of FER, attention mechanisms have been shown to significantly boost performance by emphasizing relevant facial regions that contribute most to emotion expression. So, we provided a detailed understanding of different attention mechanisms in the field of facial emotion recognition and observed how their incorporation impact the performance of the deep learning model. Occlusion, pose variance, and illumination conditions are very common hindrances in automating facial emotion recognition for the real-time environment. To address this problem, we proposed a novel framework named FMR-CapsNet for predicting basic emotions in an uncontrolled environment. This model utilizes relation-aware geometric features which provide detailed insight into facial features even for images with occlusion and pose variance. The geometric features are extracted using FaceMesh Mediapipe which provides detailed 3D Mesh representation which is better than 2D methods and hence

improves classification accuracy. The geometric features are further refined using a ResNet50 pre-trained model and refined features are finally propagated through a capsule neural network to capture inter-feature relationships and spatial information. This capability of recognizing emotions of in-the-wild facial images increases its applicability in a real-time environment. The model is evaluated on in-the-wild datasets like RAF-DB and AffectNet.

Basic emotions provide a primal emotional understanding of humans whereas compound emotion provides a richer and more accurate understanding of one's emotions. For this, a lightweight Swin Transformer-based model named LSwin-CBAM is proposed. Due to the limited size of the compound emotion dataset, the CutMix augmentation method is used to augment the data. A modified Swin Transformer model (with fewer Swin Transformer blocks) is used with the CBAM attention mechanism to recognize the compound emotions with better accuracy on RAF-DB (compound) and EmotioNet datasets. Facial emotion recognition has vast applications in the real-world. In this thesis, we worked on an application of FER i.e. driver emotion recognition which helps to alert the driver for his unusual behavior while driving. For this work, a Vision Transformer-based model is used with shifted patch tokenization (which can solve the problem of small-size datasets). The proposed method is evaluated on the driver's emotion dataset and performs better performance as compared to the state-of-the-art.

Chapter 3

Investigation of Different Mechanisms on Facial Emotion Recognition

Nidhi, Bindu Verma. “A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition”. *In 2023 Seventh International Conference on Image Information Processing (ICIIP)* (pp. 591-596) (2023, November). IEEE. (Published)

Nidhi, Bindu Verma. “Empirical Insights: Unraveling the Impact of Various Attention Mechanisms on Facial Emotion Recognition” presented in *5th International Conference on Data Analytics & Management (ICDAM-2024)*, Springer. (Accepted & Presented)

3.1 Introduction

Attention mechanisms are the methods used to divert the attention to the relevant regions of an image and neglect irrelevant ones. The neural network’s ability to focus on specific regions, eliminate irrelevant information, and effectively extract essential features from images is made possible by using attention mechanisms. This chapter presents an empirical study on the effectiveness of various attention mechanisms in facial emotion recognition (FER). Considering this aspect, four attention techniques are explored and compared: channel attention, spatial attention, CBAM attention,

self-attention, and multi-head attention (MHA). This chapter provides an insight to relevance of each attention mechanism for facial emotion recognition.

Facial emotion recognition (FER) is a substantial area of research in computer vision and artificial intelligence, aiming to understand and interpret human emotions from facial expressions. It involves the detection and classification of emotions expressed through facial features. Facial expression recognition technology offers a wide variety of practical uses such as enhancing public safety, managing and retrieving information, engaging with social media, conducting psychological research, and improving patient care. These systems can be strategically deployed in various public venues such as bus stations, railway stations, airports, or stadiums to aid security personnel in identifying potential risks. Traditional approaches relied on features extracted from handcrafted methods and machine learning algorithms, but recent advancements, especially the integration of deep learning, have significantly improved accuracy. Conventionally, research in this domain has primarily focused on controlled environments, where standardized stimuli and controlled lighting conditions enable precise analysis. However, human emotions are inherently dynamic and multifaceted, often expressed in unpredictable and uncontrolled settings, commonly referred to as “in-the-wild”. The recognition of facial emotions in such naturalistic environments presents a significant challenge due to various factors, including variability in illumination, facial expressions, occlusions, and background clutter. Despite these challenges, recent advancements in computer vision, machine learning, and deep learning techniques have spurred a growing interest in addressing facial emotion recognition in-the-wild.

In recent years, attention mechanisms have gained prominence in enhancing the performance of FER systems by selectively focusing on relevant facial regions. This chapter explores the attention mechanisms in the context of facial emotion recognition. Attention mechanisms, inspired by human perception, are techniques employed to direct focus toward pertinent areas within an image while disregarding irrelevant ones. The neural network’s ability to concentrate on specific regions, filter out irrelevant details, and adeptly distill crucial features from images is facilitated through

the application of attention mechanisms.

Furthermore, attention mechanisms contribute to the adaptability of FER models to varying facial expressions, lighting conditions, and facial occlusions. By dynamically adjusting the weights assigned to different facial regions, attention mechanisms enable models to maintain robustness in the presence of challenging factors. This adaptability is crucial for real-world applications where facial emotion recognition needs to perform accurately under diverse and unpredictable scenarios. The main contribution of this chapter is the introduction of an innovative FER model for in-the-wild datasets with empirical analysis of various attention mechanisms that systematically evaluate their efficacy in the domain of FER. This chapter introduces a capsule network-based FER model tailored for in-the-wild dataset (RAF-DB) or images, providing a hands-on evaluation of attention mechanisms, including “Channel attention”, “Spatial attention”, “Self-attention”, “CBAM attention”, and “Multihead attention” within the FER paradigm. Empirical study has been conducted to compare the effectiveness of each attention mechanism. The computational complexity and interpretability of attention mechanisms has been examined for practical feasibility.

3.2 Literature Survey

Attention mechanisms play a vital role in enhancing the classification accuracy of facial emotions by enabling models to selectively focus on salient facial features. One notable advantage of attention mechanisms is their ability to capture and prioritize essential information, thereby improving the model’s discrimination power. In the context of facial emotion recognition (FER), attention mechanisms have been shown to significantly boost performance by emphasizing relevant facial regions that contribute most to emotion expression.

For instance, studies have demonstrated the effectiveness of attention mechanisms in models such as the Residual Attention Network (RAN) and Vision Transformer (ViT). The RAN model employs spatial attention to highlight critical spatial locations

in the input image, while ViT utilizes transformer-based attention to capture intricate details across the entire facial region. These mechanisms allow the models to focus on relevant facial expressions, effectively filtering out irrelevant information and noise, leading to more accurate emotion classification.

Zhou et al. [54] developed a Facial Expression Recognition (FER) system for color images without the need for conversion to grayscale. Their approach utilized quaternion Convolutional Neural Networks (CNNs) to represent RGB color in quaternion metrics. The proposed QA-CNN (Quaternion CNN with attention mechanism) significantly reduced overhead by reducing the number of network parameters by 75% compared to traditional CNN structures [54]. Liu et al. [49] presented a patch attention unit to perceive occluded regions, employing local features for facial expression recognition. They utilized self-attention of visual transformers to emphasize relevant patches with discriminative features, effectively mitigating the impact of occlusion in facial images. Huang et al. [52] introduced a FER system incorporating two attention mechanisms: grid-wise attention for low-level feature learning and visual transformers attention mechanism for high-level feature learning. Hu et al. [64] proposed a Squeeze and Excitation model, focusing on sequentially learning weights of feature maps to enhance the quality of features while diluting irrelevant ones. Mittal et al. [65] introduced a capsule network-based model incorporating a CBAM attention module for detecting driver distraction.

3.3 Proposed work

Facial emotion recognition, a vital component of human-computer interaction and affective computing, plays a pivotal role in understanding human behavior and enhancing user experiences across various domains. In this proposed work, a novel deep learning-based model for accurate facial emotion recognition is presented. Emotion recognition from facial expressions is a quite challenging due to the variability and complexity of human emotions, as well as environmental factors such as lighting conditions and occlusions. This model addresses these challenges by leveraging attention

mechanisms in deep learning models to effectively capture and interpret subtle facial cues indicative of different emotions. The proposed model takes grayscale facial images as input (64 x 64). The input image is subjected to a 2D convolution (Conv2D) process, where a small matrix (referred to as a filter or kernel) slides across the input image. This operation entails element-wise multiplication and summation to generate a feature map. Subsequently, the ReLU (“Rectified Linear Unit”) activation function is applied to the feature map element-wise, introducing non-linearity by nullifying negative values. The attention module in neural networks focuses on specific parts of the input data while ignoring others. It assigns weights to different regions of the feature maps based on their importance in capturing emotional cues. The attentive features from the attention module are propagated through a primary capsule layer of a Capsule Neural Network (CapsNet). It groups features into capsules, where each capsule represents a set of neural units encoding different properties of the input, such as pose, texture, or facial features. Primary capsules capture spatial hierarchies and relationships within the data, preserving important structural information. The digit caps layer further refines the features propagated from the primary capsule layer. Capsules in this layer represent specific classes or categories (Class 0: “Surprise”, 1: “Fear”, 2: “Disgust”, 3: “Happiness”, 4: “Sadness”, 5: “Anger”, 6: “Neutral”). Each capsule in the digit caps layer encodes information about the presence and characteristics of a particular class, facilitating robust classification or recognition of the input data. The final output of the model is generated based on the information encoded in the digits caps layer. The Figure 3-1 illustrates the sequential flow of operations within the facial emotion recognition model, starting from the input image and progressing through convolutional and capsule layers to produce a prediction of the emotional state depicted in the input facial expression.

3.3.1 Capsule Neural Network

Capsule networks, inspired by the hierarchical structure of the human visual system, aim to address the shortcomings of traditional neural networks by introducing "capsules" as fundamental units of representation. These capsules not only encode

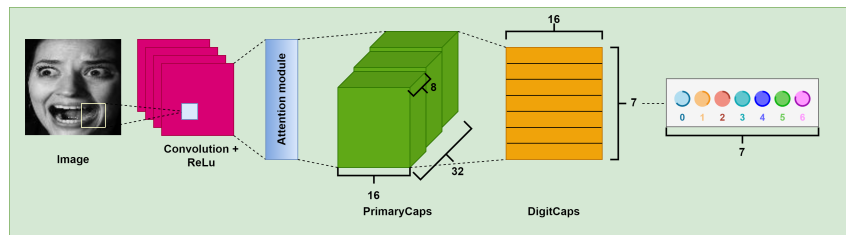


Figure 3-1: Attentive capsule neural network

the presence of features but also their instantiation parameters, such as pose and orientation, enabling the network to learn hierarchical relationships more effectively.

A capsule neural network is composed of multiple layers of capsules, which are specialized units designed to represent and learn features hierarchically. Unlike traditional neural networks where neurons in consecutive layers are densely connected, capsules in CapsNet are organized in a hierarchical structure to efficiently capture spatial relationships and pose information. The input to a CapsNet is typically an image or a set of features extracted from an image. Each input is represented as a set of vectors or feature maps. The first layer of a CapsNet often consists of convolutional capsules. These capsules employ convolutional operations to bring out the low-level features from the given input image. Each convolutional capsule consists of a set of convolutional units, each responsible for detecting specific local patterns within the input. The output of the convolutional layer is fed into primary capsules. Each primary capsule represents a higher-level feature detected by the convolutional capsules. Primary capsules are typically arranged spatially, capturing spatial hierarchies of features. Each primary capsule outputs a vector representing the instantiation parameters of a specific feature, such as pose, orientation, and scale. In CapsNet model, the PrimaryCaps layer consists of eight capsules, with each capsule containing sixteen-dimensional features. Additionally, the contribution ($\hat{u}_{j|i}$) of each capsule u_i in the PrimaryCaps layer to that of v_j DigitCaps was computed using Eq. (3.1).

$$\hat{u}_{j|i} = W_{ij} \cdot u_i \quad (3.1)$$

In the DigitCaps layer, there is a sixteen-dimensional capsule v_j allocated for each

digit class (seven classes in this experiment). These capsules obtain input from all capsules in the PrimaryCaps layer using Eq. (3.2), Eq. (3.3), Eq. (3.4).

$$cap_{ij} = \frac{\exp(b_{ij})}{\sum_l \exp(b_{il})} \quad (3.2)$$

$$c_j = \sum_l cap_{ij} \hat{u}_{j|i} \quad (3.3)$$

$$v_j = \frac{\|c_j\|^2}{1 + \|c_j\|^2} \frac{c_j}{\|c_j\|} \quad (3.4)$$

At last, the margin loss is computed for each digit capsule to classify the facial expressions using Eq. (3.5) where $Y_l = 1$ if there is relation 1, $r^+ = 0.9$, $r^- = 0.1$, and $\lambda = 0.5$.

$$D_l = Y_l \max(0, r^+ - \|v_l\|)^2 + \lambda(1 - Y_l) \max(0, \|v_l\| - r^-)^2 \quad (3.5)$$

3.3.2 Different Attention Mechanisms

Channel Attention

The concept underlying channel attention involves emphasizing informative channels while selectively suppressing less relevant ones. Within a convolutional neural network (CNN), each convolutional layer typically encompasses multiple channels, each corresponding to a distinct feature map. The channel attention mechanism calculates a series of attention weights specific to each channel within a layer’s feature maps. These weights are then utilized to adjust the feature maps, prioritizing attention to informative channels and diminishing attention to less relevant ones. To efficiently compute channel attention, it is necessary to compress the spatial dimension of the feature map, achieved through the application of an average-pooling operation as suggested by Woo et al. [66]. Additionally, for more detailed channel-wise attention, the implementation of max-pooling gathers another crucial set of information related to object features.

At the outset, spatial information from a feature map undergoes aggregation through average and max-pooling operations. This procedure yields average-pooled features represented as F_{avg}^c and max-pooled features denoted as F_{max}^c , as depicted in Eq. (3.6). These distinct features are individually input into a shared network, comprising a shared multi-layer perceptron (MLP) with a single hidden layer, resulting in output feature vectors. Subsequently, the output feature vectors are combined through element-wise summation, as detailed in Eq. (3.7).

$$F_{avg}^c = AvgPool(M), F_{max}^c = MaxPool(M), \quad (3.6)$$

$$\begin{aligned} A_c(F) &= \sigma(MLP(F_{avg}^c) + MLP(F_{max}^c)), \\ &= \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))), \end{aligned} \quad (3.7)$$

Spatial Attention

The spatial attention module utilizes both the feature maps and the channel-wise attention weights to produce a series of spatial attention maps, facilitating selective focus on different regions of the image [66]. Initially, average-pooling and max-pooling operations are conducted along the channel axis and merged to generate an efficient feature descriptor, highlighting pertinent regions. Subsequently, a convolution layer with a filter size of 7×7 ($f^{7 \times 7}$) convolves over the concatenated features to create a spatial attention map $A_s(M) \in I^{H \times W}$. This map indicates where attention should be directed or suppressed. The computation of the spatial attention map is articulated in Eq. (3.8) and Eq. (3.9).

$$F_{avg}^s = AvgPool(M), F_{max}^s = MaxPool(M), \quad (3.8)$$

$$A_s(F) = \sigma(f^{7 \times 7}([F_{avg}^s; F_{max}^s])) \quad (3.9)$$

Convolutional Block Attention Module (CBAM)

The ‘‘Convolutional Block Attention Module’’ (CBAM) attention mechanism combines both channel-wise and spatial attention mechanisms. By incorporating these attention mechanisms into CNNs, CBAM is intended to enhance the network’s capability to emphasize significant features and elevate its performance across diverse computer vision applications, including object detection and image classification.

CBAM attention sequentially produces attention maps in both channel and spatial dimensions from an intermediate feature map. Subsequently, the input feature map undergoes refinement by being multiplied with these attention maps, enhancing the adaptive features. This universal and lightweight module effortlessly integrates into any CNN-based architecture, introducing minimal additional computational overhead.

CBAM accepts an intermediate feature map represented by $M \in I^{C \times H \times W}$ and applies a 1D channel attention map $A_c \in I^{C \times 1 \times 1}$ and 2D spatial attention map $A_s \in I^{1 \times H \times W}$ in a sequential manner. The attention operation can be formulated in Eq. (3.10). The Eq. (3.10) following it describes the process of element-wise multiplication, where channel and spatial attention values are copied along spatial and channel dimensions respectively, and give M'' as the final output.

$$M' = A_c(M) \odot M, M'' = A_s(M') \odot M', \quad (3.10)$$

Self-attention

Self-attention mechanisms, also referred to as intra-attention, capture connections among various positions within a single input sequence or image [67]. In the realm of images, these mechanisms empower the model to comprehend distant relationships between pixels. Originally designed for natural language processing tasks, the transformer architecture incorporates a self-attention mechanism to assign different weights to input sequence elements, depending on their interrelations. Suppose Q, K, and V represent the query, key, and value matrices, respectively. These matrices serve

as inputs to the attention function, as depicted in the equation labeled Eq. (3.11).

$$AT(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.11)$$

Incorporating a scaling factor in the self-attention mechanism is vital to maintaining balanced gradients during back-propagation. This precaution prevents gradients from becoming excessively large (leading to exploding gradients) or overly small (resulting in vanishing gradients). This aspect is critical for fostering stability and efficacy in training deep neural networks, particularly in tasks that utilize self-attention mechanisms including textual and sequential data.

Multi-head attention exploits self-attention and the ability to focus on different parts of the input data [67]. It allows the model to jointly attend to different positions (spatial locations or channels) in different ways. This mechanism can enhance the expressive power of the model by capturing various aspects of the input data. The MHA layer divides the input into various heads to learn different levels of self-attention.

Multi-head Attention (MHA)

Multi-head attention functions by dividing the input into several sets, commonly called “heads”, each comprising queries, keys, and values. Each head independently computes attention, producing multiple sets of weighted representations for the input. These outputs from different heads are then amalgamated and subjected to a linear transformation, culminating in the ultimate output of the multi-head attention layer as shown in Eq. (3.12).

$$h_i = AT(QW_i^Q, KW_i^K, VW_i^V), MHA(Q, K, V) = Concatenate(h_1, h_2, \dots, h_h)W_0 \quad (3.12)$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W_0 \in \mathbb{R}^{hd_v \times d_{model}}$.

3.4 Experimental Results

Experiments are performed on 64-bit Windows 11, Intel Core i7 processor with 16GB RAM size, 8GB NVIDIA TITAN RTX graphics card. MSE loss function is a commonly employed in classification and regression problems. It calculates the average of the squares of the differences between the predicted and actual values. AdamW optimization algorithm combines the benefits of the Adam optimizer and weight decay. It is widely used in training neural networks due to its adaptive learning rate and momentum properties. Batch size represents the count of data samples processed before the model’s parameters are updated during the training process. A batch size of 16 means that 16 data samples are processed in each iteration. The learning rate dictates how quickly the model’s parameters are updated throughout the training process. It controls the step size in the parameter space. A smaller learning rate typically leads to slower but more stable training. Learning rate decay is the hyperparameter that controls the rate at which the learning rate decreases over time during training. A decay factor of 0.9 means that the learning rate is multiplied by 0.9 after each epoch or a certain number of iterations. Number of routings is the hyperparameter which is specific to CapsNet and determines the number of iterations (or "routings") performed in the dynamic routing algorithm during inference. It affects how capsules communicate and reach an agreement about the instantiation parameters of higher-level capsules.

3.4.1 Experimental results on RAF-DB dataset

This chapter explores how various attention mechanisms affect the performance of a neural network model. The Table 3.1 focuses on five attention mechanisms: “Channel Attention”, “Spatial Attention”, “CBAM Attention”, “Self-attention”, and “Multi-head attention”. The analysis is based on various metrics including parameter count, parameter size, model complexity, and training time per epoch. The channel Attention, spatial Attention, and CBAM attention mechanisms have similar parameter counts and sizes, indicating comparable levels of model complexity. Self-attention, however,

exhibits a significantly higher parameter count and size, suggesting a more complex model architecture. In contrast, multi-head attention has a substantially lower parameter count and size, indicating a simpler model architecture compared to the other mechanisms. Self-attention requires the highest training time per epoch, almost four times longer than the other mechanisms. This is likely due to its higher model complexity and the computational overhead associated with processing large numbers of parameters. Multi-head attention, on the other hand, exhibits the lowest training time per epoch, indicating faster convergence during training.

Table 3.1: Computational analysis of different attention mechanisms *I= Number of input channels, C= number of channels, H= height of the feature map, and W= width of the feature map.

Attention Mechanism	Parameter count	Parameters Size	Complexity	Training Time (s/epoch)
Channel Attention	26.658M	101.60MB	$O(I * C^2)$	138s
Spatial Attention	26.626M	101.57MB	$O(N * H * W)$	136s
CBAM Attention	26.659M	101.70MB	$O(N * C^2 + N * H * W)$	141s
Self-attention	27.086M	103.33MB	$O(N^2 * D)$	380s
Multi-head attention	5.914M	22.56MB	$O(H * N^2 * D)$	24s

In Table3.2, we explore the influence of different attention mechanisms within CapsNet on the task of facial emotion recognition using RAFDB dataset (in-the-wild facial emotion recognition dataset). The study focuses on five CapsNet variants, each employing a distinct attention mechanism: CapsNet-C (incorporating Channel attention), CapsNet-S (Spatial attention), CapsNet-CBAM, CapsNet-Self, and CapsNet-MHA.

Table 3.2: Comparative analysis of different attention mechanisms on RAF-DB dataset.

Model	Epochs	Accuracy	AUC Score	F1-score	Precision	Recall
CapsNet	23	75.34%	92.39	0.7537	0.75	0.75
CapsNet-C	19	75.96%	93.02	0.7593	0.80	0.80
CapsNet-S	12	76.25%	94.40	0.7622	0.83	0.83
CapsNet-CBAM	15	76.12%	94.57	0.7604	0.50	0.50
CapsNet-Self	19	71.77%	91.51	0.7170	0.50	0.50
CapsNet-MHA	27	67.54%	92.11	0.6755	0.58	0.58

3.4.2 Analysis of different Attention Mechanisms

CapsNet-S achieves the highest accuracy of 76.25%, closely followed by CapsNet-CBAM at 76.12%. CapsNet-C also performs reasonably well with an accuracy of 75.96%. Conversely, CapsNet-Self and CapsNet-MHA exhibit lower accuracies of 71.77% and 67.54% respectively. CapsNet-S demonstrates the highest F1 score of 94.40%, indicating a balanced performance in terms of precision and recall. CapsNet-CBAM follows closely with an F1 score of 94.57%. CapsNet-C also shows a respectable F1 score of 93.02%. CapsNet-Self and CapsNet-MHA exhibit lower F1 scores of 91.51% and 92.11% respectively. CapsNet-S achieves the highest precision and recall values among all variants, indicating its ability to make accurate positive predictions while capturing a high proportion of true positive instances. CapsNet-CBAM also shows competitive precision and recall values. CapsNet-Self and CapsNet-MHA exhibit lower precision and recall values, suggesting challenges in making accurate positive predictions and capturing true positive instances. CapsNet-S achieves the highest specificity of 0.83, indicating its ability to accurately identify negative instances. CapsNet-C and CapsNet-Self exhibit similar specificity values of 0.80, while CapsNet-MHA shows a slightly lower specificity of 0.58. CapsNet-CBAM demonstrates the lowest specificity of 0.50. The Figure 3-3 shows ROC_AUC curves for CapsNet (a), CapsNet-C (b), CapsNet-S (c), CapsNet-CBAM (d), CapsNet-Self (e), CapsNet-MHA (f).

Overall, CapsNet-S and CapsNet-CBAM emerge as strong contenders in terms of accuracy, F1 score, precision, recall, and specificity, showcasing their effectiveness in facial emotion recognition tasks within CapsNet. CapsNet-C also demonstrates competitive performance, while CapsNet-Self and CapsNet-MHA exhibit comparatively lower performance across various metrics. These findings highlight the significance of attention mechanisms in CapsNet and provide valuable insights for designing efficient models for facial emotion recognition.

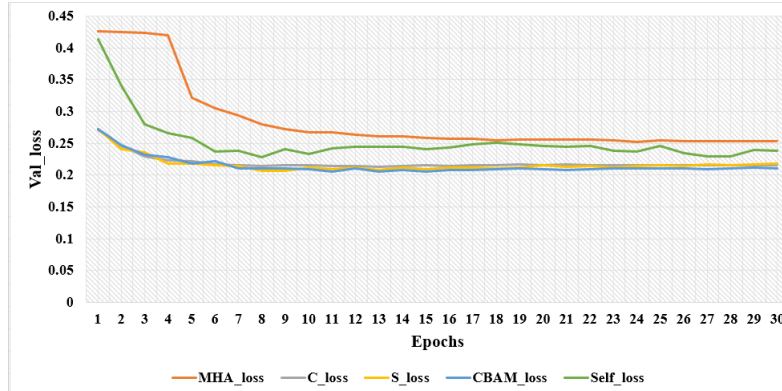


Figure 3-2: Validation loss of CapsNet with different attention mechanisms.

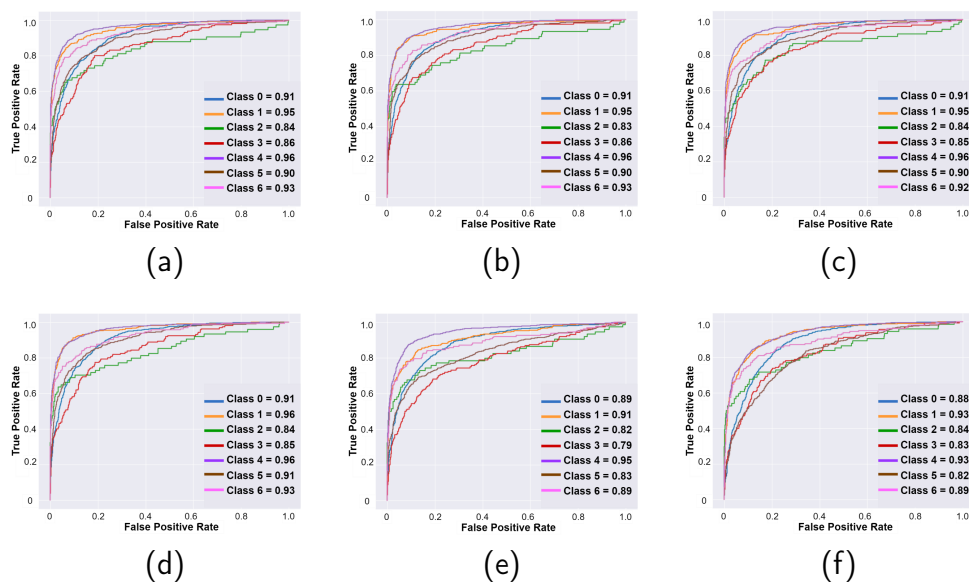


Figure 3-3: ROC_AUC Curves of applied attention mechanisms with CapsNet *(a) CapsNet (b) CapsNet-C, (c) CapsNet-S, (d) CapsNet-CBAM, (e) CapsNet-Self, (f) CapsNet-MHA.

3.5 Conclusion

In conclusion, this chapter introduces a novel deep learning-based model for accurate facial emotion recognition, focusing on in-the-wild datasets. As part of the novelty of this chapter, we conducted a comprehensive survey of various attention mechanisms within the field of Facial Emotion Recognition (FER). This chapter highlights the advancements and applications of attention mechanisms and explores how they contribute to improving the accuracy and efficiency of emotion recognition systems. Through extensive experiments, attention mechanisms are investigated within Cap-

sNet architecture, showcasing their effectiveness in enhancing model performance. CapsNet-S and CapsNet-CBAM emerge as top contenders, highlighting the significance of attention mechanisms in Capsule Networks for facial emotion recognition tasks. The study contributes valuable insights into designing efficient models for real-world applications and underscores the importance of attention mechanisms in improving FER accuracy.

Chapter 4

Design and development of a framework for a Facial Emotion Recognition system in-the-wild

Nidhi, and Bindu Verma. “In-the-Wild Facial Emotion Recognition using Relation-aware Geometric Features and CapsNet” is communicated in *Computers and Electrical Engineering* (SCIE Indexed, IF: 4.0) (Communicated)

4.1 Introduction

This chapter introduces a novel and effective framework for a facial emotion recognition system in-the-wild. In the previous chapter, we simply developed a deep-learning model to show the efficacy of different attention mechanisms that uses facial images as input, features were extracted within the model itself rather than separately. However, we explored that if the face is occluded or pose variant then the attention mechanism will not be able to focus on the defined facial emotion. Thus, in this chapter, we tried to resolve the problem of pose variations and occlusion by employing the FaceMesh model for geometric feature extraction, utilizing facial blendshape scores that adeptly capture features from pose-variant and occluded images. Addi-

tionally, a capsule neural network is employed to capture both directional and spatial information, improving the accuracy of inter-class feature differentiation.

Although humans are quite adept at identifying a person’s emotional states, but it is a highly complex process for a computer. In-the-wild refers to the complexity arises from various factors such as variations in lighting conditions, head positions, illumination, and occlusion. These challenges are usually found in unconstrained environment like images collected from internet where background noise is complimentary included to pose the challenge for automatic facial emotion recognition system. A primary solution to address these challenges is to represent facial expressions using more robust features [68]. Recently, due to the excellent performance of deep networks in various tasks of computer vision, features extracted by deep learning (DL) networks have proven to be more robust and effective. Consequently, deep learning features have steadily taken the place of traditional hand-crafted ones, leading to a significant improvement in facial expression recognition performance.

However, with non-posed images, where the entire face is not clearly visible, the resulting features propagates partial information to the classification model. Research in physiology and psychology [69] [70] shows that distinctive features of various facial expressions are not uniformly distributed across the entire face. Certain key areas, like the eyes and mouth, are particularly effective at reflecting differences between expressions. Since different facial expressions exhibit specific local variations in these critical regions, this insight encourages us to focus on these areas to enhance facial expression recognition. Similarly, for occluded facial images, where subjects are wearing glasses or some parts of the image are covered by objects, it is challenging to extract relevant facial features from the obscured areas.

It is crucial to capture facial features from all the key regions of the face to make precise recognition, but most SOTA methods for facial expression recognition works on visible part of the face which decrease the classification rate of facial expressions. Inspired by this observation, this chapter contributes to the detection of facial emotion of in-the-wild datasets such as AffectNet [71], RAF-DB [72], etc. which comprise real-time pictures including pose variant and occluded images.

In this chapter, a novel robust and effective framework (FMR-CapsNet) for facial emotion recognition is proposed. FMR-CapsNet framework is comprised of three modules feature extraction module, relation-aware blendshapes module, Resnet50 features, and capsule neural network module. The feature extraction module extracts geometric features from facial images which are further used to compute blendshape scores and hence, captures spatial relationships between them. For feature extraction, FaceMesh mediapipe is used which gives feature matrix consisting of 52 facial blendshapes for each image. Euclidean distance is then utilized to assess spatial relationships among facial landmarks, offering valuable insights into visible facial regions and enhancing the proposed model’s robustness to occlusion. These features then undergo refinement before being classified using a capsule neural network. The features are refined using ResNet50 pretrained model. Refined features are propagated to capsule neural network to capture hierarchical relationships between features, allowing it to learn hierarchical representations of facial expressions. The contribution of this chapter is as follows: To address the problem of pose-variant and occluded facial images, FaceMesh model is used for the geometric feature extraction process. The extracted features are facial blendshapes scores which can capture features from side face images and occluded images. Additionally, euclidean distance is used to evaluate the distance matrix for relation-aware blendshapes to provide the relative information between each blendshapes scores. For refinement of extracted features, pretrained ResNet50 is used to implement transfer learning on the evaluated distance matrix. Capsule neural network is used to capture the directional and spatial information to differentiate the inter-class features more accurately. The proposed method FMR-CapsNet is evaluated on two in-the-wild datasets, namely RAF-DB and AffectNet. The results demonstrate that FMR-CapsNet surpasses the SOTA methods.

4.2 Literature Survey

Facial emotion recognition has captivated significant attention in the realms of computer vision and artificial intelligence, leading to substantial contributions from nu-

merous researchers. Leveraging both ML (machine learning) and DL algorithms, these studies have explored diverse methodologies and applications to increase the accuracy and efficiency of emotion detection from facial expressions. But it still remains a significant challenge for facial images captured in uncontrolled environments due to the considerable variations in lighting, head positions, and individual attributes.

Liu et al. [73] introduces a geometry-aware conditional network (GACN) designed to retrieve long-range dependencies, enabling simultaneous pose-invariant facial emotion editing and geometry-aware facial emotion recognition (FER). Arrazola et al. [74] extracted 68 facial landmarks and identified the muscle activity and interactions on performing specific facial expressions to form a feature vector which is classified using three classifiers: Multi-layer perceptron (MLP), Support Vector Machine (SVM), Linear Discriminant (LD), etc. Yu et al. [75] proposed a joint training method for a FER model using multiple datasets. This involves selecting a subset from a supplemental dataset, producing and refining pseudo-continuous labels for the target dataset, and jointly training the model with multi-task learning.

Due to the substantial progress in computational capabilities and the development of advanced network architectures, researchers across various domains have increasingly turned their attention to deep learning techniques. Consequently, these deep learning methods have attained good classification accuracy and substantially surpassed previous facial expression recognition benchmarks (FER). Donahue et al. [76] proposed a model that integrates spatial and temporal depth, termed as long-term recursive convolutional network. This model combines the output of a CNN with an LSTM to handle visual tasks that involve time-varying inputs and outputs. Li and Deng [72] introduced a novel method called Deep Locality-Preserving CNN (DLP-CNN). This approach aimed to improve the discriminative capability of deep features by maintaining the similarity within classes while maximizing the dissimilarity between classes.

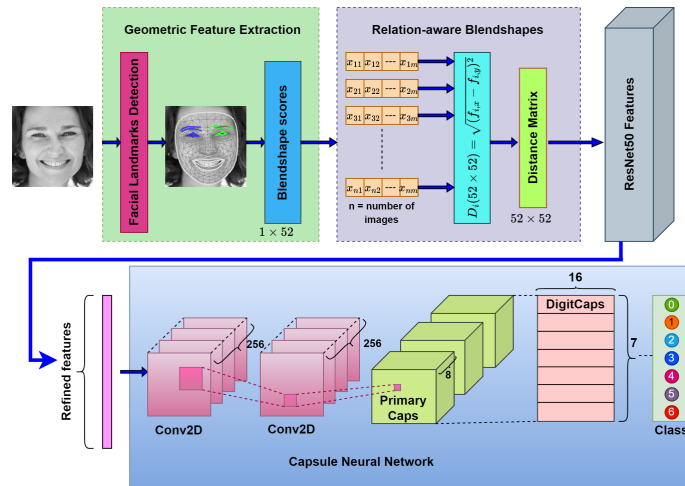


Figure 4-1: The proposed architecture of FMR-CapsNet incorporating Facemesh mediapipe for geometric feature extraction, ResNet50 for refinement of features, and CapsNet for emotion classification for facial emotion recognition in-the-wild.

4.3 Proposed Work

The convolutional neural networks for facial expression recognition (FER) overlook the face’s structural characteristics and the features’ directional information. Additionally, they fail to consider the relative relationships between different facial landmarks. To overcome these problems, Capsule neural network [77] is proposed to detect spatial information between features of different regions. Dong et al. [78] proposed Caps-BiLSTM that incorporates a capsule network with Bi-LSTM for sentiment analysis. Wang et al. [79] developed a capsule network with a novel Capsule filter routing method to discard the capsule with low activation values. To discriminate between capsules of different relevance, the Enhanced Capsule Attention Module (ECAM) is proposed to distribute the weight of each capsule between and inside the capsule dimension [79].

Considering the literature, the proposed method incorporates a capsule neural network (CapsNet) to direct the refined features and to retrieve the relative relationship between geometric features. CapsNet includes dynamic routing between capsules to encode the intermediate features and finally produce the output using squash function.

Analyzing existing literature shows that facial expression recognition is quite chal-

lenging for the images captured in an unconstrained environment due to several factors like pose variants, occlusion, lighting conditions, etc. So, the FMR-CapsNet model is proposed to overcome these challenges which can capture relation-aware geometric features of pose variant and occluded images. The architecture of the proposed model FMR-CapsNet is shown in Figure 4-1. The proposed model includes four modules: Geometric feature extraction, relation-aware blendshapes, ResNet features, and Capsule neural network.

Facemesh MediaPipe has been used for feature extraction, which excels in facial emotion recognition due to its detailed 3D mesh representation, which captures subtle facial nuances more accurately than 2D methods. Its real-time performance and robustness to lighting, pose, and occlusions make it highly effective for diverse conditions. FaceMesh provides 468 facial landmarks, as shown in Figure 4-2, which are used to derive 52 blendshape scores. A distance matrix is then created by calculating the euclidean distance between each pair of scores within a single feature vector, resulting in a 52×52 matrix for each feature vector. This feature matrix is fed into a ResNet50 pre-trained model to obtain refined features for the next module. Refined features are fed to capsule neural network.

Because the convolutional neural network overlooks the spatial and interlayer relative features of the target, Hinton et al. [77] introduced the concept of capsules to store pose knowledge for targets and utilized a dynamic routing mechanism to convey knowledge between layers. Since capsules can capture the inter-feature relationship from different local regions, the FMR-CapsNet incorporates a capsule neural network enabling features to be directional and retrieve relative relationships between features. So, this section provides insight into each module of the proposed model.

4.3.1 Feature Extraction

FMR-CapsNet utilizes geometric features (facial landmarks) extracted using Facemesh Mediapipe.

Geometric Feature Extraction

The MediaPipe FaceMesh solution serves as the tool for detecting geometric features in the form of facial landmarks in this study. Developed by Google’s MediaPipe team, it employs machine learning to detect and monitor 478 facial landmarks, encompassing features like the eyes, eyebrows, nose, mouth, and jawline [80]. It takes a single image as an input to the model to extract facial landmarks. In the proposed method, blendshapes scores are used which used 146 landmarks of Face Landmarker solution. The output from the blendshape model consists of 52 blendshape scores, each expressed as a floating-point value between 0 and 1 as shown in Figure 4-3 (Grishchenko et al., 2022). The predicted blendshapes are detailed in Table 4.1. The blendshape weights for 100 input images are visualized in Figure 4-4.

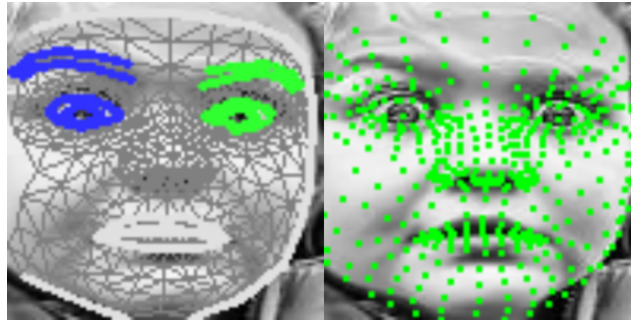


Figure 4-2: 478 Facial landmarks extracted using FaceMesh Mediapipe.

Table 4.1: Predicted blendshapes from LandMarker solution.

Blendshapes Categories			
browDownLeft	eyeLookInRight	mouthClose	mouthRollLower
browDownRight	eyeLookOutLeft	mouthDimpleLeft	mouthRollUpper
browInnerUp	eyeLookOutRight	mouthDimpleRight	mouthShrugLower
browOuterUpLeft	eyeLookUpLeft	mouthFrownLeft	mouthShrugUpper
browOuterUpRight	eyeLookUpRight	mouthFrownRight	mouthSmileLeft
cheekPuff	eyeSquintLeft	mouthFunnel	mouthSmileRight
cheekSquintLeft	eyeSquintRight	mouthLeft	mouthStretchLeft
cheekSquintRight	eyeWideLeft	mouthLowerDownLeft	mouthStretchRight
eyeBlinkLeft	eyeWideRight	mouthLowerDownRight	mouthUpperUpLeft
eyeBlinkRight	jawForward	mouthPressLeft	mouthUpperUpRight
eyeLookDownLeft	jawLeft	mouthPressRight	noseSneerLeft
eyeLookDownRight	jawOpen	mouthPucker	noseSneerRight
eyeLookInLeft	jawRight	mouthRight	tongueOut

Relation-aware Blendshapes

The extracted blendshapes score feature vector doesn’t provide the relative information between each blendshapes score. To compute the relation-aware blendshapes,

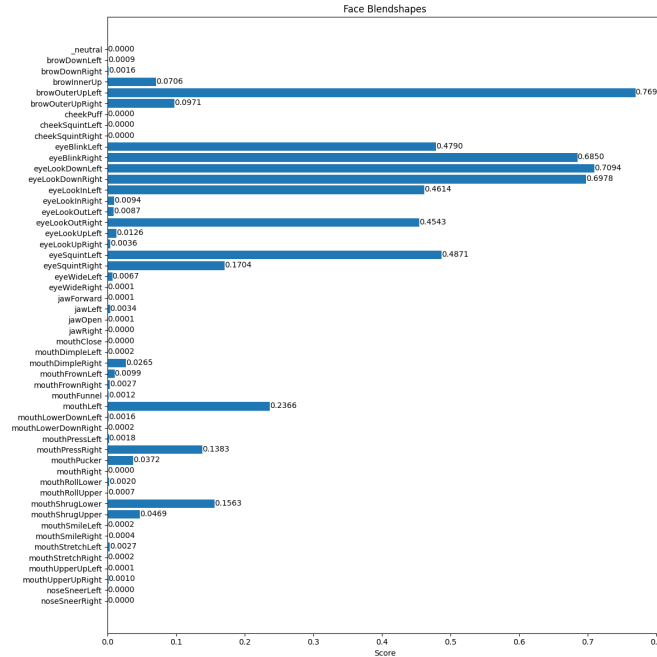


Figure 4-3: Facial blendshapes scores extracted using FaceMesh Mediapipe.

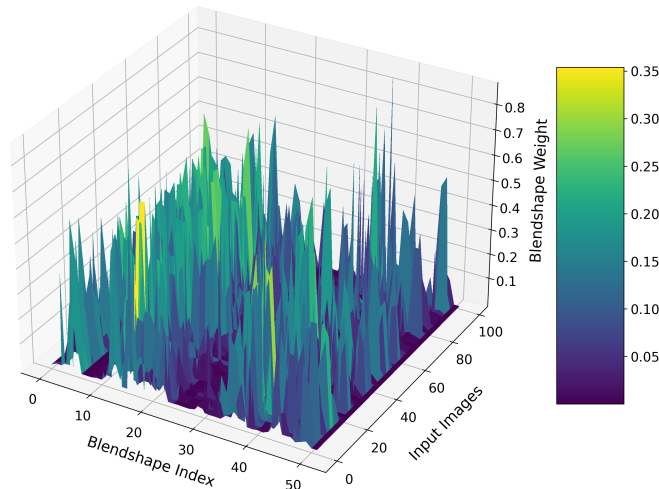


Figure 4-4: Visualization of blendshape weights of 100 random images of RAF-DB dataset.

euclidean distance is applied between each pair of blendshape scores of feature vector 1×52 to generate distance matrix 52×52 using Eq. 4.1. Euclidean distance provides a straightforward way to quantify the differences between blendshape scores. By calculating the distance between each pair of scores, we can measure how much one facial expression deviates from another, giving a numerical representation of the variation

in facial features. Calculating euclidean distances between blendshape scores helps capture these inter-feature relationships. For instance, a smile involves coordinated mouth, cheeks, and eye movements. The model can better understand the overall expression by understanding the distances between the blendshape scores corresponding to these features.

Randomly, 100 images are selected from RAF-DB dataset whose blendshape weights are extracted and their visualization is shown in Figure 4-4. The heatmaps shown in Figure 4-5 visualizes the euclidean distances between 52 facial blendshapes for RAF-DB and AffectNet datasets, where each cell represents the distance between two blendshapes. Darker colors represent small distances, whereas lighter colors denote large distances. The prominent yellow diagonal represents zero distance (each blendshape to itself). Horizontal and vertical yellow lines suggest distinct blendshapes, particularly around blendshapes 13, 16, and 17, in Figure 4-5 a) indicating these are unique or significantly different from others. Clusters of darker areas denote groups of similar blendshapes, useful for understanding and analyzing facial expression variations in the dataset.

$$D_i(52 \times 52) = \sqrt{(f_{i,x} - f_{i,y})^2}, \quad (4.1)$$

Refinement using ResNet50

ResNet [81] is renowned for its excellent performance across various computer vision tasks. Transfer learning techniques can leverage the weights of these pre-trained models for various computer vision tasks, even when resources like datasets and computing power are limited. ResNet50 has a deeper architecture compared to models like VGG16 [82] and AlexNet [83], allowing it to learn more complex features. ResNet’s introduction of residual connections mitigates the vanishing gradient problem, enabling the training of very deep networks. Thus, ResNet50’s innovative use of residual connections, balance of depth and parameter efficiency, and strong performance across various tasks contribute to its preference over other pre-trained models. In this work,

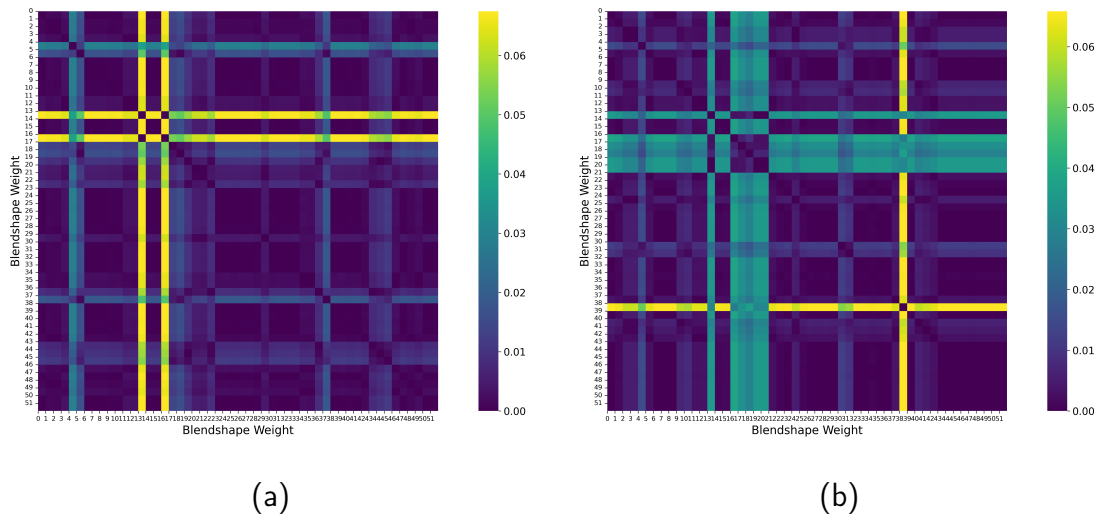


Figure 4-5: Heatmap of distance matrix generated using euclidean distance on (a) RAF-DB (b) AffectNet datasets.

ResNet50 has been utilized to implement transfer learning on the evaluated distance matrix. The design and depth of ResNet50 make it robust to variations in facial expressions, occlusions, and lighting conditions, which are common challenges in facial feature refinement.

In Figure 4-6, the initial layer has 64 filters with a 7×7 kernel size, followed by a max pooling layer of size 3×3 . The first group of layers (as represented by red color) includes three identical blocks. Similarly, the second group (represented by yellow color) contains four identical blocks, the third group (represented by green color) also has six identical blocks, and the fourth group (represented by cyan-blue color) comprises three identical blocks. The blue color curved lines depict the identity blocks, which signify the utilization of preceding layers in subsequent layers. This characteristic distinguishes ResNet50 and addresses challenges such as vanishing or exploding gradients.

4.3.2 Capsule Neural Network

Convolutional neural networks are unable to recognize object rotations and scaling variations within objects effectively. Additionally, the pooling operations in CNNs can

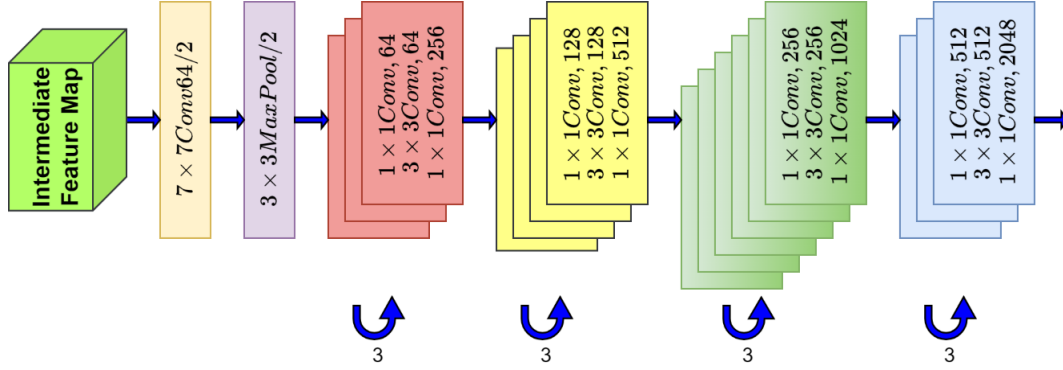


Figure 4-6: Architecture of ResNet50 pre-trained model

lead to a loss of spatial information. Capsule networks, inspired by the hierarchical structure of the human visual system, aim to address the shortcomings of traditional neural networks by introducing “capsules” as fundamental units of representation. These capsules not only encode the presence of features but also their instantiation parameters, such as pose and orientation, enabling the network to learn hierarchical relationships more effectively. Unlike the neurons in CNNs, capsules use output vectors that can capture directional information, allowing for more accurate image differentiation.

The architecture of the CapsNet is shown in Figure 4-1. The general architecture starts with a single convolutional layer but the proposed method utilizes two convolutional layers to detect local patterns from the intermediate features more effectively. The output of the convolutional layer is fed into primary capsules. Each primary capsule represents a higher-level feature detected by the convolutional capsules. Primary capsules are typically arranged spatially, capturing spatial hierarchies of features. Each primary capsule outputs a vector representing the instantiation parameters of a specific feature, such as pose, orientation, and scale. In CapsNet model, the PrimaryCaps layer consists of eight capsules, with each capsule containing sixteen-dimensional features. Additionally, the contribution ($\hat{u}_{j|i}$) of each capsule u_i in the PrimaryCaps layer to that of v_j DigitCaps was computed using Eq. (4.2).

$$\hat{u}_{j|i} = W_{ij} \cdot u_i \quad (4.2)$$

In the DigitCaps layer, there is a 16-dimensional capsule v_j allocated for each digit class (seven classes in this experiment). These capsules obtain input from all capsules in the PrimaryCaps layer using Eq. (4.3), Eq. (4.4), Eq. (4.5).

$$cap_{ij} = \frac{\exp(b_{ij})}{\sum_l \exp(b_{il})} \quad (4.3)$$

$$c_j = \sum_l cap_{ij} \hat{u}_{j|i} \quad (4.4)$$

$$v_j = \frac{\|c_j\|^2}{1 + \|c_j\|^2} \frac{c_j}{\|c_j\|} \quad (4.5)$$

At last, the margin loss is computed for each digit capsule to classify the facial expressions using Eq. (4.6) where $Y_l = 1$ if there is relation 1, $r^+ = 0.9$, $r^- = 0.1$, and $\lambda = 0.5$.

$$D_l = Y_l \max(0, r^+ - \|v_l\|)^2 + \lambda(1 - Y_l) \max(0, \|v_l\| - r^-)^2 \quad (4.6)$$

The algorithm outlines the procedure followed in the proposed model for the training of FMR-CapsNet, designed to recognize facial expressions of in-the-wild datasets. It starts by inputting the dataset and extracting features using Mediapipe’s Facemesh, creating a 1×52 feature map for each input image. For each input image, it calculates a 52×52 distance matrix by applying euclidean distance on a feature matrix of size 1×52 . These features are then refined with a pre-trained ResNet50 model and passed through two Conv2D layers. Next, it uses a capsule network’s routing procedure to determine how much each feature contributes to the final output, updating the connections between layers iteratively. This involves calculating weighted sums and applying a squash function to finalize the output vectors. After the routing iterations, the algorithm calculates the margin loss for each class to ensure accurate classification. Finally, it outputs the class probabilities, indicating the likelihood of each class for the given input.

Algorithm 4.1 Algorithm of the proposed model.**Input:**

Dataset $\in \{X_i, Y_i\}_{i=1}^n$ where n is number of images in dataset
 Model parameters θ
 Number of routings \mathbb{R}

Output:

Trained FMR-CapsNet model to recognize the facial expressions of in-the-wild datasets

- 1: Input the dataset, where Dataset $\mathbb{D}_i \in \{X_i, Y_i\}_{i=1}^n$
- 2: $Featuremap_{(1 \times 52)} = \text{Facemesh Mediapipe(D)}$
- 3: **for** $i = 1$ **to** n **do**
- 4: Distance matrix $D_{i(52 \times 52)} = \sqrt{(f_{i,x} - f_{i,y})^2}$,
- 5: **end for**
- 6: Refine the intermediate feature using pretrained Resnet50 model
- 7: Apply Conv2D layer $\times 2$
- 8: Compute the contribution for all capsule i in primary layer (shown in Eq. 4.2)
- 9: **procedure** ROUTINGS($\hat{u}_{j|i}, \mathbb{R}$)
- 10: For all capsule i in primary layer (input capsules) and j in DigitCaps layer (output capsules),
 $b_{ij} \leftarrow 0$
- 11: **for** $i = 1$ **to** \mathbb{R} **do**
- 12: Fo input capsules i $cap_i = \text{softmax}(b_i)$
- 13: For output capsules j , $c_j = \sum_l cap_{i;l} \hat{u}_{j|i}$
- 14: For output capsules j , $v_j = \text{squash}(c_j)$
- 15: For input capsules i and output capsules j , $b_{ij} = b_{ij} + \hat{u}_{j|i} \dot{v}_j$
- 16: **return** v_j
- 17: **end for**
- 18: **end procedure**
- 19: Calculate margin loss D_l for each class l
- 20: Output vector with class probabilities

4.4 Experimental Results

Experiments are performed on 64-bit Windows 11, Intel Core i7 processor with 16GB RAM size, 8GB NVIDIA TITAN RTX graphics card. The summary of hyperparameters used in FMR-CapsNet is shown in Table 4.2. The margin loss is used to produce high probabilities for the correct class and low probabilities for all other classes. This loss function helps to ensure that the capsules for the correct class have high activations while suppressing the activations of capsules for other classes. AdamW optimization algorithm combines the benefits of the Adam optimizer and weight decay. It is widely used in training neural networks due to its adaptive learning rate and momentum properties. A batch size of 16 is used and hence, 16 data samples are

processed in each iteration. The learning rate of 0.0001 is used which dictates how quickly the model's parameters are updated throughout the training process. It controls the step size in the parameter space. Learning rate decay is the hyper parameter that controls the rate at which the learning rate decreases over time during training. A decay factor of 0.9 means that the learning rate is multiplied by 0.9 after each epoch or a certain number of iterations. Number of routings is the hyper-parameter which is specific to Capsule Networks and determines the number of iterations (or "routings") performed in the dynamic routing algorithm during inference. It affects how capsules communicate and reach an agreement about the instantiation parameters of higher-level capsules.

The proposed model is evaluated on two in-the-wild facial expression datasets i.e. AffectNet and RAF-DB. The images for these datasets are collected from unconstrained environment, exposed to varying levels of lighting, head-orientation, occlusion, etc. Some samples of different facial expressions of AffectNet and RAF-DB are shown in Figure 4-7.

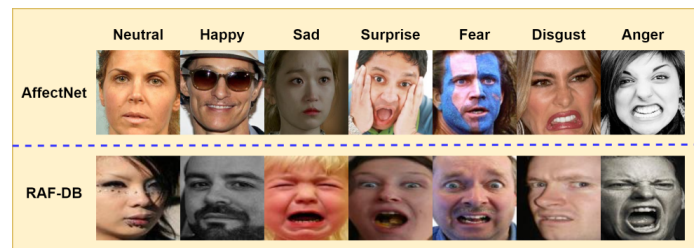


Figure 4-7: Samples of facial expressions from AffectNet and RAF-DB dataset.

RAF-DB

The RAF-DB dataset [72] contains seven basic facial expressions and twelve compound facial expressions, totaling approximately 30,000 facial images. It is considered a large-scale database as it is collected from an unconstrained environment with significant variations in subjects' attributes, lighting conditions, occlusions, and more. For the experiment, facial images with 7 categories of facial expressions are used, comprising around 12,271 images in the training set and 3,068 images in the test set.

AffectNet

The AffectNet dataset [71], which provides over 1 million in-the-wild facial images, is considered one of the largest in the research field of facial affective computing. Approximately 440,000 facial images have been manually annotated with the category or intensity of facial expression. In the experiment, 285,718 sample images with 7 facial expression categories were used. The AffectNet dataset does not include a separate testing set, so the training dataset is split into training and testing sets using a ratio of 95%:5%, resulting in 271,621 images (95%) utilized for model training and 14,097 images (5%) for testing purposes. The experiment was conducted on the testing set to evaluate the model.

The distribution of RAF-DB and AffectNet datasets is shown in Table 4.3 which shows the number of samples for each class in the training and testing datasets for both RAF-DB and AffectNet.

Table 4.2: Summary of hyperparameters used in FMR-CapsNet.

Hyperparameter	Value
Batch Size	16
Learning rate	0.0001
Learning rate decay factor	0.9
Number of routings	3
Loss function	Margin loss
Optimizer	AdamW

Table 4.3: Distribution of samples in training and testing subset of RAF-DB and AffectNet dataset.

Class	Emotion	RAF-DB		AffectNet	
		Training	Testing	Training	Testing
0	Neutral	2524	680	71156	3718
1	Happy	4772	1185	127727	6688
2	Sad	1982	478	24198	1261
3	Surprise	1290	329	13389	701
4	Fear	281	74	6062	316
5	Disgust	717	160	3615	188
6	Anger	705	162	23657	1225

4.4.1 Experimental results on RAF-DB dataset

The RAF-DB dataset contains 12,271 images for the training subset and 3,068 for the testing set. RAF-DB dataset is an in-the-wild dataset that presents several challenges, including diverse lighting conditions, varying angles, and occlusions such as glasses or hats, which complicate accurate facial recognition. FMR-CapsNet has demonstrated superior performance, achieving an accuracy of 97.01%. To address the imbalance in the RAF-DB dataset, class weights are implemented to mitigate biases towards the majority class. The general formula to calculate the class weight for each given class is shown in Eq. 4.7.

$$W_i = N_i / (C * N_i), \quad (4.7)$$

where W_i denotes weight for i^{th} class, N_i and C represents number of samples in i^{th} class and number of classes respectively. This technique assigns higher weights to minority classes, enhancing their class-wise accuracy. The specific weights assigned to each class in the experiment can be referenced in Table 4.4. The confusion metric evaluated on RAF-DB dataset using FMR-CapsNet is shown in Figure 4-8(a), the model is confused between sad and disgust, disgust and anger. Whereas, ROC curve depicts the relation between TPR (True Positive Rate) and (FPR) False Positive Rate. Figure 4-9(a) shows the ROC curve on RAF-DB dataset (different color for different emotion class).

4.4.2 Experimental results on AffectNet dataset

The AffectNet dataset does not include a separate testing set, so the training dataset is split into training and testing sets using a ratio of 95%:5%, resulting in a training dataset with 271,621 images and a testing dataset with 14,097 images. AffectNet is a large and complex in-the-wild dataset that requires substantial computational resources for implementation. To mitigate this, facial blendshapes are extracted using the Facemesh model and passed to the model as NumPy arrays, reducing resource requirements and improving classification accuracy. Additionally, since AffectNet is

an imbalanced dataset, class weights are incorporated to penalize the model’s misclassification rate. After multiple experiments, the optimized set of class weights was adopted, as shown in Table 4.4. FMR-CapsNet achieves 71.12% accuracy on the AffectNet dataset, surpassing state-of-the-art methods as illustrated in Table 4.5.

Figure 4-9 shows the ROC_AUC scores for each class of RAF-DB and AffectNet datasets. The detailed comparison of FMR-CapsNet with state-of-the-art methods regarding class-wise accuracies is shown in Table 4.6. The results indicate that FMR-CapsNet achieves the highest class-wise accuracies for six classes in the RAF-DB dataset and one class in AffectNet. The remaining class-wise accuracies are also comparable to those of other methods.

Dataset	Class weight of each emotion						
	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
RAF-DB	0.6945	0.3673	0.8844	1.3589	6.2384	2.44909	2.4860
AffectNet	0.0226	0.0125	0.0864	0.3201	0.3653	0.6449	0.0880

Table 4.4: Class weights assigned to each class of RAF-DB and AffectNet dataset.

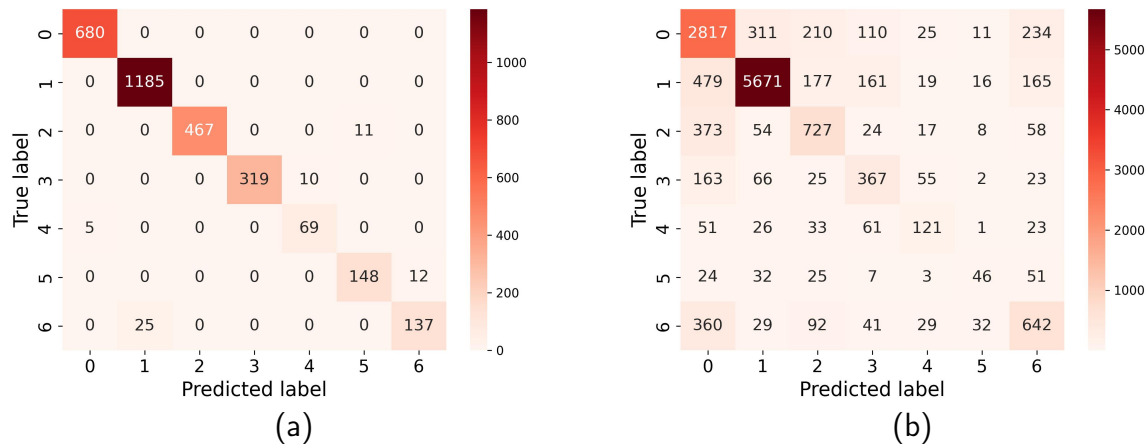


Figure 4-8: Confusion metric evaluated on (a) RAF-DB (b) AffectNet Dataset using FMR-CapsNet.

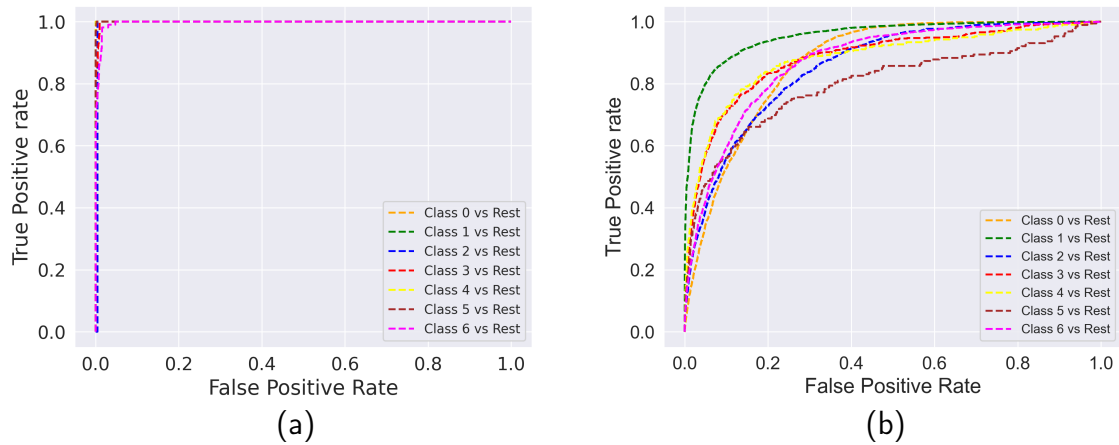


Figure 4-9: ROC curve evaluated on (a) RAF-DB (b) AffectNet Dataset using FMR-CapsNet.

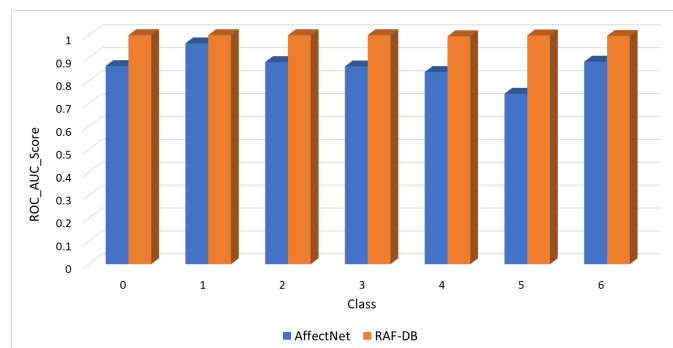


Figure 4-10: ROC_AUC score of different classes of AffectNet and RAF-DB dataset.

4.4.3 Ablation Study

The model's performance is evaluated by including various modules and results are reported in Table 4.7. A quantitative assessment is performed to validate the contribution of each module within the model. Firstly, CapsNet is evaluated alone on RAF-DB and AffectNet datasets which classify the facial expressions with an accuracy of 75.34% and 65.25% respectively. Transfer learning is applied with pre-trained weights of ResNet50 for feature extraction of input images with CapsNet which improved the model's performance by 7.71% and 5.24% for RAF-DB and AffectNet datasets respectively, which is comparable with state-of-the-art methods [31] [84] [88] [91]. CapsNet with ResNet50 improves the accuracy but was not able to perform well on occluded

Table 4.5: Performance comparison of different SOTA methods with FMR-CapsNet on RAF-DB and AffectNet datasets.

Authors	Model	Year	Input	Param(M)	Accuracy	
					RAF-DB	AffectNet
Zhang et al. [31]	IE-DBN	2020	62×64	-	84.75%	-
Wang et al. [84]	RAN-ResNet18	2020	224×224	11.19	86.90%	59.5%
Wang et al. [85]	SCN	2020	224×224	-	87.03%	-
Xia et al. [86]	ReCNN	2021	224×224	-	87.06%	-
Saurav et al. [87]	EmNet	2021	40×40	4.80	87.16%	-
Farzaneh et al. [88]	DACL	2021	224×224	11.18	87.78%	65.20%
Zhang et al. [89]	RUL	2021	224×224	-	88.98%	-
Su et al. [90]	LSGB	2022	224×224	2.5	-	58.68%
Liu et al. [91]	AMP-Net	2022	224×224	105.67	89.25%	64.54%
Wan et al. [92]	GFE2N	2022	224×224	16	82.17%	51.81%
Ryumina et al. [93]	CNN-LSTM	2022	300×300	-	-	66.4%
Gao et al. [94]	SSA-ICL-ResNet18	2023	224×224	9.02	89.44%	65.78%
Yu et al. [75]	ARM	2023	224×224	11.18	90.42%	-
Zhang et al. [95]	CF-DAN	2024	224×224	16.4	92.78%	65.58%
Zhang et al. [96]	GSDNet	2024	224×224	-	90.91%	66.11%
Ours	FMR-CapsNet	2024	100×100	34.5	97.01%	71.12%

and pose-variant images. To handle the problem of occlusion and pose-variations in facial images, the FaceMesh module is used to extract the geometric features from the facial images which capture the features from pose-variant and occluded images as well, and hence improves the classification accuracy by 16.51% and 4.86% for RAF-DB and AffectNet datasets respectively. As shown in Table 4.7, it is clear that when ResNet50 and the FaceMesh module were individually combined with CapsNet, both improved accuracy. Therefore, all three are combined to provide refined relation-aware geometric features of facial images, which greatly enhanced the accuracy and reached 97.01% for RAF-DB and 71.12% AffectNet dataset.

The model was evaluated with different numbers of routings, and 3 was found to be the optimal value for the number of routings in the capsule network module of FMR-CapsNet. The comparison of the number of trainable parameters for each configuration of FMR-CapsNet modules is shown in Figure 4-11b). The figure demonstrates that utilizing ResNet alone in the FMR-CapsNet model significantly increases the number of trainable parameters, indicating that ResNet contributes substantially to the parameter count in the proposed method. In contrast, using FaceMesh alone

Table 4.6: Detailed comparison of FMR-CapsNet with state-of-the-art methods with respect to class-wise accuracy obtained on RAF-DB and AffectNet datasets; * Highest class-wise accuracies achieved for RAF-DB and AffectNet datasets are highlighted in red and blue color respectively.

Model	Dataset	Accuracy of each emotion(%)						
		Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
IE-DBN [31]		83.8	93.8	83.1	82.4	63.5	53.1	73.5
ReCNN [86]	RAF-DB	86.62	94.35	85.36	86.93	64.86	58.75	79.01
EmNet [87]		89.0	95.0	87.0	82.0	51.0	56.0	80.0
AMP-Net [91]		89.0	96.0	87.0	86.0	65.0	65.0	82.0
RAN-ResNet18 [84]		26.0	79.0	77.0	51.0	58.0	30.0	56.0
LSGB [90]	AffectNet	38.0	66.0	48.0	52.0	59.0	48.0	54.0
CNN-LSTM [93]		68.0	88.0	64.2	63.2	60.0	62.0	59.2
SSA-ICL-ResNet18 [94]		56.8	62.0	48.0	72.0	83.8	35.6	54.6
DACL [88]		87.06	93.92	84.31	86.93	66.22	65.0	79.63
	AffectNet	64.20	87.80	68.20	61.0	58.0	61.0	56.20
	RAF-DB	82.65	94.77	76.57	72.95	63.51	46.25	67.28
GFE2N [92]	AffectNet	34.37	72.78	51.78	57.53	49.54	61.49	42.36
	RAF-DB	87.52	95.49	90.0	91.25	77.94	78.87	89.93
GSDNet [96]	AffectNet	63.29	62.45	67.88	55.95	79.64	64.4	71.35
	RAF-DB	100	100	96.79	95.71	89.58	90.48	76.19
FMR-CapsNet	AffectNet	75.77	84.80	57.66	52.41	38.24	24.21	52.41

with CapsNet results in the lowest number of trainable parameters, highlighting its minimal contribution in comparison.

Table 4.7: Evaluation of the effect of different modules used in FMR-CapsNet on RAF-DB and AffectNet datasets.

Dataset	FM	ResNet50	CapsNet	Accuracy	AUC	MCC	R	$F1_w$
RAF-DB	×	×	✓	75.34%	94.37	0.60	0.59	0.632
	×	✓	✓	83.05%	95.56	0.53	0.58	0.57
	✓	×	✓	91.85%	99.55	0.89	0.90	0.86
	✓	✓	✓	97.01%	99.81	0.93	0.92	0.93
AffectNet	×	×	✓	65.25	91.43	0.43	0.43	0.47
	×	✓	✓	70.39	83.27	0.46	0.48	0.52
	✓	×	✓	70.11	94.17	0.38	0.39	0.42
	✓	✓	✓	71.12	92.55	0.51	0.55	0.57

4.4.4 Comparative study of FMR-CapsNet with state-of-the-art methods

In recent years, several state-of-the-art methods have been proposed for improving model accuracy. But the performance of our model FMR-CapsNet, significantly sur-

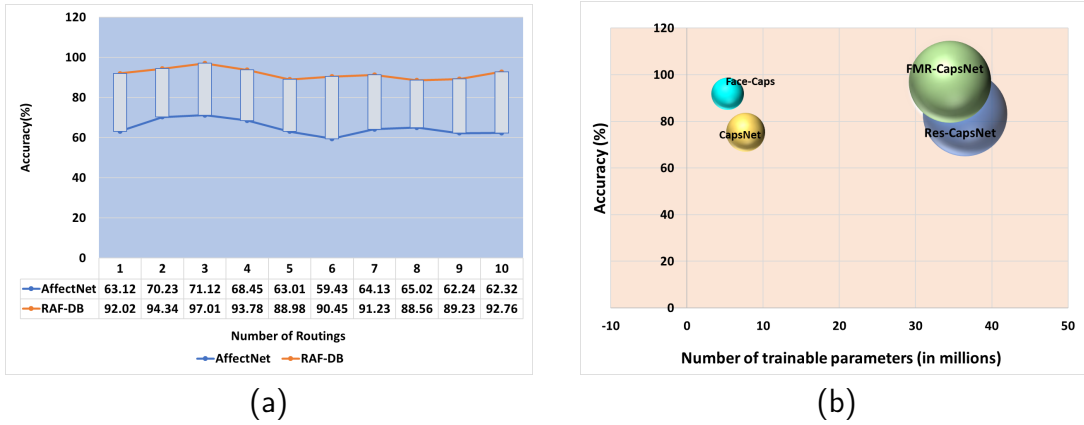


Figure 4-11: (a) Influence of number of dynamic routings in FMR-CapsNet on AffectNet and RAF-DB dataset (b) Variation in number of trainable parameters on adding different modules to FMR-CapsNet *CapsNet: Capsule neural network, Face-Caps: FaceMesh + CapsNet, Res-CapsNet: ResNet50 + CapsNet, FMR-CapsNet: FaceMesh + ResNet50 + CapsNet.

passes that of state-of-the-art models on RAF-DB and AffectNet datasets as shown in Table 4.5.

For RAF-DB dataset, IE-DBN [31] model was introduced, achieving an accuracy of 84.75%, RAN-ResNet18 [84] with 11.19M parameters, reaching an accuracy of 86.90%. Another model by Wang et al., SCN [85], reported an accuracy of 87.03%, closely followed by ReCNN [86] at 87.06%. In 2021, EmNet [87] was proposed with 4.80M parameters, achieved 87.16%, and DACL [88], with 11.18M parameters, reached 87.78%. Zhang et al.’s RUL [89] led the 2021 models with an accuracy of 88.98%. AMP-Net [91], although with a significant 105.67M parameters, achieved 89.25%, and GFE2N [92], having 16 million parameters, showed a lower performance with 82.17%. In 2023, SSA-ICL-ResNet18 [94] proposed with 9.02M parameters, reached an accuracy of 89.44%, and Yu et al.’s ARM [75], with 11.18M parameters, achieved 90.42%. The models in 2024 saw significant improvements, with CF-DAN [95] having 16.4M parameters, achieving 92.78%, and GSDNet [96] reaching 90.91%. Our proposed model, FMR-CapsNet, significantly outperforms these state-of-the-art models. With 34.5 million parameters, it achieves an outstanding accuracy of 97.01%, demonstrating substantial improvements over the existing state-of-the-art

methods.

On AffectNet dataset, with an accuracy of 71.12%, FMR-CapsNet outperforms the next best model, CNN-LSTM [93], which achieved an accuracy of 66.4%. Additionally, our model exhibits superior efficiency with a 34.5M model parameters. Other models, such as SSA-ICL-ResNet18 [94] and CF-DAN [95], achieve accuracies of 65.78% and 65.58% respectively on AffectNet dataset. AMP-Net [91] and GSD-Net [96], despite being recent models from 2022 and 2024, only achieve accuracies of 64.54% and 66.11% respectively. AMP-Net [91] is a complex model with 105.67M parameters and achieves 64.54% accuracy, which is 6.58% less than the accuracy the proposed model achieved. The GFE2N [92] model demonstrates a notably lower accuracy of 51.81%. Our FMR-CapsNet model not only sets a new benchmark for accuracy but also maintains a balanced complexity, proving its superiority in both performance and efficiency. Blurred images slightly contribute to the mis-classification rate of FMR-CapsNet on RAF-DB and AffectNet datasets, as FaceMesh Mediapipe fails to capture their geometric features, leading to incorrect classification of these images.

4.5 Conclusion

This chapter proposes a novel robust in-the-wild facial emotion recognition method utilizing relation-aware geometric features. FMR-CapsNet effectively addresses challenges posed by occluded or pose-variant images, achieving accurate results. Experimental results demonstrate that our proposed model outperforms state-of-the-art methods in facial emotion recognition tasks on RAF-DB and AffectNet datasets. Furthermore, ablation studies have elucidated how relation-aware geometric features significantly enhance the model’s classification performance.

Chapter 5

Design and development of robust and generic framework for Compound Emotion Recognition

Nidhi, and Bindu Verma. “A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition.” *Applied Intelligence* (2024): 1-17. (SCIE Indexed, IF: 3.4) DOI: 10.1007/s10489-024-05598-5 (Published)

5.1 Introduction

This chapter presents an LSwin-CBAM model for the classification of compound emotions. To address the problem of the imbalanced dataset, the proposed model exploits the CutMix augmentation technique for data augmentation. The proposed method discussed in the previous chapter is primarily designed for basic emotions. Basic emotions such as happiness and sadness do not fully encompass the complexity of human emotions, as real-life emotional states frequently involve a blend of multiple emotions, like anger intertwined with sadness or joy combined with surprise. By targeting compound emotions, the model can better capture the intricacies of human

emotional experiences. So, this chapter is extended to classify the compound emotions which are complex versions and composite of basic emotions. LSwin-CBAM also incorporates the CBAM attention mechanism to emphasize the relevant features in an image and Swin Transformer with fewer Swin Transformer blocks which leads to less computational complexity in terms of trainable parameters and improves the overall classification accuracy as well.

Authors usually focus on seven basic emotions: “Surprise”, “Anger”, “Happiness”, “Contempt”, “Fear”, and “Disgust”. However, there is a need to examine more detailed and precise facial expressions. Some authors tried to find detailed facial expressions because of recent developments in the area of compound emotions [97,98]. To examine a person’s emotional state in greater detail using facial emotion expression analysis, compound emotion categories have been proposed [3]. Such studies contribute in the context of compound emotion recognition which comprehends and identifies fine-grained facial expressions.

Transformers [67] are greatly used in Natural Language Processing (NLP) tasks. Research efforts to adapt transformers for vision tasks have been driven by the tremendous achievement of transformers in NLP. Due to its original transformer architecture and promising performance in vision tasks, Vision Transformers (ViTs) [67] received a lot of attention from AI researchers. Apart from its promising performance, there are a few limitations encountered in ViTs. The computational complexity of ViTs is quadratic to image size, which is why it particularly faces difficulties with high-quality input images. Moreover, the variable scale visual features in vision tasks make it unsuitable as it processes the fixed scale tokens.

Liu et al. [99] developed the Swin Transformer which is possibly the most intriguing concept of research following the original ViT. Swin Transformer addressed the limitations of ViT by introducing these key concepts: hierarchical feature maps and shifted windows. The key idea behind the Swin Transformer is to divide the input image into non-overlapping patches and organize them hierarchically. Unlike traditional vision transformers that operate on fixed-size patches, the Swin Transformer employs shifted windows, which means that each patch is constructed by shifting a

fixed-size window across the image. This allows the model to have a larger receptive field without increasing the computational complexity significantly. It outperforms ViTs in image classification by identifying fine-grain features in an image where each pixel is labeled as in semantic segmentation. The deep Swin Transformer model is modified by including a single stage only with additional modules which improves the classification accuracy with minimal computational overhead. The adoption of lightweight machine learning models offers several significant benefits for real-time applications. Lightweight models are designed to have fewer parameters and require less computational power compared to heavy models with more computation. Due to their reduced complexity, lightweight models often have faster inference times, allowing for real-time processing of facial expressions. Training lightweight models typically requires less time and resources compared to training larger models. This can accelerate the development and deployment process, allowing for quicker adaptation to changing requirements or datasets.

Next, the absence of significant, publicly available labeled datasets for compound emotions is one of the major barriers to advancing research on the automatic recognition of compound emotions. To overcome the above-mentioned problem, the CutMix data augmentation technique has been used in the model which increases the training set and solves the problem of information loss and inefficient behavior of other regional dropout methods. The CutMix technique picks random patches from the training images and ground truth labels are combined in the ratio corresponding to the area of patches in the images. The neural network's ability to focus on specific regions, eliminate irrelevant information, and effectively extract essential features from images is made possible by using attention mechanisms. In the proposed model, the convolutional block attention module (CBAM) attention mechanism [66] has been used which sequentially combines the channel attention module to exploit the inter-channel relationship by giving varying significance to the channels of feature maps and spatial attention module to determine which parts of a feature map are more imperative. This chapter aims to design a lightweight compound emotion recognition model named LSwin-CBAM that is based on a streamlined Swin Transformer exploit-

ing CutMix augmentation technique and CBAM attention mechanism to improve the classification rate of compound emotions. An analysis of the computational efficiency and model complexity of LSwIn-CBAM is conducted, providing insights into its practical applicability. The proposed model has experimented on compound emotion datasets such as RAF-DB, EmotioNet and outperforms state-of-the-art methods.

5.2 Literature Survey

Humans can easily recognize emotions accurately, but making a fully automated machine to recognize facial expressions is still challenging. If this capability of emotion recognition can be empowered in robots or computers then robotic applications can be developed to understand the emotions of a human with a limited understanding of the relationship between emotional expressions and body movements [9]. The facial expression recognition method is developed by deforming a few control points of the FACS-based facial motion proposed by Ekman [100]. Many studies exist [101–104] which focused on seven basic expressions but to enrich more details of human emotions, compound emotions came into the picture. Guo et al. [105] compared the top winner’s methods (used CNN) from FG 2017 workshop on an iCV-MEFED dataset for the recognition of compound emotion and observed that Ist model given by [106] has performed better with less misclassification rate (0.793 for validation set and 0.802 for test set).

Pons et al. [107] proposed a Selective Joint Multitask approach (SJMT) which considers selective loss function (sigmoid cross-entropy) for each dataset. This approach [107] is beneficial while dealing with different datasets containing unlabeled images. The proposed loss function is incorporated with ResNet-50 and VGG16 model and found more effective when applied to datasets collected in an unconstrained environment (Oulu-Casis, SFEW) and compound emotion recognition (EmotioNet) as compared with the individual ones.

Mu et al. [50] proposed a model based on visual transformers with feature fusion which is tested on in-the-wild expression datasets like RAF-DB, AffectNet, and

FERPlus. Lian et al. [108] assessed GPT-4 with vision for Generalized Emotion Recognition (GER) tasks including tweet sentiment analysis, facial emotion recognition, visual sentiment analysis, micro-expression recognition, multimodal emotion recognition, and dynamic emotion recognition. Zhu et al. [109] proposed a unified framework based on a multi-branch vision transformer for multi-task facial emotion recognition and mask-wearing classification. Dong et al. [110] introduced a novel loss function called bi-center loss, which is an extension of center loss. This new function is designed to capture compound emotion features by utilizing the centers of basic emotions. Liu et al. [99] developed a Swin Transformer that merges image patches in deeper layers to produce hierarchical feature maps and it computes self-attention operation within the local window which leads to linear computational complexity. Zhao et al. [111] proposed a Swin Transformer-based model that accepts multimodality inputs to recognize the micro-expressions. Xue et al. [112] proposed a Coarse-to-fine Cascaded Network (CFC) with Smooth Predicting (SP) and improved the classification rate by extracting both unique and universal features and by ensembling Swin Transformer with ResNet for the training of the model. The proposed LSwin-CBAM framework is based on a lightweight Swin Transformer which consists of only two consecutive Swin Transformer blocks with CBAM attention mechanism which improves the recognition capability of the network with very less computational overhead.

5.3 Proposed Work

Compound emotions are the composite of basic emotions that gives a detailed understanding of a person’s feeling. Transformers performed tremendously well for natural language processing (NLP) tasks. So researchers [113] developed a variant named “Vision Transformer” based on transformer architecture which can be applied for vision tasks. However, there are some limitations as it works on a constant scale. In contrast, the scale of visual entities in the given images can have large variations and computational complexity is quadratically increasing with image size. So, these above-mentioned limitations forced researchers [99] to develop the Swin Transformer

model to extract dense information from the given images. Swin Transformer works on each pixel and forms a hierarchical feature map that initiates with small-sized patches and progressively merges with neighboring patches as moves down in deeper transformer layers.

In our work, we proposed a transformer-based model for compound emotion recognition which used a streamlined Swin Transformer with a CBAM attention module. Swin Transformer is taken as the base model with some architectural modifications. The general architecture of the Swin transformer (ST) [99] has 4 stages where the ST block is preceded by a linear embedding layer in the first stage and a patch merging layer in the last three stages which leads to a heavy deep model that increases computational overhead with 6M parameters on experimentation. To optimize the base model, a reduced number of stages (single stage) are included, resulting in minimal computational overhead with 0.15M parameters. The adoption of a lightweight model is suitable for real-time applications. Moreover, the CBAM (Convolutional Block Attention Module) attention mechanism is integrated, introducing a trade-off between model accuracy and parameter count in the proposed model. Additionally, the utilization of CutMix augmentation enhances the classification performance, surpassing that of the base Swin model while utilizing fewer parameters (0.15M). By carefully integrating these components, the LSwin-CBAM model achieves enhanced performance in recognizing compound emotions, surpassing the capabilities of individual methods.

The proposed architecture of LSwin-CBAM is shown in Figure 5-1 that accepts the images as input which are forwarded to the CutMix augmentation module because imbalance data is a very common problem in facial expressions datasets which affects the performance of the model. The augmented dataset is fed to the Conv2D layer to produce a feature map which is given as an intermediate map to the CBAM attention module to enhance the feature representation of relevant features. CBAM is followed by a patch partitioning module to produce patches that are flattened to form a feature vector which will go through a linear embedding layer to convert it to an arbitrary dimension C . This intermediate feature map is further fed to the streamlined Swin Transformer block (consisting of a single stage only) and followed by a patch merging

operation to merge two adjacent patches for downsampling of the feature map. After patch merging, global average pooling and dense layer are applied to produce the classified emotion as an output. Below we have discussed each steps of proposed work.

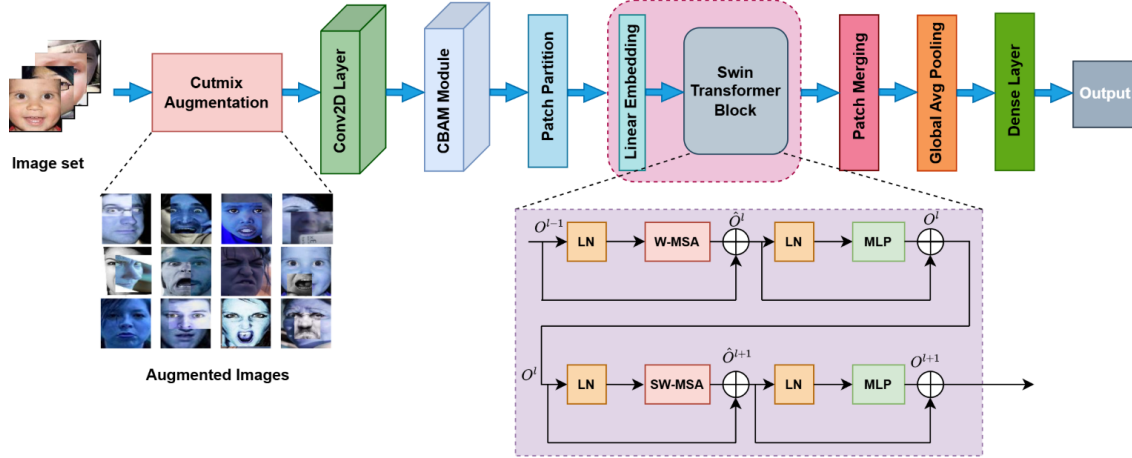


Figure 5-1: Proposed model architecture integrating CutMix augmentation, CBAM attention, and lightweight Swin Transformer module for Compound Emotion Recognition.

5.3.1 Input Data and Augmentation

Input to the proposed model is a compound emotion RGB images. CutMix augmentation is applied to increase the training samples. Data augmentation is considered as one of the important steps of data pre-processing as it greatly enhances the variety of data accessible for training our models, all without the need to gather additional data samples. Basic techniques such as flipping, random cropping, and random rotation are frequently employed to train extensive models, particularly effective for smaller datasets and simpler problems. However, significant shifts and corruptions in data can occur when faced with the complexities of real-world scenarios. So, there are advanced augmentation techniques available like Cutout [114], Mixup [115], and CutMix [116]. In Cutout augmentation [114], a square-shaped region is eliminated from the image and substituted with either black or grey pixels, or filled with Gaussian noise. The eliminated regions are typically zeroed out or populated by

random noise [114] which drastically lowers the extent of informative pixels in given images. With Mixup augmentation [115], a linear interpolation is employed between two randomly selected images. Nevertheless, the resulting images exhibit a degree of unnaturalness that can confuse the model. So, to utilize the deleted regions, Yun et al. [116] proposed the CutMix augmentation strategy. It uses a patch from a different image to restore the missing regions rather than just eliminating pixels as shown in Figure 5-2. Additionally, the ground truth labels are combined in accordance with the total number of pixels in the merged images. The additional patches make the model recognize the object from a partial context, which improves the localization capability even further.

CutMix augmentation technique [116] is applied on the given training images which produce the output images shown in Figure 5-2. The training image is indicated by $x \in I^{W \times H \times C}$ where H, W, and C represent the height, width, and channels of an image respectively. CutMix aims to produce a new training instance (\tilde{x}, \tilde{y}) by mixing any two given training samples (x_i, y_i) and (x_j, y_j) . The model is trained using the original loss function of generated training sample (\tilde{x}, \tilde{y}) . The combining or mixing operations are defined in Eq. (5.1) where $B_m \in (0, 1)^{W \times H}$ represents binary mask which suggests the pixels to keep and drop from two images.

$$\begin{aligned}\tilde{x} &= B_m \odot x_i + (1 - B_m) \odot x_j, \\ \tilde{y} &= \delta y_i + (1 - \delta) y_j,\end{aligned}\tag{5.1}$$

The CutMix method will produce a CutMix-ed augmented sample (\tilde{x}, \tilde{y}) by combining any two randomly picked training samples in each iteration as shown in Figure 5-2.

5.3.2 CBAM Attention mechanism

The origination of simulating human attention in computer vision is to improve the model's performance for image processing tasks with minimized computational complexity. Attention mechanisms concentrate on particular sections of the images instead of the complete picture and improve the interest representation, which has



Figure 5-2: Samples of Images Post-CutMix Augmentation.

gained popularity in deep learning. The convolution function extracts relevant features by incorporating spatial and cross-channel information together. The Convolutional Block Attention Module (CBAM) module sequentially infers attention maps along the two distinct dimensions of channel and spatial from an intermediate feature map and then multiplies the attention maps by the input feature map to refine the adaptive feature. As a universal and lightweight module, it can be seamlessly integrated with any CNN-based architecture with minimal overhead.

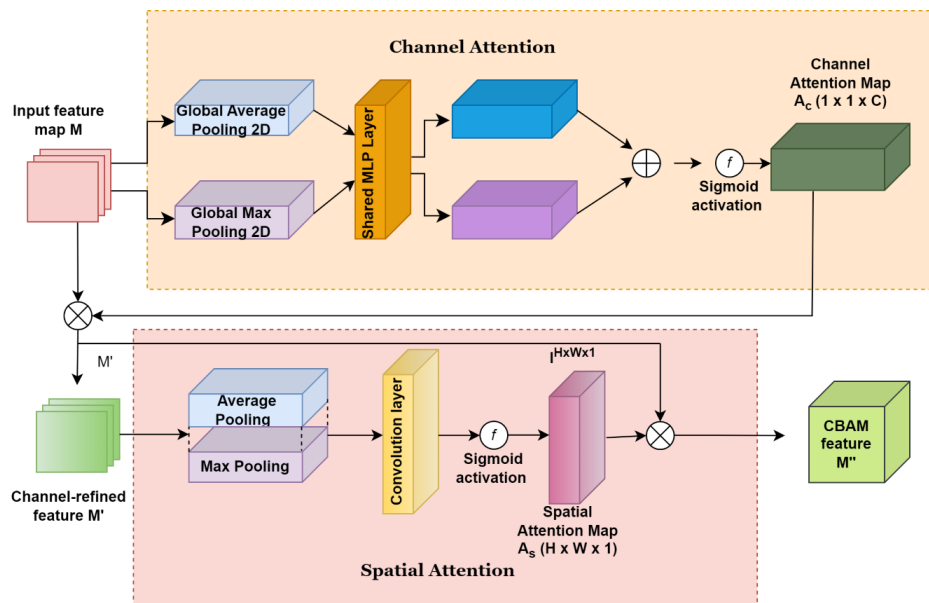


Figure 5-3: Architecture of Convolutional Block Attention Module (CBAM).

CBAM accepts an intermediate feature map represented by $M \in I^{C \times H \times W}$ and

applies a single-dimensional channel attention map $A_c \in I^{C \times 1 \times 1}$ and two-dimensional spatial attention map $A_s \in I^{1 \times H \times W}$ in a sequential manner. The attention operation can be formulated in Eq. (5.2). In element-wise multiplication, the channel and spatial attention values are copied along spatial and channel dimensions respectively, and give M'' as the final output. The visualization of learned attention weights using CBAM attention mechanism is shown in Figure 5-4 to show how CBAM highlights the relevant facial features in an original image. Brighter regions indicate greater attention weights, signifying the relevance of those regions for subsequent processing.

$$\begin{aligned} M' &= A_c(M) \odot M, \\ M'' &= A_s(M') \odot M', \end{aligned} \quad (5.2)$$

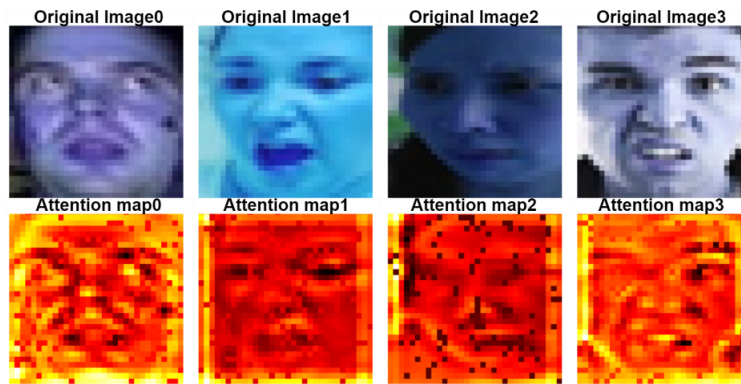


Figure 5-4: Visualization of learned attention weights of CBAM attention on input images.

5.3.3 Swin Transformer

Swin Transformer [99] is developed as a general-purpose transformer-based architecture for vision tasks. It performs local self-attention within distinct windows and attains linear computational complexity. The streamlined architecture of the Swin Transformer used in the proposed model is illustrated in Figure 5-5. Firstly, attentive features extracted from CBAM attention undergo patch partitioning process which splits input images into non-overlapping patches with a patch size of 2×2 . Each patch is considered a “token”, where its feature is configured as a concatenation of the pixel

values of an RGB image. Each feature with dimension $2 \times 2 \times 3$ will be flattened and converted into a vector of length 12. The vector is projected to an arbitrary dimension (C) using a linear embedding layer. On these patch tokens, two Swin Transformer blocks (within a single stage only) with modified self-attention mechanisms are used, and maintain the number of tokens ($H/2 \times W/2$). A transformer block with a linear embedding layer is referred to as stage 1 in Figure 5-5.

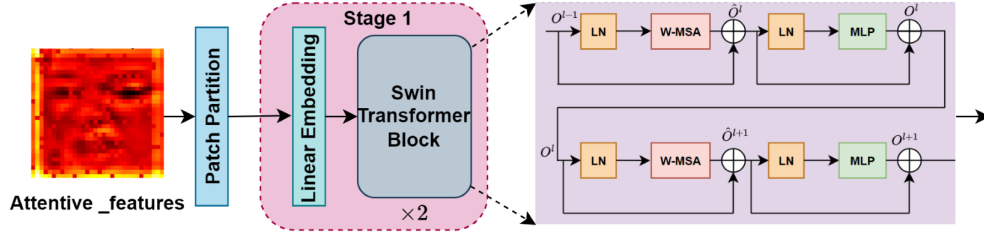


Figure 5-5: Streamlined Swin Transformer Architecture used in the proposed model.

Swin Transformer Block

The Swin Transformer block has a hierarchical structure, which allows it to process images of varying sizes. So, to create the Swin Transformer, the conventional multi-head self-attention (MSA) module in a transformer block is replaced with an SW-MSA module (Multi-head attention with shifted windows) to compute the self-attention within local windows, while other layers remain the same. It includes shifted-window-based multi-head self-attention and an MLP(2-layer) with a GELU activation function. A normalization layer is added before every MSA and MLP component, and these components are followed by the residual connection. The output of the Swin Transformer block is then passed to the patch merging layer in the network. By stacking two Swin Transformer blocks on top of each other, the model can learn increasingly complex representations of the input image.

In SW-MSA, the windows have been organized for image partitioning so that the partitions don't overlap with each other. This mechanism improves the computation complexity as shown in Eq. 5.3. Suppose the window has $W \times W$ patches, the complexity of window-based self-attention and the basic MSA module on a given

image of $h \times w$ patches can be defined as:

$$\begin{aligned}\Omega(MSA) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(W - MSA) &= 4hwC^2 + 2W^2hwC,\end{aligned}\tag{5.3}$$

A shifted window partitioning approach is followed between successive Swin Transformer blocks to incorporate cross-window links and maintain the computation efficiency of non-overlapping windows at the same time. With this approach, the output of successive Swin Transformer blocks is defined as:

$$\begin{aligned}\hat{O} &= W - MSA(LN(O^{l-1})) + O^{l-1}, \\ O^l &= MLP(LN(\hat{O}^l)) + \hat{O}^l, \\ \hat{O}^{l+1} &= SW - MSA(LN(O^l)) + O^l, \\ O^{l+1} &= MLP(LN(\hat{O}^{l+1})) + \hat{O}^{l+1},\end{aligned}\tag{5.4}$$

W-MSA denotes window-based self-attention and SW-MSA denotes shifted windows-based multi-head self-attention. In the above Eq. 5.4, \hat{O}^l and O^l represent the output of SW-MSA and MLP modules for l block, respectively.

5.3.4 Compound Emotion Classification

The output of the Swin Transformer block is fed to the patch merging layer to merge two adjacent patches for downsampling of the feature map. This layer is followed by a global averaging layer to reduce the dimensionality of the output, making it computationally efficient while preserving relevant information for downstream tasks. The output of the global averaging layer serves as input to a dense layer. In a dense layer with softmax activation, the output is produced by applying the softmax function on outputs of the global average pooling layer s_j by applying the softmax function to obtain probabilities (\mathbb{Y}) for each class(i) as shown in Eq. 5.5. The categorical cross-entropy loss (CE) is computed using the predicted probability \mathbb{Y}_i and true probability distribution Y_i for the i^{th} class in Eq. 5.5.

$$\begin{aligned} \mathbb{Y}_i &= \frac{e^{s_i}}{\sum_j^{\mathbb{C}} e^{s_j}} \\ CE &= \sum_i Y_i \log(\mathbb{Y}_i) \end{aligned} \tag{5.5}$$

The provided algorithm 5.1 outlines the processing steps for training a compound emotion recognition model, referred to as LSwin-CBAM, using a dataset (RAF-DB, EmotioNet). The algorithm starts by taking a dataset $\{X_i, Y_i\}_n$, where X_i represents facial images, and Y_i represents corresponding labels for compound emotions. The CutMix augmentation method is applied to the dataset, creating augmented samples $\{\hat{x}_i, \hat{y}_i\}$. The CBAM (Convolutional Block Attention Module) attention mechanism is applied to enhance the model’s ability to focus on important regions within facial images. A 2D convolutional layer is used to produce a feature map M from the augmented images and further processed through channel and spatial attention operations producing M' and M'' output respectively. The feature map M'' is partitioned into patches of size P_s . The algorithm proceeds through the stage of Swin Transformer block which involves a linear embedding layer, iterative application of Swin Transformer blocks $\times 2$ followed by a patch merging layer (generally applied, if the current stage is not the last one, but here single stage is used and patch merging layer is also applied exceptionally). Emphasizing efficiency without compromising accuracy, this model enables real-time processing, making it ideal for deployment in resource-constrained environments while maintaining a high level of precision in emotion detection.

5.4 Experimental Results

Experiments were performed on 64-bit Windows 11, Intel Core i7 processor with 16GB RAM size, 8GB NVIDIA TITAN RTX graphics card. Tensorflow 2.8 has been used for the implementation of the proposed framework. The selection of hyperparameters is indeed a crucial aspect of our model development. The summary of hyperparameters of the model is shown in Table 5.1. All the hyperparameters used

Algorithm 5.1 Algorithm of the proposed model.

Input:

Dataset $\in \{X_i, Y_i\}_{i=1}^n$ where n is number of images in dataset
 Model parameters θ
 Initial learning rate l_r
 Batch Size B
 Dropout Dp
 Learning rate decay factor d_r after every n epochs
 Number of epochs \mathbb{E}
 Patch size \mathbb{P}_s
 Number of stages \mathbb{N}_s
 Depth of Swin Transformer block \mathbb{D}

Output:

Trained LSwIn-CBAM model to recognize compound emotions from facial images

- 1: Input the dataset, where Dataset $\in \{X_i, Y_i\}_{i=1}^n$
- 2: Apply the CutMix augmentation method
- 3: **for** $i = 1$ **to** n **do**
 $\{\hat{x}_i, \hat{y}_i\} = CutMix(X_i, Y_i)$,
- 4: **end for**
- 5: Apply CBAM attention mechanism
- 6: Apply a 2D convolutional layer to produce a feature map M
- 7: Output of Channel attention: $M' = A_c(M) \odot M$,
- 8: Output of Spatial attention: $M'' = A_s(M') \odot M'$,
- 9: Apply patch partitioning to partition the patches of size \mathbb{P}_s
- 10: Stages:
- 11: Apply linear embedding layer
- 12: **for** $i = 1$ **to** \mathbb{N}_s **do**
- 13: **for** $j = 1$ **to** \mathbb{D} **do**
- 14: Apply Swin Transformer block
- 15: **if** $\mathbb{N}_s = 1$ **or** $i < \mathbb{N}_s - 1$ **then**
- 16: Apply patch merging layer
- 17: **end if**
- 18: **end for**
- 19: **end for**
- 20: Apply the global average pooling layer
- 21: Apply the dense layer which gives the final output

in the model are decided experimentally. These hyperparameters collectively define the configuration of a proposed model for facial emotion recognition, and these values have been meticulously selected through multiple experiments to determine the performance and behavior of the model during training. The learning rate is set to 0.001, this parameter is decided experimentally and controls the step size during the optimization process with learning rate decay of 0.0001. The dropout rate is 0.05 to prevent the model from overfitting by dropping several neurons. Several experiments were conducted using different dropout values, and the results, presented in Table 5.2 and Table 5.3, indicate that the optimal value chosen is 0.05. AdamW, is used for training LSwin-CBAM due to its improved weight decay handling (weight decay = 0.0001). Improved weight decay handling in AdamW contributes to reduce the generalization gap, making the model generalize better to unseen facial expressions and diverse datasets. Along with that, AdamW has been observed to provide more stable training and faster convergence compared to standard Adam, which is crucial for effectively training complex models like transformers used in facial emotion recognition. Categorical Cross-Entropy (CCE) loss is used for transformer models in multi-class classification due to its ability to compare predicted probability distributions with true labels, providing smooth gradients for optimization.

Table 5.1: Summary of hyperparameters used in LSwin-CBAM.

Hyperparameter	Value
Batch Size	16
Learning rate	0.001
Learning rate decay factor	0.0001
Dropout	0.05
Number of Epochs	500
Patch Size	2×2
Window Size	2×2
Number of Stages	1
Depth of Swin Transformer block	2

5.4.1 Datasets

Realworld Affective Face Database (RAF-DB)

The RAF-DB (Real-world Affective Faces Database) [72] is a face expression dataset. It contains 29672 facial photos labeled by 40 different taggers with seven basic (“Anger”, “Sadness”, “Happiness”, “Disgust”, “Fear”, “Surprise”, and “Neutral”) and eleven compound expressions (0: “Happily Surprised”, 1: “Happily Disgusted”, 2: “Sadly Fearful”, 3: “Sadly Angry”, 4: “Sadly Surprised”, 5: “Sadly Disgusted”, 6: “Fearfully Angry”, 7: “Fearfully Surprised”, 8: “Angrily Surprised”, 9: “Angrily Disgusted”, 10: “Disgustedly Surprised”). The participants’ age, gender, and ethnicity, as well as lighting conditions, head postures, occlusions (e.g. spectacles, self-occlusion, or facial hair), and post-processing processes (e.g. special effects and various filters), are all represented in this database. The number of samples in each class is depicted in Figure 5-6 which shows that the dataset is highly imbalanced.

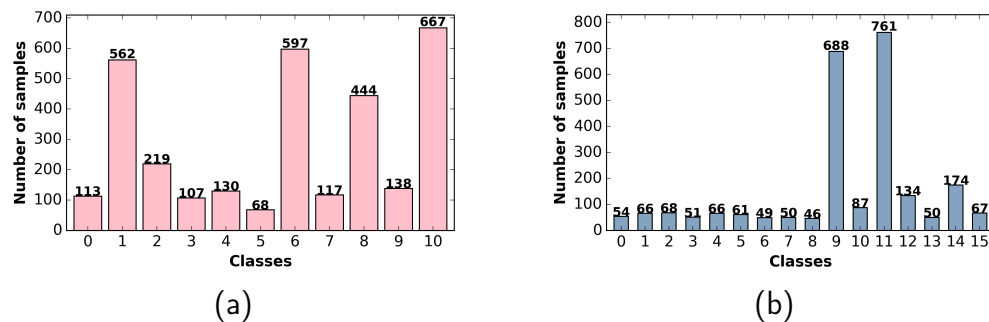


Figure 5-6: Number of samples in a) RAF-DB dataset b) EmotioNet Dataset.

EmotioNet

EmotioNet [3] is a large facial expression dataset that has nearly one million facial images and corresponding labels. It offers a wide range of facial expressions including basic and compound expressions. Images are downloaded from the internet by using different keys for 23 emotions based on action units (AUs). In our experiment, 2474 images are used which has 16 basic and compound emotion categories (0: “Angrily disgusted”, 1: “Angrily surprised”, 2: “Angry”, 3: “Appalled”, 4: “Awed”, 5: “Disgusted”,

6: “Fearful”, 7: “Fearfully angry”, 8: “Fearfully surprised”, 9: “Happily disgusted”, 10: “Happily surprised”, 11: “Happy”, 12: “Sad”, 13: “Sadly angry”, 14: “Sadly disgusted”, 15: “Surprised”). The EmotioNet dataset is highly imbalanced due to the presence of skewed class proportion as shown in Figure 5-6

The proposed model is evaluated on two compound emotion datasets: RAF-DB and EmotioNet. Both datasets are imbalanced datasets as their class distribution is very unequal which significantly affects the performance of the model.

5.4.2 Class Weights

Class imbalance is a common problem in image/video classification problems. It merely indicates that the frequency of a class is severely unbalanced, i.e., the frequency of one class is far more than that of the other classes. In other words, it introduces a bias towards the majority classes. To overcome this problem, class weights are adopted which assign weights to the classes by giving less weight to majority classes and high weights to minority classes which penalized the misclassification rate of the model. The general formula to calculate the class weight for each given class is shown in Eq. 5.6.

$$W_i = N_i / (C * N_i), \quad (5.6)$$

where W_i denotes weight for i^{th} class, N_i and C represents number of samples in i^{th} class and number of classes respectively.

5.4.3 Evaluation Metrics

As RAF-DB and EmotioNet datasets are imbalanced, accuracy is not considered a good metric for evaluating the model. This can be understood by a simple example. Suppose a dataset has two classes: Real or Fake and Real class has 1000 samples (majority class) whereas the Fake class has only 10 samples (minority class). After training on the given dataset, the model gives 99% accuracy (predicting 1000 samples correctly including 999 real and 1 fake sample or Real: 99% accuracy and Fake: 1%)

which is remarkable indeed. If we turn the spotlight on these results, it signifies that the model failed to predict the fake samples correctly which concludes that the proposed model is not a good predictor. So, it would be unfair to consider the accuracy as a performance metric for imbalanced datasets. Hence, other evaluation metrics are considered to justify the proposed model like precision (P), recall (R), f1 score ($F1_m$), ROC curve (Receiver operating characteristic curve), ROC_AUC score (Area under the Receiver Operating Characteristics Curve). ROC curve and ROC_AUC score have been considered as the optimistic metric to evaluate the performance of the classification model on imbalanced datasets because it is independent of class distribution and invariant to the threshold.

5.4.4 Experimental results on RAF-DB(Compound) dataset

The RAF-DB dataset contains 3240 images in the training subset and 840 images in the testing subset [72]. The proposed model is then trained with different configurations changing batch size, dropout, and learning rate whose results are shown in Table 5.2. The cosine learning rate decay is adopted to avoid the problem of local minima. Firstly, we experimented with the model by taking different batch sizes and dropout giving the best results with 16 and 0.05 in terms of accuracy, ROC_AUC score, precision, recall, and f1-score (51.81%, 0.78, 0.45, 0.30, and 0.30 respectively). Through different experiments, it has been observed that increasing batch size degrades the performance of the model. We also examined the influence of dropout on the model performance on different values of batch size constant (16, 32, 64, 128) and found the optimal value of dropout which is 0.05. The value of dropout above 0.05 shows degraded results on the dataset as shown in Table 5.2.

The model gives 51.81% accuracy but due to an imbalanced dataset, it cannot classify minority classes (having a very less number of samples in the training and testing subset). To overcome this problem, class weights are introduced during the training process so that the model emphasizes the minority class to improve the classification rate of minority classes as well. The class weights are initialized according to the formula given in Eq. 5.6 but with multiple experiments using different weight

sets, the optimized weight set is [0: 2.60, 1: 0.92, 2: 1.744, 3: 2.95, 4: 2.265, 5: 4.331, 6: 0.83, 7: 2.517, 8: 0.963, 9: 2.134, 10: 0.841]. The model with class weights classified the testing samples with an accuracy of 48% which is lowered but class weights are exploited to make the model learn all the classes (including the minority classes) and classify them as shown in Figure 5-7 and Figure 5-8. Figure 5-8a) shows that class weights improved the AUC of minority classes which is lower in Figure 5-7a) due to imbalanced datasets. Figure 5-8b) shows how class weights improved the corrected classified samples of minority classes as well which are lesser or zero in the confusion metric shown in Figure 5-7b).

Table 5.2: Experimental results of LSwin-CBAM on RAF-DB dataset, * indicates results of model trained with class weights.

Batch_size	Dropout	Acc(%)	RA	P	R	$F1_m$
16	0.04	49.76%	0.78	0.33	0.28	0.27
	0.05	51.81%	0.78	0.45	0.30	0.30
	0.06	50.12%	0.78	0.34	0.28	0.26
	0.07	49.64%	0.79	0.40	0.29	0.29
32	0.04	49.17%	0.77	0.32	0.30	0.29
	0.05	50%	0.77	0.35	0.29	0.27
	0.06	50.36%	0.77	0.36	0.29	0.29
	0.07	47.51%	0.76	0.25	0.25	0.23
64	0.04	47.98%	0.76	0.44	0.26	0.25
	0.05	48.46%	0.76	0.35	0.29	0.27
	0.06	46.56%	0.77	0.38	0.25	0.23
	0.07	48.34%	0.75	0.39	0.28	0.28
128	0.04	47.03%	0.76	0.37	0.26	0.26
	0.05	47.62%	0.76	0.38	0.29	0.29
	0.06	45.96%	0.74	0.40	0.26	0.25
	0.07	46.67%	0.77	0.37	0.29	0.29
*16	0.05	48%	0.74	0.36	0.31	0.31

5.4.5 Experimental results on EmotioNet dataset

In our experiment, a set of EmotioNet challenges having 2474 facial expression images have been used depicting basic and compound emotions [3]. The images are manually annotated but these are a very small number of images for the training of the model. The dataset is split in the ratio of 80:20 giving 1924 images to the training set and 481 to the testing set. With extensive experiments, the proposed model has achieved

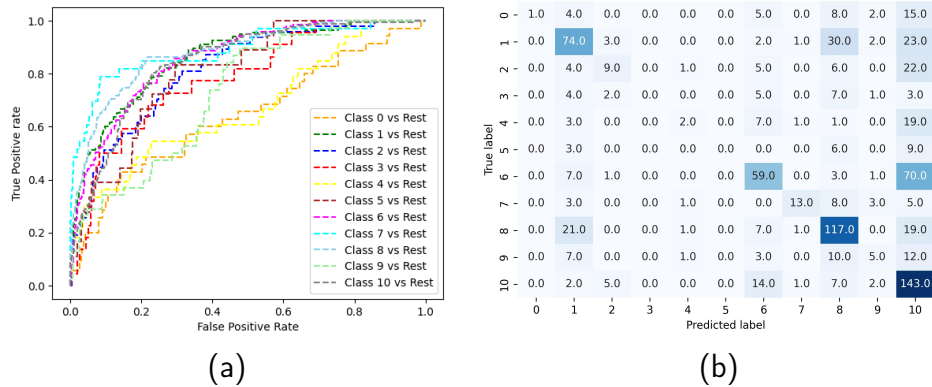


Figure 5-7: Evaluated results on RAF-DB dataset a) ROC_Curve b) Confusion metric.

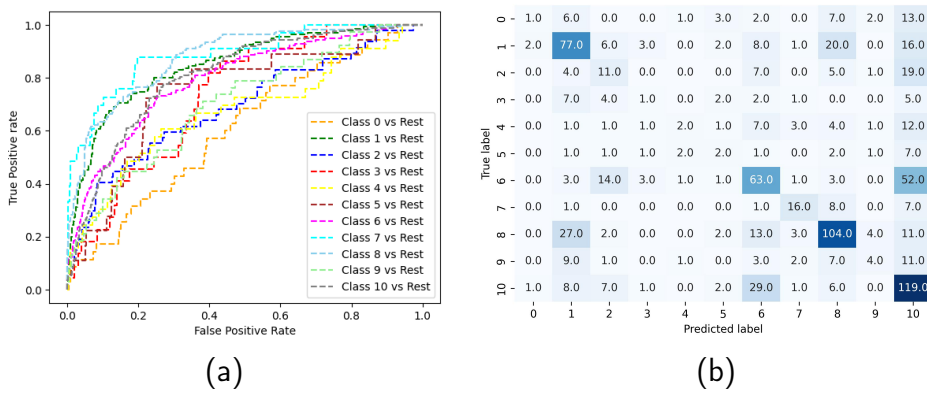


Figure 5-8: Evaluated results on RAF-DB dataset (with class weights) a) ROC_Curve b) Confusion metric.

39.67% accuracy. To the best of our knowledge, there is no work available that has considered a small set of the EmotioNet dataset for training purposes. The results are obtained by varying different batch sizes and dropouts as shown in Table 5.3. Due to an imbalanced dataset, the minority classes are not well classified so class weights are incorporated and the performance of the model without and with class weights is depicted in Figure 5-9 and Figure 5-10. By comparing Figure 5-9a) and Figure 5-10a), it is clear that adding class weights has improved the AUC of minority classes. In the confusion metric shown in Figure 5-9b), minority classes are not predicted well but adding class weights makes the model recognize the minority classes as well, shown

in diagonal entries of confusion metric in Figure 5-10.

Table 5.3: Experimental results of LSwin-CBAM on EmotioNet dataset, * indicates results of model trained with class weights.

Batch_size	Dropout	Acc(%)	RA	P	R	$F1_m$
16	0.04	36.38%	0.56	0.13	0.10	0.09
	0.05	39.67%	0.59	0.21	0.09	0.09
	0.06	35.76%	0.59	0.13	0.09	0.09
	0.07	36.38%	0.60	0.07	0.08	0.07
32	0.04	35.56%	0.59	0.12	0.09	0.09
	0.05	36.59%	0.58	0.15	0.11	0.11
	0.06	35.55%	0.57	0.15	0.09	0.10
	0.07	37.21%	0.61	0.24	0.12	0.11
64	0.04	37.42%	0.61	0.18	0.13	0.12
	0.05	36.17%	0.61	0.19	0.11	0.12
	0.06	36.38%	0.61	0.11	0.10	0.10
	0.07	36.38%	0.61	0.13	0.11	0.10
128	0.04	35.97%	0.58	0.10	0.09	0.07
	0.05	35.34%	0.60	0.18	0.13	0.13
	0.06	34.3%	0.59	0.10	0.10	0.14
	0.07	34.72%	0.59	0.18	0.11	0.11
*16	0.05	32.26	0.56	0.09	0.08	0.06

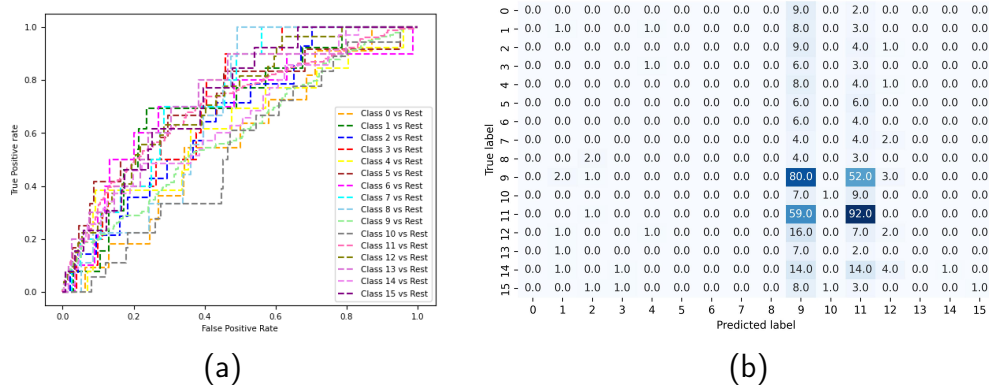


Figure 5-9: Evaluated results on Emotionet dataset a) ROC_Curve b) Confusion metric.

The imbalance in the Emotionet dataset is what prevents it from accurately predicting minority classes. There are different methods available to handle imbalanced datasets. Class weights are used here to reduce the biases of the model towards the majority class by assigning higher weights to the minority class which affects the overall performance but improves the class-wise accuracy of the minority classes.

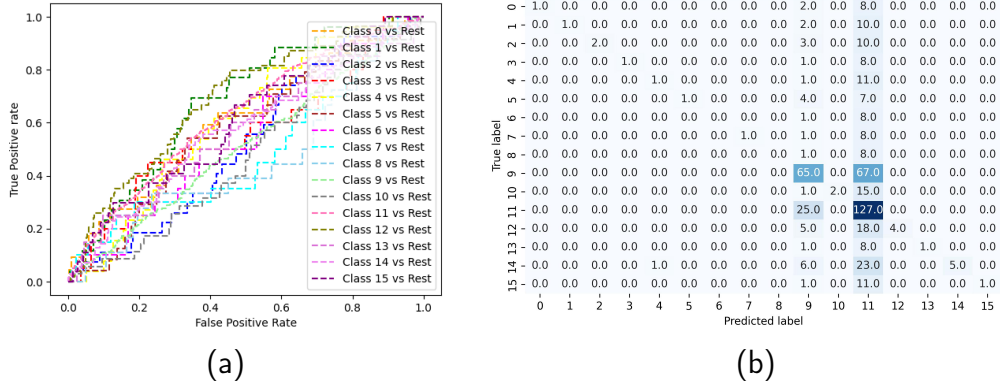


Figure 5-10: Evaluated results on Emotionet dataset (with class weights) a) ROC_Curve b) Confusion metric.

5.4.6 Ablation Study

The ablation study is provided in this section to justify the reasonability of the proposed model. Quantitative evaluation is conducted to support the contribution of each module in the model. Initially, a Swin Transformer has experimented alone on RAF-DB and EmotioNet datasets which classify the facial expression with 45.25% and 32.12% accuracy respectively. As the given datasets are imbalanced, adding the CutMix augmentation technique improved the performance by 3.02% (RAF-DB), 1.75% (EmotioNet) and hence classifies the basic and compound emotions with 47.22%(RAF-DB), and 33.95%(EmotioNet) accuracy. The CBAM attention mechanism emphasizes channel and spatial information, enhances the target regions, and extracts relevant and detailed information from the given images. Adding this module improved the classification performance improved by 4.59%(RAF-DB), 5.72%(EmotioNet) which leads to 51.81% and 39.67% accuracy for RAF-DB and EmotioNet datasets respectively. Table 5.4, 5.5 show the results of RAF-DB and EmotioNet datasets in terms of accuracy and ROC_AUC score (AUC), precision (P), recall(R), and f1-score ($F1_m$) obtained by adding each module to the model. Figure 5-11, 5-12 shows the class-wise accuracy obtained without and with class weights on RAF-DB and EmotioNet datasets respectively. Results show that incorporating class weights has improved the class-wise accuracy of both datasets.

Table 5.4: Significance of adding each module to LSwin-CBAM while experimenting on RAF-DB dataset.

Model	Acc	AUC	P	R	$F1_m$
Baseline Swin	45.25%	0.73	0.41	0.27	0.26
LSwin+CutMix	47.22%	0.77	0.41	0.31	0.33
LSwin-CBAM	51.81%	0.78	0.45	0.30	0.30

Table 5.5: Significance of adding each module to LSwin-CBAM while experimenting on EmotioNet dataset.

Model	Acc	AUC	P	R	$F1_m$
Swin	32.12%	0.55	0.13	0.10	0.11
LSwin+CutMix	33.95%	0.57	0.10	0.07	0.6
LSwin-CBAM	39.67%	0.59	0.21	0.09	0.09

Parametric and Architectural Influence

The general architecture of the Swin transformer (ST) [99] has 4 stages where the ST block is preceded by a linear embedding layer in the first stage and a patch merging layer in the last three stages. This complex architecture leads to a large number of trainable parameters which increases training time without inducing much improvement in classification performance. So, this section shows the ablation study of parametric and architectural influence by adding the number of stages in the model. Table 5.6,5.7 shows the significance of adding each stage to the model and results show the effectiveness of the lightweight LSwin-CBAM model for compound emotion recognition.

Table 5.6: Parametric and architectural influence of Swin Transformer stages on the performance of model on RAF-DB dataset.

Model	Stages	Param	Acc	AUC	P	R	$F1_m$
Swin4-CBAM	4	6M	49.29%	0.78	0.43	0.29	0.28
Swin3-CBAM	3	2.9M	49.41%	0.78	0.39	0.28	0.28
Swin2-CBAM	2	0.41M	49.64%	0.77	0.35	0.27	0.26
LSwin-CBAM	1	0.15M	51.81%	0.78	0.45	0.30	0.30

Table 5.7: Parametric and architectural influence of Swin Transformer stages on the performance of model on Emotionet dataset.

Model	Stages	Param	Acc	AUC	P	R	$F1_m$
Swin4-CBAM	4	6M	35.56%	0.57	0.08	0.08	0.06
Swin3-CBAM	3	2.9M	35.35%	0.61	0.11	0.10	0.09
Swin2-CBAM	2	0.41M	34.75%	0.58	0.10	0.09	0.09
LSwin-CBAM	1	0.15M	39.67%	0.64	0.21	0.09	0.09

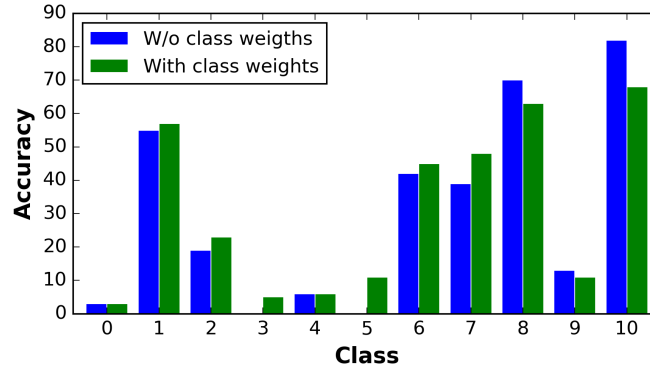


Figure 5-11: Class-wise average accuracy obtained on RAF-DB without and with class weights.

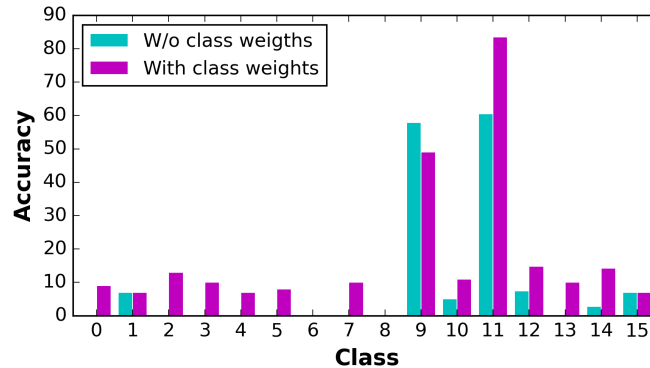


Figure 5-12: Class-wise average accuracy obtained on EmotioNet without and with class weights.

5.4.7 Comparative study of LSwin-CBAM with state-of-the-art

Firstly, the model is trained with RAF-DB and EmotioNet datasets separately. The evaluated results are compared with the state-of-the-art methods as shown in Table 5.8 and Table 5.9 for RAF-DB and EmotioNet datasets respectively. Table 5.8 shows that LSwin-CBAM outperforms state-of-the-art methods 0.1% the best per-

forming C-EXPR-NET (with parameters size of 14MB which is much higher than the parametric size of our proposed model LSwin-CBAM), 1.61% the best performing Multi-Scale Action Unit (AU)-based Network (MSAU-Net), 7.21% the DLP-CNN pre-trained on RAF-DB basic emotions and fine-tuned on RAF-DB compound emotions, 5.71% the Relation Convolutional Neural Network (ReCNN) to capture the associations between relevant regions and facial expressions, 4.81% the Pyramid with super-resolution (PSR) network for in-the-wild FER. The EmotioNet dataset has 2472 images with unequal distribution of the classes (imbalanced dataset) and there is limited literature leveraging this dataset. The Table5.9 illustrates the comparison of the proposed model with state-of-the-art methods for EmotioNet dataset, indicating its superior performance over existing models while utilizing a significantly lower number of parameters. Jiang et al. [117] used expression soft label mining (ESLM) to produce soft target labels automatically for the learning process with the ResNet50 classification model. ResNet50 was tested with LSR (Label Smoothing Regularization) [118] and PS-KD(Progressive Self-knowledge Distillation) [119], and ISLM(Iterated Soft Label Mining) [117] methods achieving classification accuracies of 37.23%, 38.95%, 38.15% respectively. In addition to the enhanced accuracy, the LSwin-CBAM model achieves a smaller parameter size of 0.581MB compared to other models (ResNet50-LSR [118], ResNet50-PS-KD [119], ResNet50-ISLM [117]) which utilize 22.53MB.

Table 5.8: Comparative analysis of state-of-the-art methods and proposed method on RAF-DB dataset.

Model	Accuracy	Parameters_size(MB)
MAE [120]	24.94%	≈ 7.65
VGG + mSVM [72]	31.6%	≈ 528
CoAtNet [121]	33.42%	≈ 52.21
baseDCNN + mSVM [72]	40.2%	≈ 5
ViT [67]	44.18%	≈ 162.24
DLP-CNN [72]	44.6%	≈ 1.34
ReCNN [86]	46.1%	≈ 4.82
VGG16+PSR [122]	46.5%	≈ 528
Fine-tuned MSAU-Net [58]	50.2%	-
Zero-shot C-EXPR-NET [55]	51.7	≈ 14
LSwin-CBAM	51.81%	≈ 0.582

Table 5.9: Comparative analysis of state-of-the-art methods and proposed method on EmotioNet dataset.

Model	Accuracy	Parameters_size(MB)
ResNet50 [81]	35.01%	≈ 22.53
ResNet50-LSR [118]	37.23%	≈ 22.53
ResNet50-PS-KD [119]	38.95%	≈ 22.53
ResNet50-ISLM [117]	38.15%	≈ 22.53
LSwin-CBAM	39.67%	≈ 0.582

The comparison of LSwin-CBAM with transformer-based models like ViT [67], MAE [120], and COAtNet [121] highlights its superior performance in image classification of compound emotions as shown in Table 5.8 and Figure 5-13. Its efficiency in utilizing limited training data makes it more suitable for scenarios with sparse data availability, as heavy models are prone to under-fitting. LSwin-CBAM’s lightweight design also enables better scalability and adaptability to smaller datasets compared to transformer models like ViT and COAtNet, which struggle with imbalanced datasets due to their demand for extensive samples per class during training. Moreover, MAE suffers from poor reconstruction quality and instability during training, while CoAtNet and ViT struggle with insufficient representative samples for each class, resulting in suboptimal performance.

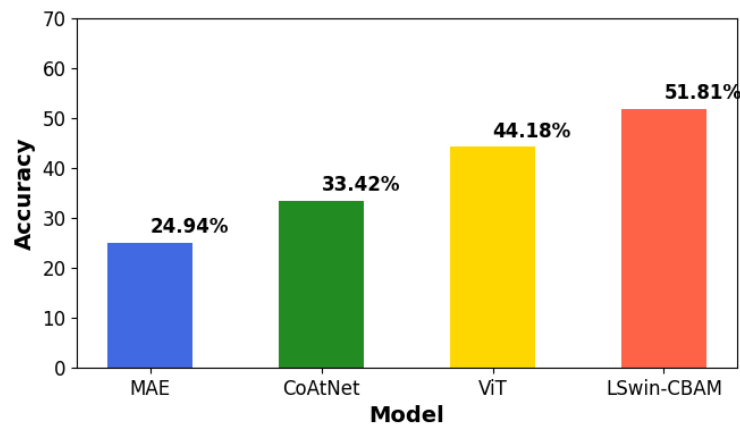


Figure 5-13: Performance Evaluation: LSwin-CBAM vs. Transformer Variants on RAF-DB dataset.

5.5 Conclusion

This study proposed LSwin-CBAM model for compound emotion recognition that incorporates CutMix augmentation to augment the training data and CBAM attention mechanism to emphasize critical sections of the image. Experimental results on RAF-DB and EmotioNet datasets show the superiority of this model in terms of different evaluation metrics (Accuracy, ROC_AUC score, Precision, Recall, F1-score) over the state-of-the-art methods. The comparative results shown in Table 5.8, 5.9 prove that LSwin-CBAM classified the compound emotion with good accuracy and with less computational complexity. The proposed model achieved 51.81% for RAF-DB and 39.67% for the EmotioNet dataset but fails to classify the minority classes as well. This problem occurs due to the imbalanced dataset which can be alleviated by incorporating class weights. Using class weights declined the overall accuracy but make it to classify minority classes for RAF-DB and EmotioNet datasets. The proposed model is very lightweight in terms of trainable parameters (0.15M) and still achieved impressive results which take it above the state-of-the-art methods.

Chapter 6

Design and development of a deep learning framework for driver’s emotion recognition

Nidhi, and Bindu Verma. “ViT-SLS: Vision Transformer with Stochastic Depth for Efficient Driver’s Emotion Recognition System” is communicated in *IEEE Transactions on Human-Machine Systems* (SCIE Indexed, IF: 3.6) (Communicated)

6.1 Introduction

This chapter presents an effective Vision Transformer-based framework termed as “ViT-SLS”, which employs a Vision Transformer with shifted patch tokenization where input images are shifted and broken into patches which are further propagated through the locality-self attention module (to enhance the performance of small size datasets) and followed with stochastic depth for regularization. In the previous chapters, we proposed framework for emotion recognition in the wild, compound emotions, analyzing the various attention mechanisms. As a conclusive part of this thesis, we worked on one of the applications of FER by developing a driver’s emotion recogni-

tion system. This system leverages FER model to monitor the emotional state of the driver, aiming to enhance road safety and driving experiences by detecting emotions such as “eye-close”, “happy”, “yawn”, etc.

According to a survey by the Indian Government’s National Crime Research Bureau (NCRB), Indian Roadways have a higher fatality rate [123]. Since 2006, the number of people killed in traffic accidents has risen steadily in India. According to the data, driver’s inattentive driving is the primary factor contributing to the 1.46 lakh deaths that occurred in road incidents in 2015. According to psychological research, unfavorable emotions including “grief”, “anger”, “disgust”, and “fear” trigger impulsive, careless, and fast driving [124]. “Anger” and “aggression” are two emotions that significantly impact driving behavior and raise the likelihood of accidents. According to Eyben et al. [125], stress and exhaustion are some factors that contribute to risky driving. Anxiety, grief, and other powerful emotions may also impair judgment when driving. Real-time emotion recognition has improved with the recent advances in sensor technology and deep learning (DL) algorithms, making it quite accurate and practical in everyday situations. Real-time emotional identification of the driver has emerged as a crucial solution, particularly for the traffic field, as it is capable of altering the driver’s behavior and lessening the likelihood of a crisis of a driver while driving.

Therefore, monitoring the driver’s emotional state is a key aspect of Advanced Driver Assistance Systems (ADAS). ADAS are currently being developed to keep drivers and passengers safe in the event of an accident by providing technologies that notify drivers of potential concerns. But, even the most recent and advanced autonomous vehicles on the road still need the driver to be alert and prepared to regain control of the vehicle in an emergency.

As humans can read the body language of others, how computers can be trained to recognize emotional state of the driver’s by observing facial expressions. This question motivates researchers to make a contribution to driver’s emotion recognition using facial expressions. In this chapter, we proposed a framework to detect the driver’s emotional states by observing the facial expression.

Due to the remarkable results of ViT in the image classification problems, the proposed model incorporates ViT with Shifted Patch Tokenization (SPT) which tokenizes the given images based on spatial feature shifting to employ the spatial information between adjacent pixels. The inclusion of locality self-attention is to eliminate the self-token relation by requiring each token to concentrate more on tokens with a close relationship to itself. The stochastic depth method has been exploited to skip a subset of layers and hence improves the performance of the model in less training time.

The proposed work has the following key points: this chapter proposes a fused Vision Transformer model termed ViT-SLS which incorporates a Vision Transformer with shifted patch tokenization, locality self-attention, and stochastic depth which comparatively improved the performance of the model. The advantage of adding a shifted patch tokenization component is to extract spatial detailing between nearby pixels and to make a broader receptive field applied to visual tokens. Locality self-attention influences the performance of the ViT-SLS by introducing irregularity in the distribution smoothing problem of attention score by highlighting the inter-token relations. Stochastic depth is used for regularization by skipping a subset of layers while training and keeping the network unchanged during testing.

6.2 Literature Survey

With the evolution of intelligent automatic human-machine interaction systems, the study of driver emotion recognition has also become an emerging paradigm. Tran et al. [126] used facial expressions and steering wheel data to recognize the driver's drowsiness using Support Vector Machine (SVM) classifier. In the proposed approach [126], facial expression recognition characterizes the drowsy and non-drowsy state of the driver. Malta et al. [127] proposed a method to analyze the driver's frustration in real-time using environmental information, the driver's emotions, and his respective responses. Xiao et al. [128] adopted a Deep CNN that has been pre-trained on CK+ and FER datasets and further fine-tuned as a backbone to recognize on-road driver's expressions. Researchers introduced transfer learning in the proposed method

FEDERNet [128] which is evaluated on their own collected dataset.

Rebolledo-Mendez et al. [129] proposed a human emotion detector that can analyze the driver's stress and tiredness of the driver which are generally associated with traffic accidents. Body sensor networks are used by the driver so that physiological signals can be transmitted to a vehicular onboard unit (OBU) that recognizes the driver's current state and sends alarm notification in case of emergency [129]. Tavakoli et al. [60] used facial expression data, gaze variability, driver's stress, and workload to examine how the external context can affect the driver's state. Wang et al. [130] developed an SVM-based driver lane-changing intention identification model considering the driver's eight emotions ("relief", "fear", "helplessness", "pleasure", "anxiety", "surprise", "anger", and "contempt"). Khandeel et al. [131] proposed a CNN model to classify drivers' emotions and evaluated it on commonly used facial expression datasets (CK+, JAFFE, KDEF) and real-time driver's emotion dataset (KMU-FED). Du et al. [132] proposed convolution based deep framework called CBLNN (Convolution Bidirectional Long Short-term Memory Neural Network) to predict the driver's emotions. Authors [132] used a multimodal approach where CNN is used to extract geometric facial features based on skin particulars and Bi-LSTM for heart rate analysis which are fused to predict the emotion classification output. In contrast to conventional approaches, Li et al. [133] proposed a model which included inputs from both the driver's facial expression and the cognitive process variables (age, gender, and driving age). Convolutional approaches were used to build the model for driver emotion recognition considering the driver's facial expression and cognitive process features as inputs. Zhang et al. [134] used Multi-task Cascaded Convolutional networks (MTCNN) for face detection and face alignment. CNN model was used to extract facial features from the input face sequences and for the classification process as well. Jeong et al. [135] proposed lightweight multi-layer random forests to recognize driver's facial emotions which consist of multiple random forests at each layer level. Authors justified the effectiveness of the proposed model by comparing the performance of renowned models like SqueezeNet [136], MobileNetV2 [137] and MobileNetV3 [138].

During the past few years, CNN have been playing a key role in solving DL’s visual task [83] [139]. This is partly because of convolutional layer’s strong inductive bias of spatial equivariance, which has been essential to learn general-purpose visual representations for simple transfer and effective performance. Interestingly, recent research has shown that transformer neural networks are equally capable of higher performance on large-scale image classification problems [140]. Along with that, it has been observed in the literature that there is very less work done on driver emotion datasets like KMU-FED [131] and D3S [141] which are captured in real-time environments. So, this study uses a vision transformer-based model named ViT-SLS which incorporates shifted patch tokenization, locality self-attention, and stochastic depth to improve the performance of driver’s emotion recognition system.

6.3 Proposed Method

Vision Transformer (ViT) [140] allows models to understand image structure independently as input images are depicted as sequences and predict the class labels for the given image. The input images are processed as a sequence of patches, where each patch has been flattened into a single vector, further concatenated by the channels of all of its pixels before projecting them linearly to the defined dimension of input. ViT has outperformed Convolutional Neural Network (CNN), and employs transformer structure to classify the images [140]. The advanced performance of ViT comes from pre-training the model using a large-scale dataset and its influence is caused by low locality inductive bias.

The proposed driver’s emotion recognition architecture is shown in Figure 6-1 which specifically incorporates three key aspects: Shifted Patch Tokenization (SPT) to retrieve the extensive spatial information, Locality self-attention (LSA) to eliminate the self-token relation i.e. diagonal entries by applying the diagonal masking, and Stochastic depth (SD) which skips a subset of layers by detouring them using the identity function during training of a mini-batch to decrease the training time of the network. First of all, the input data is fed to the network, but due to the small-

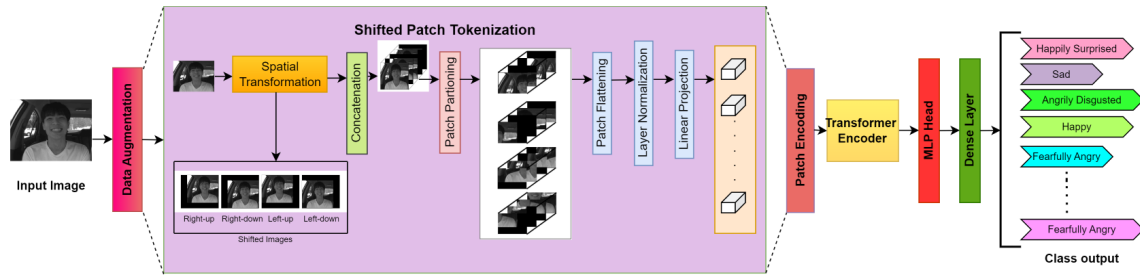


Figure 6-1: Proposed architecture of ViT-SLS (Vision Transformer integrated with Shifted patch tokenization and modified transformer encoder (including locality self-attention and stochastic depth)) for driver's emotion recognition.

size dataset, data augmentation is applied to augment the data. The proposed work applies geometry-based augmentation like random flipping, cropping, and re-scaling. This step is followed by shifted patch tokenization where the input image is shifted in each of the four diagonal directions (left-up, left-down, right-up, and right-down). Shifted images are concatenated with the original images which are further fed to the patch encoding layer. Patch encoding is done on the concatenated features where input is divided into patches first. Positional embeddings are used to add positional information to the sequence of input patches and class embedding has been applied to the sequence in accordance with the location of the image patch. This class embedding plays its key role after the self-attention mechanism to recognize the class of the input image.

Encoded patches with positional embedding are further passed to the transformer encoder whose architectural detailing is shown in Figure 6-2. The transformer encoder is made up of alternatively arranged layers of Multi-head Self-attention (MHSA) and Multi-layer perceptron (MLP) block. The multi-head attention layer divides the input into various heads in order to learn different levels of self-attention. MLP block is then applied to the concatenated output received from all heads. The applied MLP has two layers with a Gaussian Error Linear Unit (GELU) non-linear activation function. GeLU activation function is differentiable and introduces smoothness and non-linearity in the model. Stochastic depth is appended after the attention layer which skips a subset of layers during training and keeps them unchanged during testing, and MLP block with *stochastic_depth_rate* of 0.1 which signifies that selecting

any layer to be skipped with a given probability of 0.1 or 10%. Each block is followed by a layer normalization. The transformer encoder is then followed with an additional learnable classification block named the MLP Head.

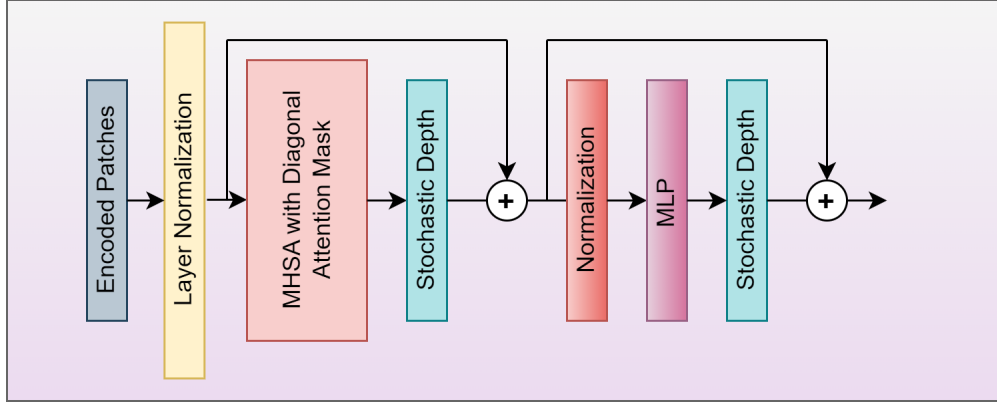


Figure 6-2: Modified transformer encoder module with locality self-attention and stochastic depth *Locality self-attention is incorporated in “MHSA with diagonal attention mask layer”.

6.3.1 Formulation of General ViT

The architectural overview of the standard ViTs is depicted in Figure 6-3 which takes input image $x \in \mathbb{R}^{H \times W \times C}$ where H, W, and C represent the height, width, and number of the channels. ViT first separates the input image into non-overlapping patches and then flattened them to produce vectors sequentially as formulated in Eq. (6.1):

$$P(x) = [x_p^1; x_p^2; x_p^3; \dots; x_p^n], \quad (6.1)$$

Here $x_p^i \in \mathbb{R}^{P^2 \cdot C}$ indicates the i-th flattened vector. N indicates the number of patches as formulated in Eq. (6.2) and P represents patch size.

$$N = HW/P^2, \quad (6.2)$$

Patch Embedding is done by applying linear projection to the flattened vectors. Embedded patches are inputs taken as visual tokens to the transformer encoder block, this process is called tokenization. The receptive fields of visual tokens and trans-

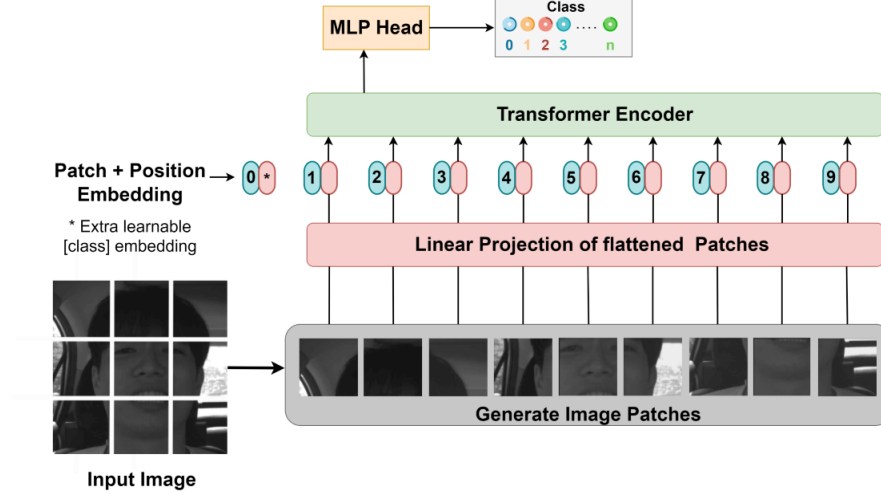


Figure 6-3: Model overview of general Vision Transformer [140].

former encoder are represented by $r_{f_{token}}$ and $r_{f_{trans}}$ respectively. Their relation can be shown in Eq. (6.3) as the tokenization operation in standard ViT is similar to the one in the convolutional layer with the same kernel size (k) and stride size (s).

$$r_{f_{token}} = r_{f_{trans}} \cdot s + (k - s), \quad (6.3)$$

The receptive field size of visual tokens r_{token} is quite smaller than that of trained features which leads to the problem of local inductive bias. Shifted Patch Tokenization is incorporated to extract extensive spatial information by escalating the receptive field of visual tokens.

The general ViT's self-attention module applies learnable linear projection to each visual token to compute query, key and value. According to Lee et al. [142] query and key values are obtained from the same visual tokens and hence have similar sizes. The dot product of similar-size vectors represented by $D \in \mathbb{R}^{(N+1) \times (N+1)}$ in Eq. (6.4) which indicates the self-token relation gives large values than that of values produced from inter-token relations. As a result, self-token relations receive comparatively high values from the softmax function in Eq. (6.5) while inter-token relations receive lower scores.

$$D = xL_q(xL_k)^T, \quad (6.4)$$

$$SF(x) = softmax(D \div \sqrt{d_k})xL_v, \quad (6.5)$$

Here, $xL_q \in \mathbb{R}^{d \times d_q}$, $xL_k \in \mathbb{R}^{d \times d_k}$, $xL_v \in \mathbb{R}^{d \times d_v}$ are linearly projected values of query, key, and value from the same token. Also, the dimensions of the query, key and value are represented by d_q , d_k and d_v respectively. In Eq. (6.5), D is divided by $\sqrt{d_k}$ so that the softmax function will not lead to a small gradient value.

According to the experimental findings of [142], d_k behaves as a high temperature of the softmax activation function and smooths out the distribution of the attention score but deteriorates the ViT's performance. So, locality self-attention is incorporated to solve the problem of smoothing caused by high temperature scaling in the softmax function.

6.3.2 Shifted Patch Tokenization

The first step is to spatially shift each input image by half the patch size in each of the four diagonal directions (left-up, left-down, right-up, and right-down). The shifted features are concatenated with the input after being cropped to the same size as the input image. The concatenated features are then flattened and separated into non-overlapping patches which are further passed to layer normalization and linear projection to obtain the visual tokens as shown in Eq. (6.6):

$$Sh(x) = L_{norm}(PE([x sh^1 sh^2 \dots sh^{N_s}]))L_s, \quad (6.6)$$

where N_s shows shifted number of images, $sh^n \in \mathbb{R}^{H \times W \times C}$ indicates the n -th shifted image, $L_s \in \mathbb{R}^{P^2 \cdot C \cdot (N_s+1) \times d_t}$ represents the learnable linear projection, and d_t shows the hidden dimension of transformer encoder.

SPT has been applied to the patch embedding layer and pooling layer [142]. The class token in the Vision Transformer is generally used to represent the information of the given image. In the patch embedding layer if the class token is used then it will be concatenated with the visual token and then applied the positional embedding to them. If a class token is not used, then a visual token will be directly added with the

positional embedding as shown in Eq. (6.7)

$$Sh_{pe}(x) = \begin{cases} [x_c \& Sh(x)] + L_{pos}, & \text{if } x_c \text{ exists} \\ Sh(x) + L_{pos}, & \text{otherwise} \end{cases} \quad (6.7)$$

Here $x_c \in \mathbb{R}^{d_t}$ is the class token and $L_{pos} \in \mathbb{R}^{(N+1) \times d_t}$ is the learnable positional embedding.

SPT is also used in the pooling layer to reduce the size of 3D features during the tokenization process because tokenization converts 3-D tensor features into 2D-matrix features i.e., x into $z = \mathbb{T}(x) \in \mathbb{R}^{N \times d}$ and hence reduce the visual token count. Class tokens and visual tokens are first separated, and 2D matrices that represent visual tokens are then transformed into 3D tensors containing the spatial structure. This means $\mathcal{T} : \mathbb{R}^{N \times d}$ is transformed to $\mathbb{R}^{(H/P) \times (W/P) \times d}$. At last, linearly projected class token will be concatenated with the embedded and reduced visual tokens. But if the class token doesn't exist, then \mathcal{T} is imposed to the SPT output as shown in Eq. (6.8).

$$Sh_{po}(z) = \begin{cases} [x_c L_c \& Sh(\mathcal{T}(z))], & \text{if } x_c \text{ exists} \\ Sh(\mathcal{T}(z)), & \text{otherwise} \end{cases} \quad (6.8)$$

Here, $L_c \in \mathbb{R}^{d \times d_t}$ is a learnable linear projection and d'_t indicates the hidden dimension of the upcoming stage.

6.3.3 Locality Self-Attention Method

The locality self-attention method is used to eliminate the self-token relation i.e. diagonal entries by applying the diagonal masking and for learnable temperature scaling.

Self-token relations are excluded from the softmax operation using diagonal masking by replacing diagonal entries with $-\infty$, hence increasing the scores of inter-token relations. This causes ViT to pay more attention to other tokens than to its own tokens. The formulation of diagonal masking is shown in Eq. (6.9).

$$D_{m,n}^M(x) = \begin{cases} D_{m,n}(x), & \text{if } m \neq n \\ \infty, & \text{otherwise} \end{cases} \quad (6.9)$$

Here, $D_{m,n}^M$ represents the masked entry of the similarity metric.

Locality self-attention incorporates learnable temperature scaling to find softmax temperature. According to the experimental findings by [142], the mean learned temperature is lesser than the stable temperature of general ViT. Generally, the reduced temperature of softmax operation enhances the attention score distribution. The application of the diagonal masking mechanism and learnable temperature scaling is formulated in Eq. (6.10), where the learnable temperature is represented by l_t .

$$\mathcal{L}(x) = \text{softmax}(D^M(x)/l_t) xL_v, \quad (6.10)$$

After the application of the diagonal masking mechanism and learnable temperature scaling, the $\mathcal{L}(x)$ is denoted as *inp* for the next step.

6.3.4 Stochastic Depth

The proposed fused strategy also incorporates stochastic depth as a regularization technique. This technique skips a subset of layers by detouring them using the identity function during the training of a mini-batch to decrease the training time of the network [143]. It is quite similar to the dropout method, with the exception that it works with a whole block of layers rather than the individual nodes that make up a layer. Stochastic depth rate defines the probability of dropping the block of layers represented by p_{drop} and the probability of keeping the block active during the training p_{keep} is formulated in Eq. (6.11). The formulation of stochastic depth in the training phase is shown in Eq. (6.12).

$$p_{keep} = 1 - p_{drop}, \quad (6.11)$$

$$out = GELU(b * f(inp) + identity(inp)), \quad (6.12)$$

where b is drawn from the Bernoulli random variable with the hyperparameter p , and also, $f(inp)$ represents the optional one or two convolutions. The value of b equals 1 indicates that the block survived and 0 indicates the block is skipped. The average survival probability of a block is represented by p . The formulation for the testing process must be adjusted for missing blocks during training as shown in Eq. (6.13). If all of them are activated simultaneously, it may result in strong signals [143].

$$out = GELU(p * f(inp) + inp), \quad (6.13)$$

Emotion Classification

The dense layer receives the row vector output from the preceding layer to compute the class probabilities or predictions where each value represents the likelihood of the corresponding class and eventually provides the emotion class as the final output. The sparse categorical cross-entropy has been used to compute the loss by measuring the discrepancy between the predicted probability distribution \mathbb{Y} and the true class label Y as shown in Eq. 6.14. The detailed algorithm of the proposed model is shown in Algorithm 6.1.

$$SE = \sum_i^c Y_i \log(\mathbb{Y}_i) \quad (6.14)$$

6.4 Experimental Results

The system requirements for the evaluation of the proposed model are 64-bit Windows 11, Intel Core i7 processor with 16GB RAM size, 8GB NVIDIA TITAN RTX graphics card. The proposed framework has been implemented using Tensorflow 2.8. The detailing of datasets is illustrated in Subsection 6.4.1. The class distribution of each dataset is shown in Figure 6-4 which shows the number of samples for each class in

Algorithm 6.1 Algorithm of the proposed model.

```

1: Input the data
2: Apply data augmentation
3: Apply Shifted patch tokenization method to shift the images
4:  $N$  = Number of images
5: for  $i = 1$  to  $N$  do
6:    $Sh(x) = [x, sh^1, sh^2, \dots, sh^{N_s}]$ 
7: end for
8: Perform patch encoding on shifted images and flatten them,
9:    $Sh(x) = PE[x, sh^1, sh^2, \dots, sh^{N_s}]$ 
10: Normalize the flattened patches,
11:    $Sh(x) = L_{norm}(PE[x, sh^1, sh^2, \dots, sh^{N_s}])$ 
12: Apply linear projection,
13:    $Sh(x) = L_{norm}(PE[x, sh^1, sh^2, \dots, sh^{N_s}])L_s$ 
14: Add the Patch Embedding layer
15: if class token exists then
16:    $Sh_{pe}(x) = [x_c, Sh(x)] + L_{pos}$   $\triangleright$  concatenate with the visual token and add positional embedding
17: else    $Sh(x) + L_{pos}$   $\triangleright$  directly add the positional embedding to the visual token
18: end if
19: Add the transformer encoder module (Step 19-38)
20: Normalize the encoded patches
21: Apply multi-head attention layer with locality-self attention (diagonal masking)
22: for matrix  $D_{m,n}^M(x)$  do
23:   if  $m \neq n$  then    $D_{m,n}^M(x) = D_{m,n}(x)$ 
24:   else    $D_{m,n}^M(x) = \infty$ 
25:   end if
26: end for
27:  $inp = \mathcal{L}(x)$ 
28: Apply stochastic depth to skip a subset of layers while training (Step 28-34)
29:  $stochastic\_depth\_rate = 0.1$ ,
30:   average survival probability =  $p$ ,
31:    $b = \text{Bernoulli}(p)$ 
32: if  $training\_process = true$  then    $out = GELU(b * f(inp) + identity(inp))$ 
33: else    $out = GELU(p * f(inp) + inp)$ 
34: end if
35: Normalize the outputs
36: Apply MLP feed-forward layer
37: Repeat Step 27
38: Apply MLP head
39: Apply a dense layer to produce the final output

```

the D3S, KMU-FED, CK+, and JAFFE dataset. For the experiment of the proposed study, small datasets are used because, as per the literature, two or three driver emotion recognition datasets are available like D3S [141] and KMU-FED [131] which have 1319 and 1106 images respectively. While conducting experiments, the optimal data splitting ratio was found to be 80-20%. Therefore, the model has been trained using 80% of the dataset while 20% is used for testing the model. The training subset

is further split into 70% training and 10% validation data. During the experiment, optimal configurations were determined: the patch size was set to 6, the count of transformer layers was set to 8 with an embedding dimension of 128, and the number of heads was set to 4. Input images are fed to the network in batches with a batch size of 16. AdamW optimizer has been used in the experiment, as it improves the model’s ability to generalize the samples in a better way [144]. Along with that, sparse categorical cross-entropy has been used as a loss function. The base models have been fused and modified to meet the objective, and from the experimental results for D3S, KMU-FED, CK+, and JAFFE datasets shown in Subsections 6.4.2, 6.4.3, and 6.4.4, it can be stated that the fused model gives better results than those obtained from any of the techniques used alone.

6.4.1 Datasets

Keimyung University Facial Expression of Drivers (KMU-FED) Dataset

Keimyung University Facial Expression of Drivers (KMU-FED) [131] contains the driver’s facial images which are captured in a real-time environment. The dataset consists of six basic emotion (“Anger”, “Disgust”, “Fear”, “Happy”, “Sad”, and “Surprise”) that were photographed with a Near Infrared (NIR) camera mounted on the steering wheel or the dashboard. It has 55 image sequences captured from 12 participants giving 1106 frames in total. The duration of image sequences varies from 10 to 26 frames with different lightning variations (back, right, left light, and front) and partially occluded images with hair or sunglasses.

Driver Drowsiness Dataset (D3S) Dataset

D3S Dataset is created to detect the driver’s drowsiness while driving [141]. RGB Camera is used to capture video streams from four subjects in the real environment. The dataset consists of four emotions: “Eyeclose”, “Happy”, “Yawn”, and “Neutral”.

Extended Cohn Kanade (CK+) Database

The CK+ dataset [145] consists of 593 video sequences collected from 123 participants of age ranging from 18 to 15 years from different genders and culture. Each video indicates a facial change from the neutral expression to a selected peak expression, recorded at 30 frames per second (FPS) in 640×480 or 640×490 pixels resolution. 327 videos of the total videos are annotated with one of the seven basic emotions: “Surprise”, “Anger”, “Happiness”, “Contempt”, “Fear”, and “Disgust”.

The Japanese Female Facial Expression (JAFFE)

This dataset [146] contains 213 images from 10 distinct Japanese female participants. Each participant was instructed to pose for seven emotions which include six basic emotions (“Anger”, “Fear”, “Disgust”, “Surprise”, “Happiness”, and “Sadness”) and one neutral which are further labeled by the average semantic rating given by 60 annotators on each emotion.

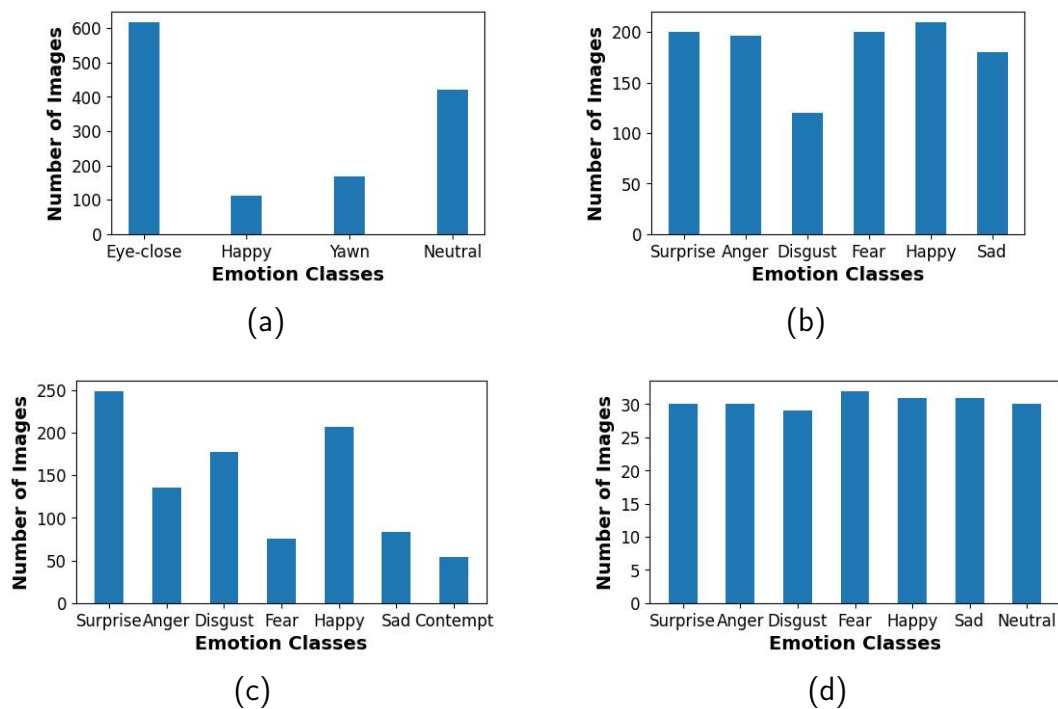


Figure 6-4: Distribution of classes for (a) D3S, (b) KMU-FED, (c) CK+, and (d) JAFFE datasets.

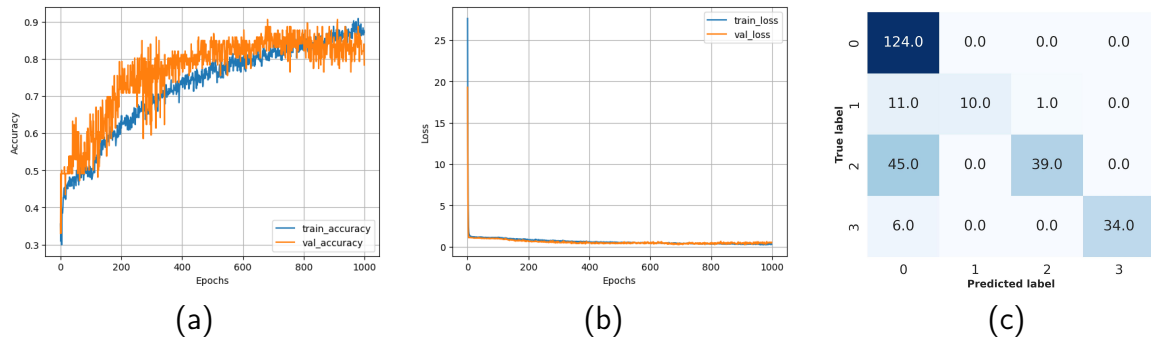


Figure 6-5: Performance of the general Vision Transformer on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.

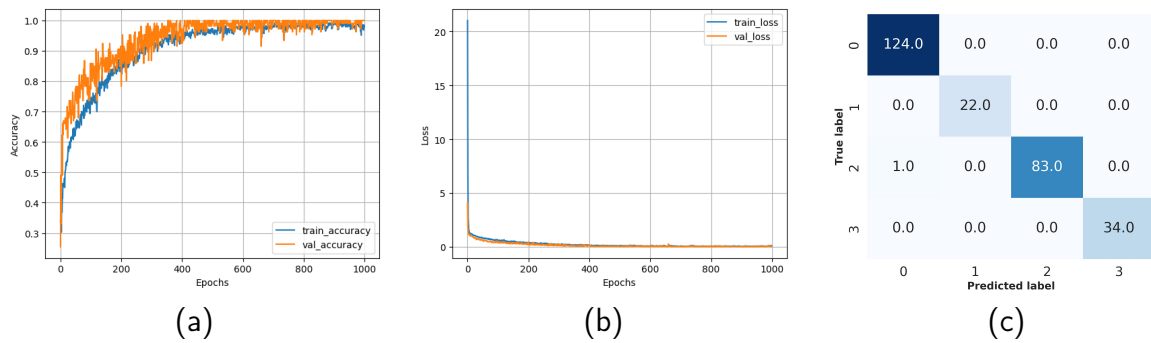


Figure 6-6: Performance of ViT-SL on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.

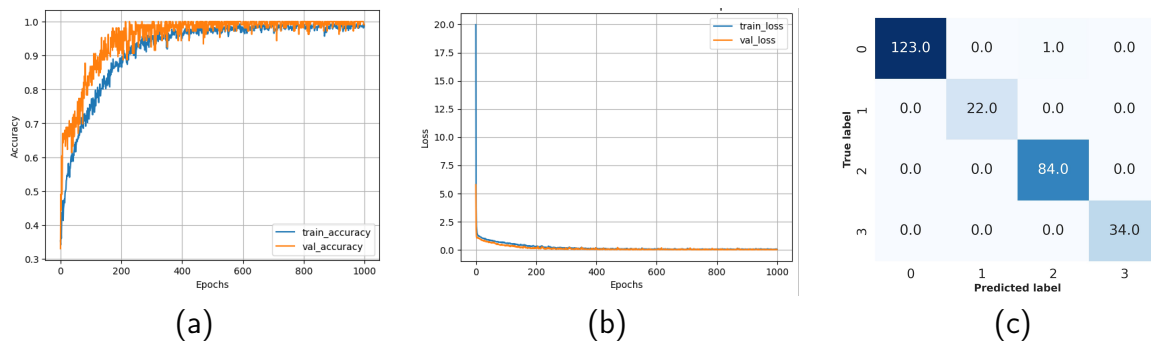


Figure 6-7: Performance of the proposed model (ViT-SLS) on D3S dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.

6.4.2 Experimental results on D3S dataset

For the D3S dataset, it has been observed that the standard Vision Transformer gives 76.14% testing accuracy when trained for 1000 epochs. Testing accuracy improved to 99.4% when trained using a Vision Transformer with shifted patch tokenization and locality self-attention (ViT-SL), and it reached 100% when trained on the proposed model (ViT-SLS) which incorporates additional stochastic depth for regularization and to converge towards global minima faster. When the model proposed by Lee et al. [142] is evaluated on the D3S dataset, the testing accuracy reaches 99.4% in 354 epochs, whereas the proposed model (ViT-SLS) achieves 99.7% testing accuracy in just 215 epochs which shows that adding stochastic depth reduces the convergence time. It is clearly derived from the results shown in Figure 6-5, Fig. 6-6, and Figure 6-7 that the fused model (ViT-SLS) outperforms the base models ViT [140] and ViT-SL [142]. In Figure 6-5c), Figure 6-6c), and Figure 6-7c), 0, 1, 2, 3 on the x and y-axis represents “eye-close”, “happy”, “yawn” and “neutral” emotions respectively.

6.4.3 Experimental results on KMU-FED dataset

For the KMU-FED dataset, the performance of the general vision Transformer is manifested in Figure 6-8, which has achieved 99.4% accuracy on the testing dataset. Whereas, the ViT-SLS has recognized the emotions of the testing dataset with 99.9% accuracy as shown in Figure 6-9. General ViT misinterpreted 3 surprise emotion images as fear emotion and 1 fear image as happy emotion which declines its average accuracy. Surprise emotion is generally misinterpreted as fear emotion because they share some common action units (AU1, AU2, AU26) [147]. But ViT-SLS misinterpreted only a single fear image as a happy emotion and achieved 99.9% accuracy which depicts that the proposed model can accurately recognize the driver’s emotions in the real-time framework. In Figure 6-8 c) and Figure 6-9 c), 0, 1, 2, 3, 4, 5 on the x and y-axis represent “surprise”, “anger”, “disgust”, “fear”, “happy”, and “sad” emotions respectively.

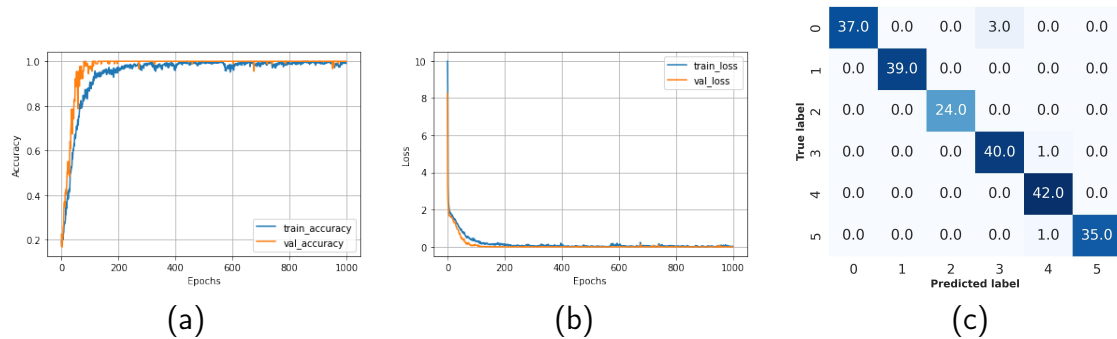


Figure 6-8: Performance of the General Vision Transformer on KMU-FED dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.

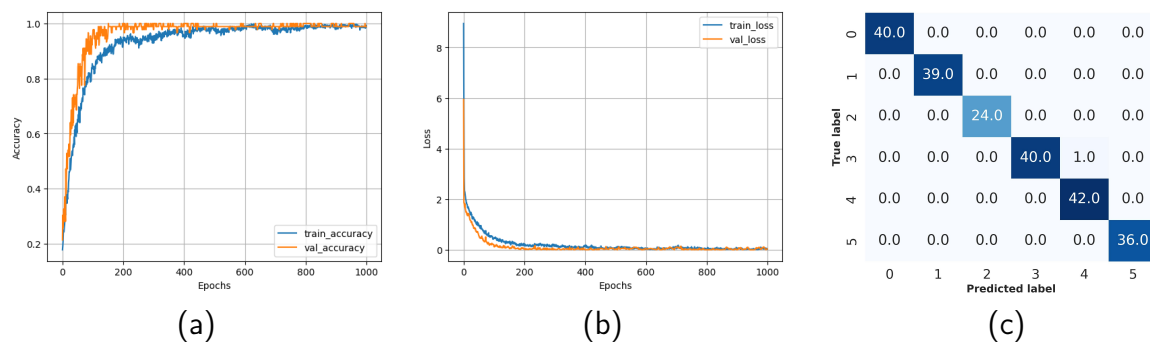


Figure 6-9: Performance of the proposed model (ViT-SLS) on KMU-FED dataset: a) Accuracy vs Epoch graph b) Loss vs Epoch Graph c) Confusion metric.

6.4.4 Experimental results on CK+ and JAFFE datasets

The proposed model is evaluated on two commonly used facial emotion datasets i.e., CK+ [145] and JAFFE [146]. For CK+ datasets, the performance of ViT-SLS on the testing set is 99.49% which is better than the state-of-the-art methods as shown in Table 6.3. And for the JAFFE dataset, ViT-SLS recognize the images of the test set with 93.02% accuracy in a very short time span(minutes). The ViT-SLS model has approximately 42M parameters, which corresponds to a storage size of 162.45 megabytes(MB). Additionally, it requires an epoch runtime of 927 milliseconds(ms). Class-wise accuracy of each dataset is depicted in Figure 6-10.

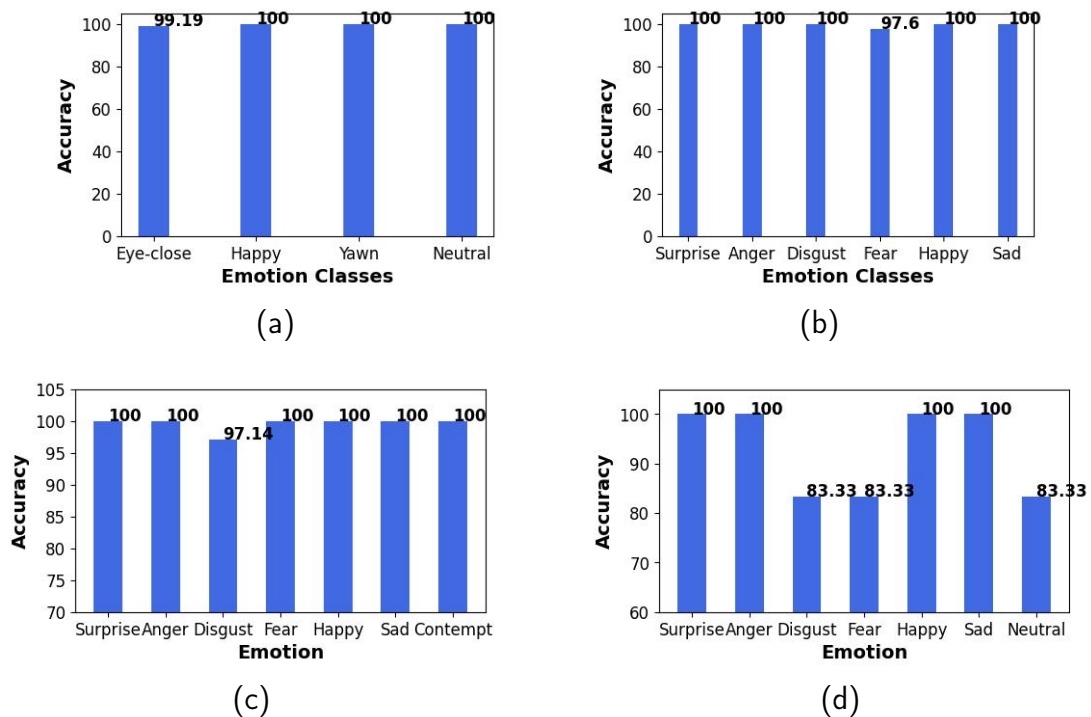


Figure 6-10: Class-wise accuracy of ViT-SLS on (a) D3S, (b) KMU-FED, (c) CK+, and (d) JAFFE.

6.4.5 Ablation Study

This section contains a detailed study of how the addition of a stochastic depth component along with shifted patch tokenization and locality self-attention [142] to the Vision Transformer [140] outperforms the state-of-the-art methods on D3S and KMU-FED datasets. Stochastic depth introduces regularization by skipping a set of layers during the training of the model. Due to the availability of small scale driver datasets, shifted patch tokenization is applied, which extracts the rich spatial information and locality self-attention, neglecting the self-token relations and highlighting the inter-token ones. The fused model ViT-SLS outperforms the base models ViT [140] producing 76.14% testing accuracy and ViT-SL [142] producing 98.2% on the D3S dataset. Adding the stochastic depth component increases the testing accuracy by 1.5% which makes it 99.7% accurate for the driver’s emotion recognition. Similar experiments are performed on the KMU-FED dataset to evaluate the proposed model which shows 0.4% improvement in the testing accuracy and eventually predicts the

Table 6.1: Comparative analysis of state-of-the-art methods and proposed method on KMU-FED dataset.

Model	Accuracy
SqueezeNet [135]	89.7%
MobileNetV2 [135]	93.8%
MobileNetV3 [135]	94.9%
Lightweight Multilayer Random Forests [135]	95.1%
SqueezeNet 1.1 [148]	95.83%
MTCNN [134]	97.3%
CNN + SVM [149]	98.64%
VGG19 [150]	99.7%
ViT-SLS	99.9%

Table 6.2: Comparative analysis of state-of-the-art methods and proposed method on D3S dataset.

Model	Accuracy
SVM [141]	90%
CNN [151]	97%
ViT-SLS	99.7%

driver’s emotions with 99.9% accuracy. Moreover, the performance of these models is depicted in Figure 6-11

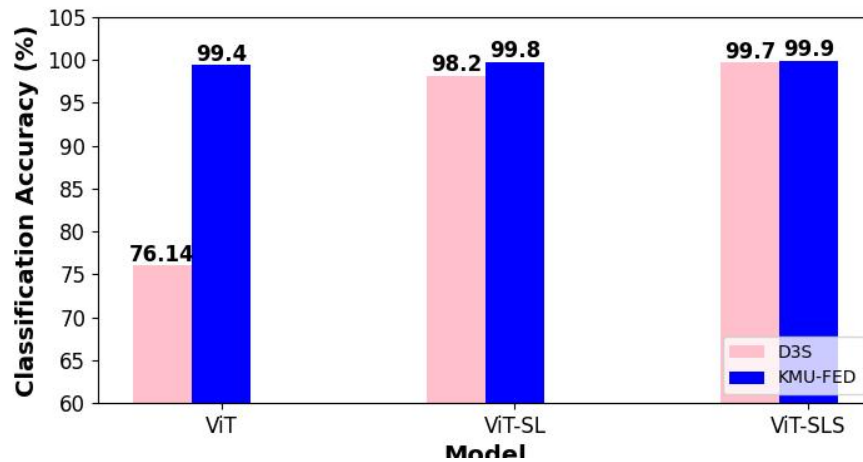


Figure 6-11: Ablation study on D3S and KMU-FED datasets.

Table 6.3: Comparative analysis of state-of-the-art methods and proposed method on CK+ dataset.

Model	Accuracy
Deep CNN [152]	93.24%
VGG16 [153]	94.8%
CNN + SVM [149]	95.05%
Swin Transformer [99]	95.52%
CNN [154]	97.38%
CNN with data augmentation [155]	97.69%
ACNN [156]	98.0%
ConvNet [157]	98.38%
Semi-Supervised Deep Belief Network [158]	98.57%
CoT_AdaptiveViT [159]	99.20%
ViT-SLS	99.49%

Table 6.4: Comparative analysis of state-of-the-art methods and proposed method on JAFFE dataset.

Model	Accuracy
Swin Transformer [99]	73.17%
ConvNet [157]	77.14%
ACNN [156]	92.8%
CNN [160]	92.83%
ResNet50 [161]	92.86%
Face-STN [162]	93.40%
VGG16 [153]	93.7%
ViT-SLS	93.02%

6.4.6 Comparative study of ViT-SLS with state-of-the-art

In this section, the proposed method is compared with the state-of-the-art methods that came across the literature. For KMU-FED, Table 6.1 shows that ViT-SLS achieved 99.9% accuracy which is the highest among other proposed methods in the literature. Whereas, other proposed methods [135], [134], [148] have achieved accuracy below 97.5% only. Similarly, for the D3S dataset, Table 6.2 shows that ViT-SLS achieved the gain of 2.7% over the comparative methodologies which makes it 99.7% accurate for driver’s emotion recognition.

Li et al. [154] used CNN for facial emotion recognition and achieved 97.38% but it is further increased by augmenting the dataset as CK+ dataset is an imbalanced dataset. Umer et al. [155] used CNN with data augmentation and improved the accuracy to 97.69%. But ViT-SLS achieves 99.49% testing accuracy which shows that

ViT-SLS outperforms other state-of-the-art methods on the CK+ dataset. Xiong et al [159] proposed CoT_AdaptiveViT using contextual transformer between CNN and ViT and achieved 99.20% on CK+ dataset. Minaee et al. [156] proposed an attentional convolutional network(ACNN) to recognize facial emotions and used CK+, JAFFE, and FER-2013 datasets to evaluate the proposed model. ViT-SLS has obtained competitive results on the JAFFE dataset as well. Barros et al. [162] used a Spatial transformer network (Face-STN) made up of set of convolutional layers and obtained 93.40% accuracy on the JAFFE dataset.

6.5 Conclusion

In order to accurately classify the driver's emotions in a real-time environment, this chapter has presented a computationally effective driver's emotion recognition system. Vision Transformer based method is proposed which can be trained on small-size datasets. ViT-SLS consists of shifted patch tokenization to extract the extensive spatial information, locality self-attention to highlight the inter-token relations, and stochastic depth for regularization. The optimal implementation of the fused model (ViT-SLS) on D3S (100%), KMU-FED (99.9%), CK+ (99.49%), and JAFFE (93.02%) datasets has given promising results.

Chapter 7

Conclusion & Future Directions

The facial emotion recognition system is a powerful tool to have notable communication and can improve human-computer interaction as well. Facial expression plays an important role in analyzing human emotions. As humans perceive information about other person based on his body language, facial expressions, and gestures. In the same way, computers can be trained to recognize emotions by observing facial expressions. Many researchers have contributed their work in this area but as conclusion, it has been observed that deep learning algorithms outperform traditional machine learning algorithms [163]. During experimentation, different challenges arise due to the small size of datasets, imbalanced datasets, different illumination conditions.

7.1 Summary and Contribution of the Thesis

In this thesis, models are proposed primarily for facial emotion recognition. Four frameworks have been proposed for evaluation for attention mechanisms, basic emotion recognition for in-the-wild dataset, compound emotion recognition, and driver's emotion recognition.

- In the first framework, attention mechanisms for facial emotion recognition are mainly focused. Different attention mechanisms are explored to evaluate their efficacy in facial emotion recognition. The capsule neural network-based

model is proposed where each attention mechanism is added separately to pay attention to relevant regions of an image and discard the irrelevant ones. Adding an attention mechanism improves the model’s classification accuracy with less computational overhead.

Capsule neural network-based model is introduced, integrating an attention mechanism to enhance classification accuracy. The work explores the effectiveness of different attention mechanisms in improving facial emotion recognition systems. Specifically, it examines and compares four types of attention mechanisms: channel attention, spatial attention, CBAM attention, self-attention, and multi-head attention (MHA). Experiments were conducted using a CapsNet architecture on an in-the-wild facial emotion recognition dataset. The findings highlight the unique contributions of each attention mechanism in enhancing recognition performance across various facial expressions.

We witness that attention may not focus on the active facial regions when face is occluded or pose variant. Thus, in the next framework we focus on the in-the-wild emotions where occlusion and pose variant are main concerns.

- In the second framework, various challenges in facial expression recognition (FER), such as pose variation, occlusion, and illumination changes, are effectively addressed. To enhance feature extraction, FaceMesh from MediaPipe is used to capture geometric features from in-the-wild images. These relation-aware geometric features help mitigate issues related to occlusion and pose variations to a significant extent. Additionally, a capsule neural network is employed, incorporating dynamic routing between capsules to encode intermediate features more effectively.

A robust Facial Expression Recognition (FER) method is introduced to address challenges in in-the-wild images caused by pose variations and occlusions. The proposed approach, FMR-CapsNet, leverages the FaceMesh model for geometric feature extraction, using facial blendshape scores to effectively capture features from side-facing or occluded images. The novelty of this work is use of

relation-aware geometric features for facial emotion recognition which improved the classification accuracy under unconstrained environment as well. For this, a distance matrix is constructed based on the euclidean distance matrix to represent relations between blendshape scores, providing relative information. To enhance these features further, transfer learning is applied using a pretrained ResNet50 on the distance matrix. Additionally, a capsule neural network is utilized to capture both directional and spatial information, improving inter-class feature differentiation and overall accuracy.

Basic emotions do not fully encompass the complexity of human emotions, as real-life emotional states frequently involve a blend of multiple emotions. By targeting compound emotions, the model can better capture the intricacies of human emotional experiences. So, the next framework is designed to classify compound emotions which are complex versions and composite of basic emotions.

- In the third framework, we propose a model for compound emotion recognition, an area that remains relatively under explored in research. Given the limited dataset size, the CutMix augmentation technique is applied to increase the training samples, enhancing the model’s learning capability. A lightweight Swin Transformer with a CBAM attention mechanism is used to improve the representation of relevant features. Since available datasets are highly imbalanced, class weights are introduced to address this issue, assigning higher weights to minority classes to reduce the model’s mis-classification rate.

A model is designed for classifying compound emotions. To tackle the challenge of imbalanced datasets, the model uses the CutMix augmentation technique for data augmentation. It integrates the CBAM attention mechanism to highlight relevant image features and employs a simplified Swin Transformer with fewer blocks, reducing computational complexity and the number of trainable parameters while enhancing classification accuracy. The novelty of this work is using a modified Swin Transformer with CutMix and CBAM attention which

results into a lightweight and efficient model. Experimental results on the RAF-DB and EmotioNet datasets demonstrate that this transformer-based network effectively recognizes compound emotions.

As a conclusive part of this thesis, the next framework is based on one of the applications of Facial Emotion Recognition (FER) by developing a driver’s emotion recognition system.

- In the fourth framework, we explored an application of facial emotion recognition for detecting driver emotions. We propose an efficient, lightweight Vision Transformer-based model (ViT-SLS) that accurately classifies driver emotions in real-time settings. ViT-SLS incorporates shifted patch tokenization to capture detailed spatial information and locality self-attention to emphasize inter-feature relationships. Given the limited size of driver emotion datasets, stochastic depth regularization is applied in the proposed model to enhance robustness.

ViT-SLS employs shifted patch tokenization, where input images are shifted and divided into patches, processed through a locality-self attention module to improve performance on limited data. Stochastic depth is incorporated for regularization, enabling robust performance across images of varying resolutions. The proposed model was evaluated on four datasets—D3S, KMU-FED, CK+, and JAFFE—achieving remarkable accuracies, demonstrating its effectiveness in recognizing driver emotions.

7.2 Future Directions

- There is potential for future research to further enhance the robustness of emotion recognition systems, particularly in addressing the class-wise accuracy challenges posed by highly imbalanced datasets.
- While the model performs admirably across various conditions, ongoing efforts can focus on optimizing its handling of blur and partial-face images to broaden

its applicability in real-world settings.

- There are very less facial datasets available for compound emotions which limits the training of the network. As deep neural networks are data-hungry, the more the training data is provided, the more accurately it would classify the compound emotions. So, there is a need for balanced and large datasets for compound emotions which can improve the learning process and further improve the classification rate as well.
- There is still scope for improvements that can be considered for future work in the area of compound emotion recognition.
- As driver's emotion datasets are small-scale datasets, it limits their application on deep-learning models which require large-scale datasets. Application-specific, large-scale datasets can be developed to enable the implementation of advanced transformer-based models for improved feature extraction and representation.
- Most of the available datasets for FER are imbalanced which poses a challenge to the implementation of FER. Therefore, there is a need for techniques that can address the issue of dataset imbalance more effectively.

References

- [1] M. Shiffrar, M. D. Kaiser, and A. Chouhourelou, “Seeing human movement as inherently social,” *The science of social vision*, pp. 248–264, 2011.
- [2] P. Ekman, “Are there basic emotions?” 1992.
- [3] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.
- [4] A. G. Reece and C. M. Danforth, “Instagram photos reveal predictive markers of depression,” *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017.
- [5] L. Manikonda and M. De Choudhury, “Modeling and understanding visual attributes of mental health disclosures in social media,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 170–181.
- [6] P. Kindness, J. Masthoff, and C. Mellish, “Designing emotional support messages tailored to stressors,” *International Journal of Human-Computer Studies*, vol. 97, pp. 1–22, 2017.
- [7] R. e. KALIOUBY, R. Picard, and S. Baron-Cohen, “Affective computing and autism,” *Annals of the New York Academy of Sciences*, vol. 1093, no. 1, pp. 228–248, 2006.
- [8] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder,”

-
- International journal of human-computer studies*, vol. 66, no. 9, pp. 662–677, 2008.
- [9] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, “Arbee: Towards automated recognition of bodily expression of emotion in the wild,” *International journal of computer vision*, vol. 128, no. 1, pp. 1–25, 2020.
- [10] A. Mostafa, M. I. Khalil, and H. Abbas, “Emotion recognition by facial features using recurrent neural networks,” in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. IEEE, 2018, pp. 417–422.
- [11] E. A. Boyle, A. H. Anderson, and A. Newlands, “The effects of visibility on dialogue and performance in a cooperative problem solving task,” *Language and speech*, vol. 37, no. 1, pp. 1–20, 1994.
- [12] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, “A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots,” *Neural Computing and Applications*, vol. 29, no. 7, pp. 359–373, 2018.
- [13] K. Kulkarni, C. A. Corneanu, I. Ofodile, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, “Automatic recognition of facial displays of unfelt emotions,” *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 377–390, 2018.
- [14] M. N. Ab Wahab, A. Nazir, A. T. Z. Ren, M. H. M. Noor, M. F. Akbar, and A. S. A. Mohamed, “Efficientnet-lite and hybrid cnn-knn implementation for facial expression recognition on raspberry pi,” *IEEE Access*, vol. 9, pp. 134 065–134 080, 2021.
- [15] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, “Emotion recognition using facial expressions,” *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

-
- [16] F. Z. Salmam, A. Madani, and M. Kissi, “Facial expression recognition using decision trees,” in *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. IEEE, 2016, pp. 125–130.
- [17] K. Bailly, S. Dubuisson *et al.*, “Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests,” *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 167–181, 2017.
- [18] T. Nguyen, I. Bass, M. Li, and I. K. Sethi, “Investigation of combining svm and decision tree for emotion classification,” in *Seventh IEEE International Symposium on Multimedia (ISM’05)*. IEEE, 2005, pp. 5–pp.
- [19] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, “A deep neural network-driven feature learning method for multi-view facial expression recognition,” *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.
- [20] B. Verma and A. Choudhary, “Affective state recognition from hand gestures and facial expressions using grassmann manifolds,” *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 14 019–14 040, 2021.
- [21] —, “Deep learning based real-time driver emotion monitoring,” in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. IEEE, 2018, pp. 1–6.
- [22] M. Das and S. K. Ghosh, “Deep-step: A deep learning approach for spatiotemporal prediction of remote sensing data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1984–1988, 2016.
- [23] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

-
- [24] L. Shao, D. Wu, and X. Li, “Learning deep and wide: A spectral method for learning deep networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2303–2308, 2014.
- [25] Y. Wang, Y. Li, Y. Song, and X. Rong, “The application of a hybrid transfer algorithm based on a convolutional neural network model and an improved convolution restricted boltzmann machine model in facial expression recognition,” *IEEE Access*, vol. 7, pp. 184 599–184 610, 2019.
- [26] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, “An efficient deep learning technique for facial emotion recognition,” *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 1649–1683, 2022.
- [27] M. Bentoumi, M. Daoud, M. Benaouali, and A. Taleb Ahmed, “Improvement of emotion recognition from facial images using deep learning and early stopping cross validation,” *Multimedia Tools and applications*, pp. 1–31, 2022.
- [28] J. Zhi, T. Song, K. Yu, F. Yuan, H. Wang, G. Hu, and H. Yang, “Multi-attention module for dynamic facial emotion recognition,” *Information*, vol. 13, no. 5, p. 207, 2022.
- [29] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Benamoun, “Dynamic facial expression recognition under partial occlusion with optical flow reconstruction,” *IEEE Transactions on Image Processing*, vol. 31, pp. 446–457, 2021.
- [30] D. Kamińska, K. Aktas, D. Rizhinashvili, D. Kuklyanov, A. H. Sham, S. Escalera, K. Nasrollahi, T. B. Moeslund, and G. Anbarjafari, “Two-stage recognition and beyond for compound facial emotion recognition,” *Electronics*, vol. 10, no. 22, p. 2847, 2021.
- [31] H. Zhang, W. Su, J. Yu, and Z. Wang, “Identity–expression dual branch network for facial expression recognition,” *IEEE transactions on cognitive and developmental systems*, vol. 13, no. 4, pp. 898–911, 2020.

-
- [32] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, “Deep convolution network based emotion analysis towards mental health care,” *Neurocomputing*, vol. 388, pp. 212–227, 2020.
- [33] D. Sen, S. Datta, and R. Balasubramanian, “Facial emotion classification using concatenated geometric and textural features,” *Multimedia Tools and Applications*, vol. 78, no. 8, pp. 10 287–10 323, 2019.
- [34] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, “Deep spatial-temporal feature fusion for facial expression recognition in static images,” *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.
- [35] J. A. Aghamaleki and V. Ashkani Chenarlogh, “Multi-stream cnn for facial expression recognition in limited training data,” *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22 861–22 882, 2019.
- [36] F. Xu, J. Zhang, and J. Z. Wang, “Microexpression identification and categorization using a facial dynamics map,” *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.
- [37] N. Perveen, D. Roy, and C. K. Mohan, “Spontaneous expression recognition using universal attribute model,” *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5575–5584, 2018.
- [38] A. I. Jabbooree, L. M. Khanli, P. Salehpour, and S. Pourbahrami, “A novel facial expression recognition algorithm using geometry β -skeleton in fusion based on deep cnn,” *Image and Vision Computing*, vol. 134, p. 104677, 2023.
- [39] X. Jin, Z. Lai, and Z. Jin, “Learning dynamic relationships for facial expression recognition based on graph convolutional network,” *IEEE Transactions on Image Processing*, vol. 30, pp. 7143–7155, 2021.
- [40] G. Wen, T. Chang, H. Li, and L. Jiang, “Dynamic objectives learning for facial expression recognition,” *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2914–2925, 2020.

-
- [41] D. H. Nguyen, S. Kim, G.-S. Lee, H.-J. Yang, I.-S. Na, and S. H. Kim, "Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks," *IEEE Transactions on Affective Computing*, 2019.
- [42] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE access*, vol. 7, pp. 41 273–41 285, 2019.
- [43] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Transactions on affective computing*, vol. 12, no. 2, pp. 544–550, 2018.
- [44] K. Fujii, D. Sugimura, and T. Hamamoto, "Hierarchical group-level emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3892–3906, 2020.
- [45] 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.
- [46] S. Xie and H. Hu, "Facial expression recognition with fir-cnn," *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.
- [47] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster r-cnn," *Procedia Computer Science*, vol. 107, pp. 135–140, 2017.
- [48] J. Kim, J.-K. Kang, and Y. Kim, "A resource efficient integer-arithmetic-only fpga-based cnn accelerator for real-time facial emotion recognition," *IEEE Access*, vol. 9, pp. 104 367–104 381, 2021.
- [49] C. Liu, K. Hirota, and Y. Dai, "Patch attention convolutional vision transformer for facial expression recognition with occlusion," *Information Sciences*, vol. 619, pp. 781–794, 2023.

- [50] F. Ma, B. Sun, and S. Li, “Facial expression recognition with visual transformers and attentional selective fusion,” *IEEE Transactions on Affective Computing*, 2021.
- [51] F. Xue, Q. Wang, and G. Guo, “Transfer: Learning relation-aware facial expression representations with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610.
- [52] Q. Huang, C. Huang, X. Wang, and F. Jiang, “Facial expression recognition with grid-wise attention and visual transformer,” *Information Sciences*, vol. 580, pp. 35–54, 2021.
- [53] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” in *Workshop on faces in ‘Real-Life’ Images: detection, alignment, and recognition*, 2008.
- [54] Y. Zhou, L. Jin, H. Liu, and E. Song, “Color facial expression recognition by quaternion convolutional neural network with gabor attention,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 4, pp. 969–983, 2020.
- [55] D. Kollias, “Multi-label compound expression recognition: C-expr database & network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5589–5598.
- [56] S. K. Jarraya, M. Masmoudi, and M. Hammami, “Compound emotion recognition of autistic children during meltdown crisis based on deep spatio-temporal analysis of facial geometric features,” *IEEE Access*, vol. 8, pp. 69 311–69 326, 2020.
- [57] S. Shaila, D. Shivamma, U. Monica, and K. Tejashree, “Facial expression recognition for compound emotions using mobile net architecture,” in *2022 Inter-*

- national Conference on Artificial Intelligence and Data Engineering (AIDE)*.
IEEE, 2022, pp. 187–190.
- [58] L. Liang, C. Lang, Y. Li, S. Feng, and J. Zhao, “Fine-grained facial expression recognition in the wild,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 482–494, 2020.
- [59] A. Tawari and M. M. Trivedi, “Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos,” in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 344–349.
- [60] A. Tavakoli, S. Boker, and A. Heydarian, “Driver state modeling through latent variable state space framework in the wild,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [61] L. Yang, H. Yang, B.-B. Hu, Y. Wang, and C. Lv, “A robust driver emotion recognition method based on high-purity feature separation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 15 092–15 104, 2023.
- [62] K. Zaman, S. Zhaoyun, B. Shah, T. Hussain, S. M. Shah, F. Ali, and U. S. Khan, “A novel driver emotion recognition system based on deep ensemble classification,” *Complex & Intelligent Systems*, vol. 9, no. 6, pp. 6927–6952, 2023.
- [63] L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain, and W. Gao, “Driver emotion recognition with a hybrid attentional multimodal fusion framework,” *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2970–2981, 2023.
- [64] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

- [65] H. Mittal and B. Verma, “Cat-capsnet: A convolutional and attention based capsule network to detect the driver’s distraction,” *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [68] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.
- [69] E. Friesen and P. Ekman, “Facial action coding system: a technique for the measurement of facial movement,” *Palo Alto*, vol. 3, no. 2, p. 5, 1978.
- [70] J. C. Hager, P. Ekman, and W. V. Friesen, “Facial action coding system,” *Salt Lake City, UT: A Human Face*, p. 8, 2002.
- [71] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
- [72] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.
- [73] T. Liu, J. Li, J. Wu, B. Du, J. Wan, and J. Chang, “Confusable facial expression recognition with geometry-aware conditional network,” *Pattern Recognition*, vol. 148, p. 110174, 2024.

- [74] M. A. Solis-Arrazola, R. E. Sanchez-Yañez, C. H. Garcia-Capulin, and H. Rostro-Gonzalez, “Enhancing image-based facial expression recognition through muscle activation-based facial feature extraction,” *Computer Vision and Image Understanding*, vol. 240, p. 103927, 2024.
- [75] C. Yu, D. Zhang, W. Zou, and M. Li, “Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition,” *IEEE Transactions on Affective Computing*, 2024.
- [76] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.
- [77] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” *Advances in neural information processing systems*, vol. 30, 2017.
- [78] Y. Dong, Y. Fu, L. Wang, Y. Chen, Y. Dong, and J. Li, “A sentiment analysis method of capsule network based on bilstm,” *IEEE Access*, vol. 8, pp. 37 014–37 020, 2020.
- [79] W. Wang, F. Lee, S. Yang, and Q. Chen, “An improved capsule network based on capsule filter routing,” *IEEE Access*, vol. 9, pp. 109 374–109 383, 2021.
- [80] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, “Real-time facial surface geometry from monocular video on mobile gpus,” *arXiv preprint arXiv:1907.06724*, 2019.
- [81] 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.
- [82] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14124313>

- [83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [84] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region attention networks for pose and occlusion robust facial expression recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.
- [85] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, “Suppressing uncertainties for large-scale facial expression recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.
- [86] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, “Relation-aware facial expression recognition,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1143–1154, 2021.
- [87] S. Saurav, R. Saini, and S. Singh, “Emnet: a deep integrated convolutional neural network for facial emotion recognition in the wild,” *Applied Intelligence*, vol. 51, no. 8, pp. 5543–5570, 2021.
- [88] A. H. Farzaneh and X. Qi, “Facial expression recognition in the wild via deep attentive center loss,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2402–2411.
- [89] Y. Zhang, C. Wang, and W. Deng, “Relative uncertainty learning for facial expression recognition,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.
- [90] C. Su, J. Wei, D. Lin, and L. Kong, “Using attention lsgb network for facial expression recognition,” *Pattern Analysis and Applications*, pp. 1–11, 2022.
- [91] H. Liu, H. Cai, Q. Lin, X. Li, and H. Xiao, “Adaptive multilayer perceptual attention network for facial expression recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6253–6266, 2022.

- [92] F. Wan and R. Zhi, “Gaussian distribution-based facial expression feature extraction network,” *Pattern Recognition Letters*, vol. 164, pp. 104–111, 2022.
- [93] E. Ryumina, D. Dresvyanskiy, and A. Karpov, “In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study,” *Neurocomputing*, vol. 514, pp. 435–450, 2022.
- [94] H. Gao, M. Wu, Z. Chen, Y. Li, X. Wang, S. An, J. Li, and C. Liu, “Ssa-icl: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition,” *Neural Networks*, vol. 158, pp. 228–238, 2023.
- [95] F. Zhang, G. Chen, H. Wang, and C. Zhang, “Cf-dan: Facial-expression recognition based on cross-fusion dual-attention network,” *Computational Visual Media*, pp. 1–16, 2024.
- [96] X. Zhang, J. Zhu, D. Wang, Y. Wang, T. Liang, H. Wang, and Y. Yin, “A gradual self distillation network with adaptive channel attention for facial expression recognition,” *Applied Soft Computing*, p. 111762, 2024.
- [97] A. Bellocchi, “Methods for sociological inquiry on emotion in educational settings,” *Emotion Review*, vol. 7, no. 2, pp. 151–156, 2015.
- [98] C. Loob, P. Rasti, I. Lüsi, J. C. Jacques, X. Baró, S. Escalera, T. Sapinski, D. Kaminska, and G. Anbarjafari, “Dominant and complementary multi-emotional facial expression recognition using c-support vector classification,” in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 833–838.
- [99] 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,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [100] P. Ekman, “Facial expression and emotion.” *American psychologist*, vol. 48, no. 4, p. 384, 1993.

-
- [101] M. G. Calvo, A. Fernández-Martín, A. Gutiérrez-García, and D. Lundqvist, “Selective eye fixations on diagnostic face regions of dynamic emotional expressions: Kdef-dyn database,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.
- [102] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, “Cross-subject multimodal emotion recognition based on hybrid fusion,” *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.
- [103] M. Alam, L. S. Vidyaratne, and K. M. Iftekharuddin, “Sparse simultaneous recurrent deep learning for robust facial expression recognition,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 10, pp. 4905–4916, 2018.
- [104] Nidhi and B. Verma, “From methods to datasets: a detailed study on facial emotion recognition,” *Applied Intelligence*, vol. 53, no. 24, pp. 30 219–30 249, 2023.
- [105] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baro, H. Demirel *et al.*, “Dominant and complementary emotion recognition from still images of faces,” *IEEE Access*, vol. 6, pp. 26 391–26 403, 2018.
- [106] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, “Multi-modality network with visual and geometrical information for micro emotion recognition,” in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 814–819.
- [107] G. Pons and D. Masip, “Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition,” *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 4764–4771, 2020.
- [108] Z. Lian, L. Sun, H. Sun, K. Chen, Z. Wen, H. Gu, B. Liu, and J. Tao, “Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition,” *Information Fusion*, vol. 108, p. 102367, 2024.

-
- [109] A. Zhu, K. Li, T. Wu, P. Zhao, and B. Hong, “Cross-task multi-branch vision transformer for facial expression and mask wearing classification,” *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 1, p. 46–53, Apr. 2024. [Online]. Available: <https://www.suaspress.org/ojs/index.php/JCTAM/article/view/v1n1a07>
- [110] R. Dong and K.-M. Lam, “Bi-center loss for compound facial expression recognition,” *IEEE Signal Processing Letters*, 2024.
- [111] X. Zhao, Y. Lv, and Z. Huang, “Multimodal fusion-based swin transformer for facial recognition micro-expression recognition,” in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2022, pp. 780–785.
- [112] F. Xue, Z. Tan, Y. Zhu, Z. Ma, and G. Guo, “Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2412–2418.
- [113] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [114] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [115] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations*, 2018, pp. 1–13.
- [116] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *Proceedings*

- of the *IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.
- [117] J. Jiang, M. Wang, B. Xiao, J. Hu, and W. Deng, “Joint recognition of basic and compound facial expressions by mining latent soft labels,” *Pattern Recognition*, vol. 148, p. 110173, 2024.
- [118] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [119] K. Kim, B. Ji, D. Yoon, and S. Hwang, “Self-knowledge distillation with progressive refinement of targets,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6567–6576.
- [120] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [121] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes,” *Advances in neural information processing systems*, vol. 34, pp. 3965–3977, 2021.
- [122] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, “Pyramid with super resolution for in-the-wild facial expression recognition,” *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.
- [123] S. Gopalakrishnan, “A public health perspective of road traffic accidents,” *Journal of family medicine and primary care*, vol. 1, no. 2, p. 144, 2012.
- [124] C. E. Izard, “Emotion theory and research: Highlights, unanswered questions, and emerging issues,” *Annual review of psychology*, vol. 60, pp. 1–25, 2009.
- [125] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien, “Emotion on the road: necessity, acceptance, and feasibility

- of affective computing in the car,” *Advances in human-computer interaction*, vol. 2010, pp. 1–17, 2010.
- [126] D. Tran, J. Du, W. Sheng, D. Osipychov, Y. Sun, and H. Bai, “A human-vehicle collaborative driving framework for driver assistance,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3470–3485, 2018.
- [127] L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda, “Analysis of real-world driver’s frustration,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 109–118, 2010.
- [128] H. Xiao, W. Li, G. Zeng, Y. Wu, J. Xue, J. Zhang, C. Li, and G. Guo, “On-road driver emotion recognition using facial expression,” *Applied Sciences*, vol. 12, no. 2, p. 807, 2022.
- [129] G. Rebolledo-Mendez, A. Reyes, S. Paszkowicz, M. C. Domingo, and L. Skrypchuk, “Developing a body sensor network to detect emotions during driving,” *IEEE transactions on intelligent transportation systems*, vol. 15, no. 4, pp. 1850–1854, 2014.
- [130] X. Wang, Y. Guo, C. Bai, Q. Yuan, S. Liu, and J. Han, “Driver’s intention identification with the involvement of emotional factors in two-lane roads,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6866–6874, 2020.
- [131] M. Jeong and B. C. Ko, “Driver’s facial expression recognition in real-time for safe driving,” *Sensors*, vol. 18, no. 12, p. 4270, 2018.
- [132] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, “A convolution bidirectional long short-term memory neural network for driver emotion recognition,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4570–4578, 2020.
- [133] W. Li, G. Zeng, J. Zhang, Y. Xu, Y. Xing, R. Zhou, G. Guo, Y. Shen, D. Cao, and F.-Y. Wang, “Cogemonet: A cognitive-feature-augmented driver emotion

- recognition model for smart cockpit,” *IEEE Transactions on Computational Social Systems*, vol. 9, no. 3, pp. 667–678, 2021.
- [134] J. Zhang, X. Mei, H. Liu, S. Yuan, and T. Qian, “Detecting negative emotional stress based on facial expression in real time,” in *2019 IEEE 4th international conference on signal and image processing (ICSIP)*. IEEE, 2019, pp. 430–434.
- [135] M. Jeong, J. Nam, and B. C. Ko, “Lightweight multilayer random forests for monitoring driver emotional status,” *Ieee Access*, vol. 8, pp. 60 344–60 354, 2020.
- [136] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [137] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [138] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [139] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, “Big transfer (bit): General visual representation learning,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.
- [140] 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.
- [141] I. Gupta, N. Garg, A. Aggarwal, N. Nepalia, and B. Verma, “Real-time driver’s drowsiness monitoring based on dynamically varying threshold,” in *2018*

-
- Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, 2018, pp. 1–6.
- [142] S. H. Lee, S. Lee, and B. C. Song, “Vision transformer for small-size datasets,” *arXiv preprint arXiv:2112.13492*, 2021.
- [143] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 646–661.
- [144] Z. Zhuang, M. Liu, A. Cutkosky, and F. Orabona, “Understanding adamw through proximal methods and scale-freeness,” *Transactions on machine learning research*, 2022. [Online]. Available: <https://par.nsf.gov/biblio/10396293>
- [145] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, “Automated face analysis by feature point tracking has high concurrent validity with manual faces coding,” *Psychophysiology*, vol. 36, no. 1, pp. 35–43, 1999.
- [146] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.
- [147] C.-H. Hjortsjö, *Man’s face and mimic language*. Studentlitteratur Lund, Sweden, 1970.
- [148] G. K. Sahoo, S. K. Das, and P. Singh, “Deep learning-based facial emotion recognition for driver healthcare,” in *2022 National Conference on Communications (NCC)*, 2022, pp. 154–159.
- [149] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, “A hybrid model for driver emotion detection using feature fusion approach,” *International journal of environmental research and public health*, vol. 19, no. 5, p. 3085, 2022.

-
- [150] G. K. Sahoo, S. K. Das, and P. Singh, "Performance comparison of facial emotion recognition: A transfer learning-based driver assistance framework for in-vehicle applications," *Circuits, Systems, and Signal Processing*, pp. 1–28, 2023.
- [151] V. VIJAYPRIYA and M. UMA, "An effective hybrid features for driver fatigue detection using convolutional neural network," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 4, 2023.
- [152] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019.
- [153] A. K. Dubey and V. Jain, "Automatic facial recognition using vgg16 based transfer learning model," *Journal of Information and Optimization Sciences*, vol. 41, no. 7, pp. 1589–1596, 2020.
- [154] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *The visual computer*, vol. 36, no. 2, pp. 391–404, 2020.
- [155] S. Umer, R. K. Rout, C. Pero, and M. Nappi, "Facial expression recognition with trade-offs between data augmentation and deep learning features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 721–735, 2022.
- [156] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.
- [157] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

-
- [158] A. R. Kurup, M. Ajith, and M. M. Ramón, “Semi-supervised facial expression recognition using reduced spatial features and deep belief networks,” *Neurocomputing*, vol. 367, pp. 188–197, 2019.
- [159] L. Xiong, J. Zhang, X. Zheng, and Y. Wang, “Context transformer and adaptive method with visual transformer for robust facial expression recognition,” *Applied Sciences*, vol. 14, no. 4, p. 1535, 2024.
- [160] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos, “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order,” *Pattern recognition*, vol. 61, pp. 610–628, 2017.
- [161] E. Tsalera, A. Papadakis, M. Samarakou, and I. Voyiatzis, “Feature extraction with handcrafted methods and convolutional neural networks for facial emotion recognition,” *Applied Sciences*, vol. 12, no. 17, p. 8455, 2022.
- [162] P. Barros and A. Sciutti, “Across the universe: Biasing facial representations toward non-universal emotions with the face-stn,” *IEEE Access*, vol. 10, pp. 103 932–103 947, 2022.
- [163] G. Pons and D. Masip, “Supervised committee of convolutional neural networks in automated facial expression analysis,” *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343–350, 2017.

Appendix A

Capsule Neural Network

In this appendix, we explain the capsule neural network.

A.1 Capsule Neural Network

A capsule neural network is a variant of a neural network that represents a hierarchical relationship between the extracted features. Convolutional neural networks are unable to recognize object rotations and scaling variations within objects effectively. Additionally, the pooling operations in CNNs can lead to a loss of spatial information. Capsule networks, inspired by the hierarchical structure of the human visual system, aim to address the shortcomings of traditional neural networks by introducing “capsules” as fundamental units of representation.

These capsules not only encode the presence of features but also their instantiation parameters, such as pose and orientation, enabling the network to learn hierarchical relationships more effectively. Unlike the neurons in CNNs, capsules use output vectors that can capture directional information, allowing for more accurate image differentiation.

The architecture of the Capsule neural network is shown in Figure A-1. The general architecture starts with a single convolutional layer but the proposed method utilizes two convolutional layers to detect local patterns from the intermediate features more effectively. The output of the convolutional layer is fed into primary

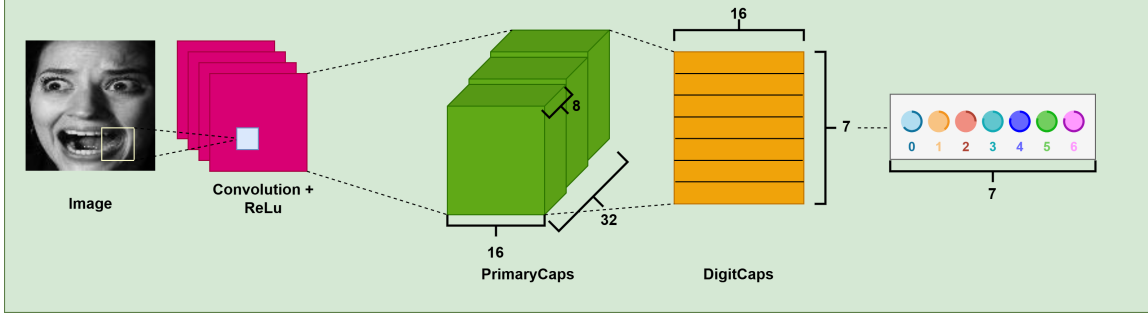


Figure A-1: Capsule neural network

capsules. Each primary capsule represents a higher-level feature detected by the convolutional capsules. Primary capsules are typically arranged spatially, capturing spatial hierarchies of features. Each primary capsule outputs a vector representing the instantiation parameters of a specific feature, such as pose, orientation, and scale. In CapsNet model, the PrimaryCaps layer consists of eight capsules, with each capsule containing sixteen-dimensional features. Additionally, the contribution ($\hat{u}_{j|i}$) of each capsule u_i in the PrimaryCaps layer to that of v_j DigitCaps was computed using Eq. (A.1).

$$\hat{u}_{j|i} = W_{ij} \cdot u_i \quad (\text{A.1})$$

In the DigitCaps layer, there is a 16-dimensional capsule v_j allocated for each digit class (seven classes in this experiment). These capsules obtain input from all capsules in the PrimaryCaps layer using Eq. (A.2), Eq. (A.3), Eq. (A.4).

$$cap_{ij} = \frac{\exp(b_{ij})}{\sum_l \exp(b_{il})} \quad (\text{A.2})$$

$$c_j = \sum_l cap_{ij} \hat{u}_{j|i} \quad (\text{A.3})$$

$$v_j = \frac{\|c_j\|^2}{1 + \|c_j\|^2} \frac{c_j}{\|c_j\|} \quad (\text{A.4})$$

At last, the margin loss is computed for each digit capsule to classify the facial expressions using Eq. (A.5) where $Y_l = 1$ if there is relation 1, $r^+ = 0.9$, $r^- = 0.1$, and $\lambda = 0.5$.

$$D_l = Y_l \max(0, r^+ - \|v_l\|)^2 + \lambda(1 - Y_l) \max(0, \|v_l\| - r^-)^2 \quad (\text{A.5})$$

Appendix B

Vision Transformer

In this appendix, we explain the vision transformer in detail.

B.1 Vision Transformer

Vision Transformer (ViTs) [140] allows models to understand image structure independently as input images are depicted as sequences and predict the class labels for the given image. The input images are processed as a sequence of patches, where each patch has been flattened into a single vector, further concatenated by the channels of all of its pixels before projecting them linearly to the defined dimension of input. ViT has outperformed Convolutional Neural Network (CNN), and employs transformer structure to classify the images [140]. The advanced performance of ViT comes from pre-training the model using a large-scale dataset and its influence is caused by low locality inductive bias.

The architectural overview of the standard ViTs is depicted in Figure B-1 which take input image $x \in R^{H \times W \times C}$ where H, W and C represent the height, width, and number of the channels. ViT first separates the input image into non-overlapping patches and then flattened them to produce vectors sequentially as formulated in Eq. (B.1):

$$P(x) = [x_p^1; x_p^2; x_p^3; \dots; x_p^n], \quad (\text{B.1})$$

Here $x_p^i \in \mathbb{R}^{P^2 \cdot C}$ indicates the i -th flattened vector. N indicates the number of patches as formulated in Eq. (B.2) and P represents patch size.

$$N = HW/P^2, \quad (\text{B.2})$$

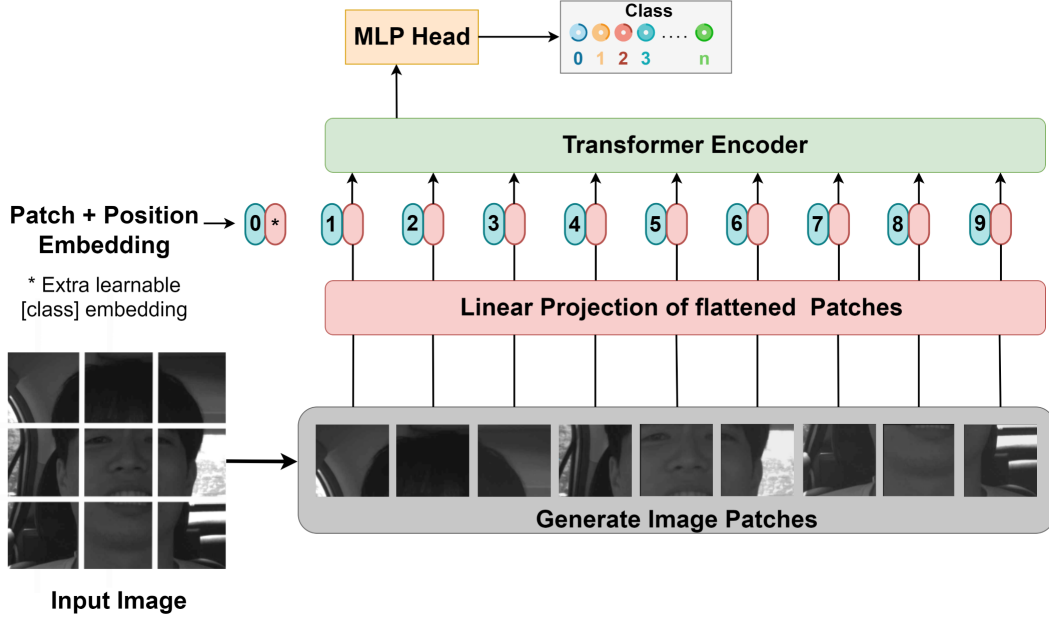


Figure B-1: Model Overview of General Vision Transformer [140]

Patch Embedding is done by applying linear projection to the flattened vectors. Position embeddings are incorporated into the patch embeddings to preserve positional information as shown in Eq. (B.3). Standard learnable single-dimensional embedding is utilized as there is no improvement in the performance of the model using 2D-aware position embeddings. The final sequence of embedding vectors is then fed into the encoder. The transformer encoder of ViT is made up of alternating layers of multi-head self-attention (MSA) and MLP blocks. Layer normalization (LN) is applied before each block, and residual connections are added after each block.

The self-attention mechanism identifies the relationships among encoded patches in the input sequence as shown in Eq. (B.4). For each patch embedding, it calculates a weighted sum of all the patch embeddings, with the weights determined by how relevant each patch is to the current one. This enables the model to prioritize important patches while taking into account both local and global contexts. Multi-

head attention enhances this process by using multiple sets of learnable parameters (attention heads) to extract various types of relationships.

Following the self-attention step, the output from each patch’s self-attention process is sent through a feedforward neural network as shown in Eq. (B.5). This network usually includes a fully connected layer, followed by an activation function such as ReLU (Rectified Linear Unit). The role of the feedforward network is to introduce non-linearity, enabling the model to capture more complex interactions and relationships between patches.

The outputs from both the self-attention mechanism and the feedforward network are followed by layer normalization and residual connections as depicted in Eq. (B.6). Layer normalization ensures more stable and faster training by normalizing the inputs to each sub-layer. Meanwhile, residual connections, or skip connections, add the original input embeddings to the output of each sub-layer, aiding gradient flow during training and mitigating the vanishing gradient issue.

$$\mathbf{z}_0 = \left[\mathbf{e}_{\text{class}}; \mathbf{p}_1^{\text{embed}}; \mathbf{p}_2^{\text{embed}}; \dots; \mathbf{p}_N^{\text{embed}} \right] + \mathbf{P}_{\text{pos}}, \quad \mathbf{P} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad \mathbf{P}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \quad (\text{B.3})$$

$$\mathbf{z}_{\text{attn}}^{(\ell)} = \text{MHA}(\text{Norm}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1, 2, \dots, L \quad (\text{B.4})$$

$$\mathbf{z}_{\text{ff}}^{(\ell)} = \text{MLP}(\text{Norm}(\mathbf{z}_{\text{attn}}^{(\ell)})) + \mathbf{z}_{\text{attn}}^{(\ell)}, \quad \ell = 1, 2, \dots, L \quad (\text{B.5})$$

$$\mathbf{y}_{\text{output}} = \text{Norm}(\mathbf{z}_{(L)}^0) \quad (\text{B.6})$$

Appendix C

Swin Transformer

C.1 Swin Transformer

Swin transformer [99] is developed as a general-purpose transformer-based architecture for vision tasks. It performs local self-attention within distinct windows and attains linear computational complexity. The streamlined architecture of the swin transformer used in the proposed model is illustrated in Figure C-1. Firstly, attentive features extracted from CBAM attention undergo patch partitioning process which splits input images into non-overlapping patches with a patch size of 2×2 . Each patch is considered a "token," where its feature is configured as a concatenation of the pixel values of an RGB image. Each feature with dimension $2 \times 2 \times 3$ will be flattened and converted into a vector of length 12. The vector is projected to an arbitrary dimension (C) using a linear embedding layer. On these patch tokens, two swin transformer blocks (within a single stage only) with modified self-attention mechanisms are used, and maintain the number of tokens ($H/2 \times W/2$). A transformer block with a linear embedding layer is referred to as stage 1 in Figure C-1. The tokens are processed through stage 2 which includes a patch merging layer to produce a reduced number of tokens as the network becomes larger and swin transformer block for feature transformation. Stage 2 is repeated twice to form Stage 3 and Stage 4.

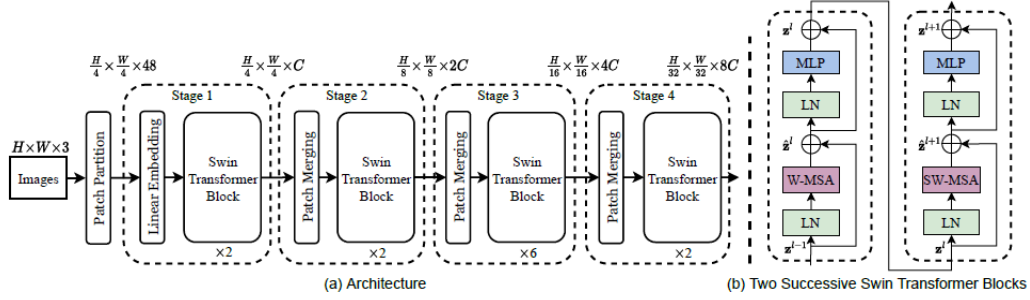


Figure C-1: Swin Transformer architecture

Swin Transformer Block

The Swin Transformer block has a hierarchical structure, which allows it to process images of varying sizes. So, to create the Swin Transformer, the conventional multi-head self-attention (MSA) module in a transformer block is replaced with an SW-MSA module (Mutli-head attention with shifted windows) to compute the self-attention within local windows, while other layers remain the same. It includes shifted-window-based multi-head self-attention and an MLP(2-layer) with a GELU activation function. A normalization layer is added before every MSA and MLP component, and these components are followed by the residual connection. The output of the Swin Transformer block is then passed to the patch merging layer in the network. By stacking two swin transformer blocks on top of each other, the model can learn increasingly complex representations of the input image.

In SW-MSA, the windows have been organized for image partitioning so that the partitions don't overlap with each other. This mechanism improves the computation complexity as shown in Eq (C.1). Suppose the window has $W \times W$ patches, the complexity of window-based self-attention and the basic MSA module on a given image of $h \times w$ patches can be defined as:

$$\begin{aligned}\Omega(MSA) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(W - MSA) &= 4hwC^2 + 2W^2hwC,\end{aligned}\tag{C.1}$$

A shifted Window partitioning approach is followed between successive swin transformer blocks to incorporate cross-window links and maintain the computation effi-

ciency of non-overlapping windows at the same time. With this approach, the output of successive swin transformer blocks is defined as:

$$\begin{aligned}
 \hat{O} &= W - MSA(LN(O^{l-1})) + O^{l-1}, \\
 O^l &= MLP(LN(\hat{O}^l)) + \hat{O}^l, \\
 \hat{O}^{l+1} &= SW - MSA(LN(O^l)) + O^l, \\
 O^{l+1} &= MLP(LN(\hat{O}^{l+1})) + \hat{O}^{l+1},
 \end{aligned}
 \tag{C.2}$$

W-MSA denotes window-based self-attention and SW-MSA denotes shifted windows-based multi-head self-attention. In the above Eq (C.2), \hat{O}^l and O^l represent the output of SW-MSA and MLP modules for l block, respectively.

In the Swin Transformer, self-attention is not applied globally as in traditional transformers; instead, it is performed within smaller, localized windows. This approach limits the interaction of tokens to those within the same window, thereby reducing the computational load.

To allow tokens from different windows to interact across layers, the Swin Transformer employs a “shifted” window strategy. After completing self-attention calculations in one layer using localized windows, the windows are shifted for the next layer. As a result, tokens in one layer’s window overlap with tokens from adjacent windows in the following layer, enabling cross-window interaction.

LIST OF PUBLICATION AND THEIR PROOFS

LIST OF JOURNALS

Journal Paper 1:

Nidhi, & Verma, B. (2023). From methods to datasets: a detailed study on facial emotion recognition. *Applied Intelligence*, 53(24), 30219-30249.
DOI: 10.1007/s10489-023-05052-y (Published, IF: 3.4)

Applied Intelligence (2023) 53:30219–30249
<https://doi.org/10.1007/s10489-023-05052-y>



From methods to datasets: a detailed study on facial emotion recognition

Nidhi¹ · Bindu Verma¹

Accepted: 26 September 2023 / Published online: 15 November 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Human ideas and sentiments are mirrored in facial expressions. Facial expression recognition (FER) is a crucial type of visual data that can be utilized to deduce a person's emotional state. It gives the spectator a plethora of social cues, such as the viewer's focus of attention, emotion, motivation, and intention. It's said to be a powerful instrument for silent communication. AI-based facial recognition systems can be deployed at different areas like bus stations, railway stations, airports, or stadiums to help security forces identify potential threats. There has been a lot of research done in this area. But, it lacks a detailed review of the literature that highlights and analyses the previous work in FER (including work on compound emotion and micro-expressions), and a comparative analysis of different models applied to available datasets, further identifying aligned future directions. So, this paper includes a comprehensive overview of different models that can be used in the field of FER and a comparative study of the traditional methods based on hand-crafted feature extraction and deep learning methods in terms of their advantages and disadvantages which distinguishes our work from existing review studies. This paper also brings you to an eye on the analysis of different FER systems, the performance of different models on available datasets, evaluation of the classification performance of traditional and deep learning algorithms in the context of facial emotion recognition which reveals a good understanding of the classifier's characteristics. Along with the proposed models, this study describes the commonly used datasets showing the year-wise performance achieved by state-of-the-art methods which lacks in the existing manuscripts. At last, the authors itemize recognized research gaps and challenges encountered by researchers which can be considered in future research work.

Keywords Facial expression analysis · Emotion recognition · Classification · Deep learning · Facial action unit

1 Introduction

Emotion plays a vital role in human behavior. Body movement and postures can be used to predetermine rich information about a human's status, awareness, intention, and emotional state [1]. Human activity can be recognized from body movements, but emotion recognition is also a constraint. Facial expressions bring out 55% of the conveyed message, which is relatively higher than the part communicated by the combination of voice and language [2]. As humans can read the body language of others, how computers

can be trained to recognize emotional expressions by observing body movements. This question motivates researchers to contribute to facial emotion recognition. Several psychological studies have suggested that features like body movement and postures can be used for emotion recognition [3–6]. Even researchers had also found that the efficiency of emotion recognition was less when face features were used alone. Still, the system performed better when body movements were also used with face features [6].

Research done in social psychology explains that facial expression helps in coordinating the conversation between speaker and listener that can perceive the interest of the listener [7]. Many researchers have developed different models for emotion recognition which are based on a kNN classifier, decision tree, random forest classifier, convolutional neural network, multilayer perceptron, etc. Traditional machine learning techniques such as Bayesian classifier and Support Vector Machine, k-NN, and decision tree do not provide the

✉ Bindu Verma
bindu.cvvision@gmail.com
Nidhi
nidhi.kanwar189@gmail.com

¹ Department of Information Technology, Delhi Technological University, New Delhi 110042, Delhi, India

Journal Paper 2:

Nidhi, & Verma, B. (2024). A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition. *Applied Intelligence*, 1-17.

DOI:10.1007/s10489-024-05598-5 (Published, IF: 3.4)

Applied Intelligence (2024) 54:7793–7809
https://doi.org/10.1007/s10489-024-05598-5



A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition

Nidhi¹ · Bindu Verma¹

Accepted: 6 June 2024 / Published online: 14 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Facial emotion recognition has become a complicated task due to individual variations in facial characteristics, as well as racial and cultural variances. Different psychological studies show that there are complex expressions other than basic emotions which are made up of two basic emotions like “Happily Disgusted”, “Happily Surprised”, “Sadly Surprised”, etc. Compound emotion recognition is challenging due to very less publicly available compound emotion datasets which are imbalanced too. In this paper, we have proposed an LSwin-CBAM for the classification of compound emotions. To address the problem of the imbalanced dataset, the proposed model exploits the cutmix augmentation technique for data augmentation. It also incorporates the CBAM attention mechanism to emphasize the relevant features in an image and swin transformer with fewer swin transformer blocks which leads to less computational complexity in terms of trainable parameters and improves the overall classification accuracy as well. The experimental results of LSwin-CBAM on RAF-DB and EmotioNet datasets show that the proposed transformer-based network can well recognize compound emotions.

Keywords Swin transformer · CutMix augmentation · CBAM · Compound emotion recognition

1 Introduction

Face expressions are an efficient way for people to communicate their emotions. It gives the spectator a plethora of social cues, such as the viewer’s focus of attention, emotion, motivation, and intention and brings out 55% of the conveyed message, which is relatively higher than the part communicated by the combination of voice and language [1]. Research done in social psychology explains that facial expression helps in coordinating the conversation between speaker and listener that can perceive the interest of the listener [2, 3]. In most of the studies, macro-expressions (happiness, surprise, sadness, anger, fear, and disgust) are subjects of concern that are most commonly observed in most cultures. But, sometimes humans can express mixed emotions. For example, a person can have feelings of “sad” and “surprise” components together which forms a compound emotion named “Sadly

Surprised”. The combination of two or more emotions is represented by a compound emotion where one emotion is dominant and the other is complementary. Face recognition in the real environment is a challenging task due to the variation in head orientation, illumination conditions, and head poses. For compound emotion recognition where dominant and complementary emotions need to be recognized, the task becomes more difficult [4].

Authors usually focus on seven basic emotions: ‘Surprise’, ‘Anger’, ‘Happiness’, ‘Contempt’, ‘Fear’, and ‘Disgust’. But there is a need to examine more detailed and precise facial expressions. Some authors tried to find out detailed facial expressions because of recent developments in the area of compound emotions [5, 6]. To examine a person’s emotional state in greater detail using facial emotion expression analysis, compound emotion categories have been proposed [7]. Such studies contribute in the context of compound emotion recognition which comprehends and identifies fine-grained facial expressions.

Transformers [8] are greatly used in Natural Language Processing (NLP) tasks. Research efforts to adapt transformers for vision tasks have been driven by the tremendous achievement of transformers in NLP. Due to its original trans-

✉ Bindu Verma
bindu.cvision@gmail.com

Nidhi
nidhi.kanwar189@gmail.com

¹ Department of Information Technology, Delhi Technological

Journal Paper 3:

Nidhi, and Bindu Verma. "In-the-Wild Facial Emotion Recognition using Relation-aware Geometric Features and CapsNet" is communicated in *Computers and Electrical Engineering* (SCIE Indexed, IF: 4.0) **(Communicated)**

em Computers and Electrical Engineering Bindu Verma | Logout

Home Main Menu Submit a Manuscript About Help

← Submissions Being Processed for Author

Page: 1 of 1 (1 total submissions)

Results per page 10

Action	Manuscript Number	Title	Initial Date Submitted	Status Date	Current Status
Action Links	COMPELECENG-D-24-07317	In-the-Wild Facial Emotion Recognition using Relation-aware Geometric Features and CapsNet	Nov 14, 2024	Nov 28, 2024	Under Review

Page: 1 of 1 (1 total submissions)

Results per page 10



Journal Paper 4:

Nidhi, and Bindu Verma. “ViT-SLS: Vision Transformer with Stochastic Depth for Efficient Driver’s Emotion Recognition System” is communicated in *IEEE Transactions on Human-Machine Systems* (SCIE Indexed, IF: 3.6) **(Communicated)**

ScholarOne Manuscripts™ Nidhi Kanwar ▾ Instructions & Forms Help Log Out

 IEEE Transactions on Human-Machine Systems

Home Author Review

Author Dashboard

Author Dashboard

- 1 Manuscripts I Have Co-Authored
- Legacy Instructions
- 5 Most Recent E-mails

Manuscripts I Have Co-Authored

ATTENTION AUTHORS!

This site is no longer used for new submissions, please visit the [IEEE THMS Author Portal](#) to submit your manuscript.

STATUS	ID	TITLE	CREATED	SUBMITTED
Contact Journal	THMS-24-04-0169 (REX-PROD-2-B11418C4-CEE5-4479-9621-813D03971775-E15427EC-F4E6-4FE2-9F97-DD6BE8EF0515-90298)	ViT-SLS: Vision Transformer with Stochastic Depth for Efficient Driver’s Emotion Recognition System	07-Apr-2024	08-Apr-2024
• Under Review		View Submission Submitting Author: Verma, Bindu		



LIST OF CONFERENCES:

Conference Paper 1:

Nidhi, Bindu Verma. "A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition". In 2023 Seventh International Conference on Image Information Processing (ICIIP) (pp. 591-596) (2023, November). IEEE. **(Published)**

Conferences > 2023 Seventh International Co... 

A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition

Publisher: IEEE

[Cite This](#)

[PDF](#)

Nidhi ; Bindu Verma [All Authors](#)

14

Full

Text Views



Abstract

Document Sections

I. Introduction

II. Literature Survey

III. Advantages

IV. Challenges

V. Limitations

[Show Full Outline](#) ▼

[Authors](#)

[Figures](#)

[References](#)

[Keywords](#)

[Metrics](#)

Abstract:

Facial emotion recognition, a vital aspect of human-computer interaction, has witnessed significant advancements with the integration of attention mechanisms in recent years. This paper presents a comprehensive survey of attention mechanisms in facial emotion recognition. Various types of attention mechanisms have been explored including channel attention, spatial attention, CBAM (Convolutional Block Attention Module), self-attention, and multi-head attention. The survey delves into the basic understanding of each attention mechanism. By examining a wide range of recent studies and methodologies, this research paper synthesizes the advantages, challenges, and limitations of different attention mechanisms, shedding light on their interpretability, computational complexity, and adaptability to diverse facial expressions.

Published in: 2023 Seventh International Conference on Image Information Processing (ICIIP)

Date of Conference: 22-24 November 2023

DOI: 10.1109/ICIIP61524.2023.10537662

Date Added to IEEE Xplore: 28 May 2024

Publisher: IEEE

► **ISBN Information:**

Conference Location: Solan, India

▼ **ISSN Information:**

I. Introduction

Attention mechanisms are the methods used to divert the attention to the relevant regions of an image and neglect irrelevant ones. The neural network's ability to focus on specific regions, eliminate irrelevant information, and effectively extract essential features from images is made possible by using a [Sign in to Continue Reading](#) have proposed different variants of attention mechanisms over the literature which outperform object detection [1], [2], image classification [3], [4], face recognition [5], [6], action recognition [7], [8], person re-identification [9], [10], image generation [11], [12], pose estimation [13], etc.

Conference 1: Certificate



JAYPEE
GROUP



ICIIP



JUIT
जिज्ञासा तन्त्राभिराम

ICIIP2023/#61524/PP/CRN-0096



IEEE

Certificate of Participation

This is to certify that

Nidhi

has participated and presented a paper entitled

A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition

in **2023 Seventh International Conference on Image Information Processing (ICIIP-2023)**
organised by the Department of Computer Science & Engineering,
Jaypee University of Information Technology, Wahnaghat,
Solan, Himachal Pradesh, India during 22nd-24th November 2023.



Prof. (Dr.) Vivek Sehgal
Principal General Chair
ICIIP-2023



Dr. Ruchi Verma
Conference General Chair
ICIIP-2023



Dr. Vipul Sharma
Conference Chair
ICIIP-2023



Dr. Pankaj Dhiman
Conference Chair
ICIIP-2023

Conference Paper 2:

Nidhi, Bindu Verma. "Empirical Insights: Unraveling the Impact of Various Attention Mechanisms on Facial Emotion Recognition" presented in 5th International Conference on Data Analytics & Management (ICDAM-2024), Springer. (Accepted & Presented)





DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daultapur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: Development of Framework for Facial Emotion Recognition

Total Pages: 141

Name of the Scholar: Nidhi

Supervisor: Dr. Bindu Verma

Department: Information Technology

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: Turnitin

Similarity Index: 8%

Word Count: 36,574 Words

Date: 2/12/2024

Candidate's Signature:

Supervisor's Signature:

Nidhi Thesis

intro_Ph_D_Thesis__Nidhi__final__Copy_ (11).pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid::27535:72812592

Submission Date

Dec 2, 2024, 6:50 AM GMT+5:30

Download Date

Dec 2, 2024, 6:57 AM GMT+5:30

File Name

intro_Ph_D_Thesis__Nidhi__final__Copy_ (11).pdf

File Size

10.1 MB

141 Pages

36,574 Words

202,846 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.





Filtered from the Report

- ▶ Bibliography
- ▶ Small Matches (less than 10 words)




Exclusions

- ▶ 3 Excluded Sources

Match Groups

-  **195** Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%
Matches that are still very similar to source material
-  **5** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 2%  Internet sources
- 7%  Publications
- 3%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

NIDHI

Full Time PhD Scholar, Delhi Technological University
New Delhi, Pincode-110042
Phone- +918395998421
Email_id- nidhi.kanwar189@gmail.com



RESEARCH INTEREST:

- Having research interest in Machine Learning (Neural Network) specifically working on Image classification.
- Working on “Facial Emotion Recognition”.

EDUCATION QUALIFICATION:

Course	College/Institute	Board/ University	Percentage/CGPA	Year
M.Tech.(CS)	Banasthali Vidyapith, Banasthali	Banasthali Vidyapith	8.77 CGPA	2019
B.Tech(CS)	Banasthali Vidyapith, Banasthali	Banasthali Vidyapith	71.44%	2017
Intermediate(12 th)	Delhi Public School	CBSE	71.8%	2013
High School(10 th)	Dharam Public School	CBSE	9.4	2011

PROJECTS:

B.TECH PROJECTS:

Major Project: Worked on Academy for Tally Course Project

- **Project Overview:** In this we had to make application form for Farmer Producer Organization where FPOs will register with their details and those details will be shown to District Level Officer for Verification and verified FPOs details will further shown to State Level Officer for final verification.
- **Environment:** JAVA, MySQL

Minor Project: Worked on a minor project on Mess Management

- **Project Overview:** Mess Management System is to get the current status of the mess and meals per day, to manage details regarding the stock of groceries, diet, worker's record based on daily fluctuating rates.
- **Environment:** JAVA, MySQL

M.TECH PROJECT:

- **Project Overview:** Research based project on Neural Network based approach for data classification and performance of network by varying the design parameters
- **Environment:** Python, Tensorflow

PhD PROJECT:

- **Project Overview:** Worked on Facial Emotion Recognition
- **Environment:** Python, Tensorflow, PyTorch

INTERNSHIP EXPERIENCE:

CSC E-Governance Services India Limited, New Delhi

RESEARCH EXPERIENCE

- Defence Research & Development Organization(DRDO), Delhi
- Worked as an Assistant Professor at GLA University, Mathura

RESEARCH WORK:

Publications:

Journals:

- Nidhi, and Bindu Verma. "From methods to datasets: a detailed study on facial emotion recognition." Applied Intelligence 53, no. 24 (2023): 30219-30249. **(SCIE Indexed, IF: 3.4) DOI: 10.1007/s10489-023-05052-y (Published)**
- Nidhi, and Bindu Verma. "A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition." Applied Intelligence (2024): 1-17. **(SCIE Indexed, IF: 3.4) DOI:10.1007/s10489-024-05598-5 (Published)**
- Nidhi, and Bindu Verma. "ViT-SLS: Vision Transformer with Stochastic Depth for Efficient Driver's Emotion Recognition System" is communicated in IEEE Transactions on Human-Machine Systems **(SCIE Indexed, IF: 3.6) (Communicated)**

- Nidhi, and Bindu Verma. "In-the-Wild Facial Emotion Recognition using Relation-aware Geometric Features and CapsNet" is communicated in Computers and Electrical Engineering (**SCIE Indexed, IF: 4.0**) (**Communicated**)

Conferences:

- Nidhi, Bindu Verma. "A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition". In 2023 Seventh International Conference on Image Information Processing (ICIIP) (pp. 591-596) (2023, November). IEEE (**Published**)
- Kanwar, N., Goswami, A. K., Mishra, S. P. (2019). Design Issues in Artificial Neural Network (ANN). *4th IEEE International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU 2019)*. (**Published**)
- Kanwar, N., Goswami, A., & Mishra, S. P. (2019). Deep Learning-An Emerging Paradigm. *Available at SSRN 3355272*. (**Published**)
- Kanwar, N., Goswami, A. K., & Malhotra, S. (2022). An Empirical Analysis of Multilayer Perceptron Based on Neural Network using Image Segment Dataset. *In Advances in Computational Intelligence and Communication Technology (pp. 423-431)*. Springer, Singapore. (**Published**)
- Paper Title: "A Comprehensive Exploration: Attention Mechanisms in Facial Emotion Recognition" in 2023 Seventh International Conference on Image Information Processing (ICIIP -2023) (**Accepted & Presented**)

AWARDS AND ACHIEVEMENTS:

- Topper in M.Tech (Computer Science)
- Organized a Value Added Course of "Shell Programming" in GLA University, Mathura
- Member of NAAC team in GLA University, Mathura
- Core team member of Mayukh 2k17 & Mayukh 2k18 Management team (National Level Fest)
- Conducted workshop on "What's Beauty without Brain" in Mayukh 2k17.
- Merit certified in National Knowledge Olympiad.
- Sports Captain in school.
- Won Debate Competition at School Level.
- Won in Interschool Dance Competition.

- Won in “Bharat ko Jano” quiz Competition.

TEACHING EXPERIENCE:

- 18 months (1.5yr) experience of Teaching as an Assistant Professor in GLA University, Mathura, Uttar Pradesh

PERSONAL STRENGTH:

- Hardworking
- Optimistic
- Confident
- Problem Solving Skills
- Excellent time manager

HOBBIES AND INTERESTS:

Reading, Sports, Music, Dance

PERSONAL PROFILE:

Date of Birth: 18-sep-1996

Father’s Name- Dinesh Thakur

Mother’s Name-Indra

Nationality: Indian

Languages Known: Hindi, English

Declaration:

I hereby declare that the above mentioned details are true to the best of my knowledge.