Thèse de doctorat

Pour obtenir le grade de Docteur de

l'UNIVERSITÉ POLYTECHNIQUE HAUTS-DE-FRANCE

et de l'INSA HAUTS-DE-FRANCE

et de l'UNIVERSITÉ de technologie du Brandebourg

Discipline: **Electronique, microélectronique, nanoélectronique et micro-ondes**

Présentée et soutenue par IBTISSAM SAADI.

Le 01/12/2025, à Valenciennes, France

École doctorale :

École Doctorale Polytechnique Hauts-de-France (ED PHF n°635)

Unité de recherche :

Institut d'Electronique de Microélectronique et de Nanotechnologie – Site de Valenciennes (IEMN – UMR CNRS 8520)

Amélioration de la reconnaissance des expressions faciales dans des environnements réalistes à l'aide d'approches d'apprentissage profond

JURY

Président:	Mr. Yassine ruichek	Prof, Université de technologie de Belfort-Montbéliard
Rapporteur:	Mr. Frédéric Chausse	Prof, Université Clermont Auvergne
Rapporteur:	Mrs. Susana Castillo	Dr, Université technique de Braunschweig
Examinateur:	Mr. Christian Wallraven	Prof, Université de Korea
Directeur:	Mr. Abdelmalik TALEB-AHMED	Prof, Université Polytechnique Hauts-de- France
Directeur:	Mr. Douglas W. Cunningham	Prof, Université de technologie du Brandebourg
Co-encadrant:	Mr. Yassin Elhillali	Prof, Université Polytechnique Hauts-de- France
Invité:	Mr. Salim Bitam	Prof, Université Mohamed Khider Biskra
Invité:	Mrs. Atika Rivenq	Prof, Université Polytechnique Hauts-de- France

PhD Thesis

Submitted for the degree of Doctor of Philosophy from

UNIVERSITÉ POLYTECHNIQUE HAUTS-DE-FRANCE

And INSA HAUTS-DE-FRANCE

and Brandenburg University of Technology

Subject: **Electronics, Microelectronics, Nanoelectronics and Micro-waves**

Presented and defended by IBTISSAM SAADI.

On 01/12/2025, Valenciennes, France

Doctoral school:

Doctoral School Polytechnique Hauts-de-France (ED PHF n°635)

Research unit:

Institute of Electronics Microelectronics and Nanotechnology – Valenciennes site (IEMN – UMR CNRS 8520)

Improving Facial Expression Recognition in Realistic Environments Using Deep Learning Approaches

JURY

President:	Mr. Yassine ruichek	Prof, Université de technologie de Belfort-Montbéliard
Reviewer:	Mr. Frédéric Chausse	Prof, University Clermont Auvergne
Reviewer:	Mrs. Susana Castillo	Dr, Technical University of Braunschweig
Examiner:	Mr. Christian Wallraven	Prof, Korea University
Director:	Mr. Abdelmalik TALEB-AHMED	Prof, Polytechnic University of Hauts-de-France
Director:	Mr. Douglas W. Cunningham	Prof, Brandenburg University of Technology
Co-supervisor:	Mr. Yassin Elhillali	Prof, Polytechnic University of Hauts-de-France
Invited:	Mr. Salim Bitam	Prof, Mohamed Khider University Biskra
Invited:	Mrs. Atika Rivenq	Prof, Polytechnic University of Hauts-de-France

Abstract

Facial expressions are vital channels of non-verbal communication, conveying rich information about emotional and cognitive states in social interactions. Enabling intelligent systems to automatically recognize these expressions makes Facial Expression Recognition (FER) a key task in affective computing. Despite significant progress, existing FER methods still face limitations in effectively and efficiently recognizing facial expressions under real-world conditions, such as those found in driving environments or dynamic, in-the-wild scenarios.

This thesis addresses these challenges by proposing several novel deep learning-based models that improve the performance, robustness, and efficiency of FER systems. To this end, several contributions are presented. The work begins with a comprehensive review of FER, highlighting persistent challenges in the field, particularly within the driving context. Subsequently, DFER-GCViT, a Vision Transformer-based architecture tailored for driver FER, is introduced to improve recognition accuracy under conditions of occlusion, pose variation, and lighting changes. To enhance computational efficiency, ShuffViT-DFER is proposed, a lightweight hybrid model that combines convolutional and transformer-based pretrained networks. Furthermore, multimodal approaches are explored using Vision-Language Models (VLMs), particularly CLIP. CLIVP-FER is developed to integrate visual and textual features, enhancing semantic understanding in the driving context. The research then shifts toward general dynamic FER, leveraging parameter-efficient fine-tuning and temporal modeling of CLIP, and maintains strong performance with reduced computational cost through the proposed PE-CLIP.

Extensive experiments on benchmark datasets demonstrate the effectiveness of the proposed models across diverse and challenging conditions. The results underscore the importance of optimizing architectures, incorporating multimodal cues, and enabling lightweight FER systems suitable for practical deployment.

Keywords: Facial Expression Recognition (FER), Emotion Recognition, Driving Context, Dynamic FER, Deep Learning, Vision Transformers, Vision-Language Models (VLMs), Multimodal Learning, Computational Efficiency.

Kurzfassung

Gezichtsuitdrukkingen zijn essentiële kanalen voor non-verbale communicatie, die rijke informatie over emotionele en cognitieve toestanden in sociale interacties overbrengen. Het mogelijk maken dat intelligente systemen deze uitdrukkingen automatisch herkennen, maakt gezichtsherkenning (FER) tot een belangrijke taak in affectieve informatica. Ondanks aanzienlijke vooruitgang hebben bestaande FER-methoden nog steeds te maken met beperkingen bij het effectief en efficiënt herkennen van gezichtsuitdrukkingen onder reële omstandigheden, zoals in rijomgevingen of dynamische, natuurlijke scenario's.

Dit proefschrift pakt deze uitdagingen aan door verschillende nieuwe, op deep learning gebaseerde modellen voor te stellen die de prestaties, robuustheid en efficiëntie van FER-systemen verbeteren. Daartoe worden verschillende bijdragen gepresenteerd. Het werk begint met een uitgebreid overzicht van FER, waarbij de aandacht wordt gevestigd op hardnekkige uitdagingen in het veld, met name in de context van het autorijden. Vervolgens wordt DFER-GCViT geïntroduceerd, een op Vision Transformer gebaseerde architectuur die is afgestemd op FER voor automobilisten, om de herkenningsnauwkeurigheid te verbeteren onder omstandigheden van occlusie, variatie in houding en veranderingen in de belichting. Om de computationele efficiëntie te verbeteren, wordt ShuffViT-DFER voorgesteld, een lichtgewicht hybride model dat convolutional en op transformers gebaseerde vooraf getrainde netwerken combineert. Verder worden multimodale benaderingen onderzocht met behulp van Vision-Language Models (VLMs), met name CLIP. CLIVP-FER is ontwikkeld om visuele en tekstuele kenmerken te integreren, waardoor het semantisch begrip in de rijcontext wordt verbeterd. Het onderzoek verschuift vervolgens naar algemene dynamische FER, waarbij gebruik wordt gemaakt van parameter-efficiënte fine-tuning en temporele modellering van CLIP, en behoudt sterke prestaties met lagere rekenkosten door middel van de voorgestelde PE-CLIP.

Uitgebreide experimenten met benchmarkdatasets tonen de effectiviteit van de voorgestelde modellen aan in diverse en uitdagende omstandigheden. De resultaten onderstrepen het belang van het optimaliseren van architecturen, het integreren van multimodale signalen

en het mogelijk maken van lichtgewicht FER-systemen die geschikt zijn voor praktische implementatie.

Sleutelwoorden: Gezichtsherkenning (FER), Emotieherkenning, Rijcontext, Dynamische FER, Deep learning, Vision Transformers, Vision-Language Models (VLMs), Multimodaal leren, Computationele efficiëntie.

Résumé

Les expressions faciales sont des canaux essentiels de communication non verbale, qui transmettent de nombreuses informations sur les états émotionnels et cognitifs dans les interactions sociales. Permettre aux systèmes intelligents de reconnaître automatiquement ces expressions fait de la reconnaissance des expressions faciales (FER) une tâche clé dans le domaine de l'informatique affective. Malgré des progrès significatifs, les méthodes FER existantes se heurtent encore à des limites pour reconnaître efficacement et efficacement les expressions faciales dans des conditions réelles, telles que celles rencontrées dans les environnements de conduite ou dans des scénarios dynamiques et naturels.

Cette thèse aborde ces défis en proposant plusieurs modèles novateurs basés sur l'apprentissage profond qui améliorent les performances, la robustesse et l'efficacité des systèmes FER. À cette fin, plusieurs contributions sont présentées. Le travail commence par une revue complète de la FER, soulignant les défis persistants dans ce domaine, en particulier dans le contexte de la conduite. Ensuite, DFER-GCViT, une architecture basée sur Vision Transformer et adaptée à la FER des conducteurs, est présentée afin d'améliorer la précision de la reconnaissance dans des conditions d'occlusion, de variation de pose et de changements d'éclairage. Afin d'améliorer l'efficacité computationnelle, ShuffViT-DFER est proposé, un modèle hybride léger qui combine des réseaux pré-entraînés convolutifs et basés sur des transformateurs. En outre, des approches multimodales sont explorées à l'aide de modèles vision-langage (VLM), en particulier CLIP. CLIVP-FER est développé pour intégrer des caractéristiques visuelles et textuelles, améliorant ainsi la compréhension sémantique dans le contexte de la conduite. La recherche s'oriente ensuite vers la FER dynamique générale, en tirant parti du réglage fin efficace en termes de paramètres et de la modélisation temporelle de CLIP, et maintient des performances élevées avec un coût de calcul réduit grâce au PE-CLIP proposé.

Des expériences approfondies sur des ensembles de données de référence démontrent l'efficacité des modèles proposés dans des conditions diverses et difficiles. Les résultats

soulignent l'importance d'optimiser les architectures, d'intégrer des indices multimodaux et de permettre la mise en place de systèmes FER légers adaptés à un déploiement pratique.

Mots-clés : Reconnaissance des expressions faciales (FER), reconnaissance des émotions, contexte de conduite, FER dynamique, apprentissage profond, transformateurs de vision, modèles vision-langage (VLM), apprentissage multimodal, efficacité computationnelle.

Acknowledgements

This thesis marks the end of a long and meaningful journey, one that would not have been possible without the support and guidance of many people. I am deeply grateful to all who accompanied me along the way. Above all, I thank Allah for granting me the strength and perseverance to carry this work to completion.

I would like to express my sincere appreciation to my supervisors, Prof. Abdelmalik Taleb-Ahmed (UPHF, France) and Prof. Douglas Cunningham (BTU, Germany), for their trust from the very beginning and for giving me the opportunity to pursue this research. Their expertise, guidance, and continuous support have been instrumental throughout this journey.

My heartfelt thanks go to Prof. Abdenour Hadid (Sorbonne University Abu Dhabi), whose support from the very beginning of this journey made a real difference. His dedicated mentorship, timely guidance, and generous advice at key moments greatly shaped both the direction of this work and my personal development. I am also grateful to Prof. Yassin Elhillali (UPHF, France) for welcoming me into the Syfra Lab and providing a collaborative environment, as well as for his valuable feedback and encouragement throughout this research.

I would also like to extend my sincere thanks to the EUNICE Alliance – *European University for Customised Education* for their financial support. Their contribution made this work possible, and I am truly grateful for the opportunities provided through this program. I also gratefully acknowledge the financial support provided by Brandenburg University of Technology (BTU) Cottbus–Senftenberg .

My sincere thanks go to the members of my defense committee for agreeing to evaluate my thesis. I am truly honored by their presence and grateful for the time and expertise they dedicated to reviewing and discussing my work.

I am thankful to my colleagues at the IEMN Laboratory (UPHF) and the TIKI Lab (BTU) for the engaging work environment, helpful exchanges, and shared moments that made this experience both productive and enjoyable. Working with students at BTU was also a rewarding part of this journey; their curiosity and involvement contributed to the thesis, and mentoring them helped me grow as both a researcher and a communicator.

Lastly, I extend my deepest thanks to my family and close friends for their constant love, patience, and encouragement. Your support gave me the strength to continue. I am also grateful to all who supported me along the way, whether through collaboration, friendship, or simple acts of kindness.

Funding Acknowledgement

This doctoral research was supported by the EUNICE Alliance – European University for Customised Education as well as Brandenburg University of Technology (BTU) Cottbus–Senftenberg . I sincerely acknowledge their financial contribution, which played an important role in enabling the completion of this thesis.

Contents

List of Figures i

List of Tables v

Acronyms viii

Academic Activities During the PhD x

1 Introduction 1

1.1 Context and Motivation 1
1.2 Research Objectives and Questions 2
1.3 Research Contributions 3
1.4 Thesis Outline 5

2 Facial Expression Recognition: Background & Literature Review 7

2.1 Introduction 7
2.2 Fundamentals of Facial Expression Recognition 7
2.2.1 Key Concepts and Terminologies 9
2.2.2 Application Domains of FER 12
2.2.3 General workflow of FER systems 14
2.3 Datasets, Evaluation Metrics, and Interpretability in FER 18
2.3.1 Commonly Used FER Datasets 19
2.3.2 Evaluation Metrics and Visualization Techniques 27
2.4 Modeling Techniques for FER 30
2.4.1 Classical Machine Learning Models 30
2.4.2 Deep Learning Models 32
2.4.3 Transfer Learning Strategies 39
2.5 Review of Existing FER Methods 41
2.5.1 Approaches in General-Purpose FER 41

2.5.2 FER in Driving environments . . . 47
2.5.3 Limitations and Open Challenges in Facial Expression Recognition (FER) . . . 53
2.6 Conclusion . . . 55

3 A Global Context Vision Transformer for Driver Facial Expression Recognition 57
3.1 Introduction . . . 57
3.2 Motivation . . . 57
3.3 Related Work . . . 59
3.4 Overview of Global Context Vision Transformer (GC-ViT) . . . 61
3.5 Proposed Methodology: DFER-GCViT . . . 63
3.5.1 Overall Architecture . . . 63
3.5.2 Pre-processing Module . . . 63
3.5.3 Feature Extraction Module . . . 66
3.5.4 Driver Facial Expression Recognition Module . . . 66
3.6 Experimental Analysis . . . 67
3.6.1 Dataset Description . . . 67
3.6.2 Implementation Setup . . . 68
3.6.3 Results and Discussion . . . 68
3.7 Conclusion . . . 73

4 Lightweight, Fast and Efficient Recognition of Driver Facial Expression 75
4.1 Introduction . . . 75
4.2 Motivation . . . 75
4.3 Related Work . . . 76
4.4 Pre-trained Models for FER: An Overview . . . 78
4.4.1 Pre-trained CNN Architectures . . . 78
4.4.2 Pre-trained Vision Transformers architectures . . . 81
4.4.3 The Selected Pre-trained Models for Driver's FER . . . 82
4.5 Proposed Model: ShuffViT-DFER . . . 85
4.5.1 Overall architecture . . . 85
4.5.2 Pre-processing . . . 86
4.5.3 Lightweight and efficient feature extraction . . . 87
4.5.4 Classification . . . 88

4.6 Experimental Analysis 89
4.6.1 Experimental Data 89
4.6.2 Experimental Setup 89
4.6.3 Results and Discussion 90
4.7 Conclusion 95

5 Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment 96
5.1 Introduction 96
5.2 Motivation and Related Work 97
5.3 Contrastive Language-Image Pre-Training (CLIP): An overview 98
5.4 Proposed Method: CLIVP-FER 101
5.4.1 Overall Architecture 101
5.4.2 Data Pre-processing 101
5.4.3 Extracting and Refining Features Using CLIP and Average Pooling . 103
5.4.4 Emotion Classification 105
5.5 Experimental Analysis 105
5.5.1 Experimental Data 106
5.5.2 Experimental Setup 106
5.5.3 Results and Discussion 106
5.6 Conclusion 111

6 A Parameter-Efficient Fine-Tuning of VLMs for Dynamic Facial Expression Recognition 113
6.1 Introduction 113
6.2 Motivation and Related Work 113
6.3 Proposed Framework 117
6.3.1 Overview of PE-CLIP 117
6.3.2 Visual Feature Extraction and Temporal Modeling 119
6.3.3 Textual Feature Extraction and Alignment 122
6.3.4 Classification via Cosine Similarity 125
6.4 Experimental Analysis 125
6.4.1 Experimental Data 126
6.4.2 Implementation Setup 126
6.4.3 Experimental Results 127

6.4.4 Ablation Analysis 130
6.4.5 Visualization-Based Results 136
6.5 Conclusion 141

7 Conclusion and Future Work **142**
7.1 Summary of Contributions 142
7.2 Future Work 143

Résumé substantiel en français **145**
1 Contexte et Motivation 145
2 Problématique et objectifs 146
3 Fondements théoriques et méthodologiques 147
3.1 Applications de la reconnaissance des expressions faciales en contexte de conduite 148
3.2 Workflow général d'un système de reconnaissance des expressions faciales 148
3.3 Jeux de données et métriques d'évaluation pour FER 151
3.4 Métriques d'évaluation et des techniques de visualisation 152
4 l'État de l'art 153
5 Contributions 156
5.1 Modélisation du contexte global pour la reconnaissance des expressions faciales du conducteur 156
5.2 Architecture légère et efficace pour la reconnaissance des expressions faciales en conduite 157
5.3 Reconnaissance multimodale des expressions faciales du conducteur par modèles vision–langage 157
5.4 Adaptation efficace en paramètres des modèles vision–langage pour la reconnaissance dynamique des expressions faciales 158
6 Conclusion et perspectives 159

Bibliography **161**

List of Figures

2.1 Examples of facial expressions induced through electrical stimulation by Duchenne de Boulogne. 8
2.2 Examples of discrete emotion models: Ekman's six basic emotions, Plutchik's wheel of emotions. 10
2.3 Dimensional emotion models. (a) Pleasure-Arousal-Dominance (PAD) and (b) Valence-Arousal (V-A) models 11
2.4 Illustration of lower and upper facial AUs as defined in the FACS. 11
2.5 Example of driver's facial expression recognition using a camera installed inside a car. The figure is only for illustration and is partially generated using OpenAI's ChatGPT-4. 13
2.6 General workflow of a facial expression recognition (FER) system. The pipeline includes input data, pre-processing (face detection, alignment, augmentation, and normalization), feature extraction using either handcrafted or deep learning-based methods, and emotion classification using classical or deep learning classifiers. 14
2.7 Overview of the facial expression dataset characteristics. 19
2.8 Some samples from FER datasets used in this thesis. 23
2.9 A representative CNN architecture illustrating a layered structure comprising convolutional, activation, and pooling operations, followed by fully connected layers used for classification. 33
2.10 Illustration of RNN, LSTM, and GRU cells architecture. 35
2.11 The architecture of vision transformer (ViT). 36
2.12 Deep learning-based transfer learning strategies 40

3.1 Facial expression recognition in driving environments presents additional challenges, including varying lighting conditions, occlusions, pose variations, and motion blur. 58

3.2 The GCViT architecture, showing the hierarchical structure with local and global multi-head self-attention blocks. 61
3.3 Overall workflow of the proposed DFER-GCViT methodology. The process begins with the application of a series of preprocessing techniques to the input facial images, including face detection, cropping, resizing, normalization, and data augmentation. The preprocessed images are then passed through the pre-trained GCViT model for feature extraction, producing both feature maps and attention heatmaps. Finally, the extracted features are input into a classification head to predict facial expression classes. 64
3.4 Detailed view of the feature extraction and classification modules in the GCViT-based model for driver facial expression recognition. The figure illustrates the removal of the original fully connected (FC) layer, its replacement with a new FC layer in the enhanced classifier head, and the subsequent classification of driver facial expressions. 65
3.5 Confusion matrices of the KMU-FED dataset and FER2013 dataset. 72
3.6 Heatmap visualization of DFER-GCViT using Grad-CAM technique for facial expressions on KMU-FED and FER2013 datasets. 73

4.1 Core Architectural Units of ShuffleNet V1 and ShuffleNet V2. 84
4.2 The architecture of EfficientViT. 85
4.3 The workflow of our proposed model for driver facial expression recognition. The pipeline begins with face detection using the MTCNN, followed by data augmentation techniques applied to the cropped face. The preprocessed images are then processed in parallel by two pretrained models, ShuffleNet-V2 and EfficientViT-M2. The extracted features are fused and passed through a custom classifier to predict the driver's facial expression accurately. 86
4.4 Confusion matrices of the proposed model for the KMU-FED and KDEF datasets. 93
4.5 XGrad-CAM attention maps for different facial expressions, generated using (a) ShuffleNet V2, (b) EfficientViT-M2, and (c) the proposed ShuffViT-DFER model on the KMU-FED dataset. 94

5.1 The Overall CLIP workflow. 99

5.2 The figure illustrates our proposed CLIVP-FER architecture. The visual inputs are processed with Contrastive Language-Image Pretraining(CLIP) encoder into a vector f_{v}. The textual inputs are processed into a vector f_{t}. The visual feature vector f_{v} is then processed through an average pooling layer to produce f_{v1}, which is subsequently combined with f_{t} to form f_{z}. This combined vector is further processed by an additional average pooling layer, resulting in the final feature vector f_{final}, used to classify the facial expressions by a Multilayer Perceptron (MLP). The figure highlights the frozen (non-trainable) and trainable components. 102
5.3 Confusion matrix of our CLIVP-FER model on the KMU-FED dataset. . . 108

6.1 Performance comparison of dynamic facial expression recognition on the DFEW dataset. 116
6.2 Overall architecture of the proposed PE-CLIP model. The model takes tokenized AU-based textual descriptions C_t and embedded image sequences X_t, enriched with MaPLe learnable tokens as inputs. These representations are processed through CLIP encoders, where the Shared Adapter (ShA) improves the representation learning across textual-visual modalities. Additionally, the Temporal Dynamic Adapter (TDA) with dynamic scaling, captures the key temporal dependencies in the vision branch, while the Textual Adapter (TA) enhances the textual representations. The resulting visual (f_v) and textual (f_t) embeddings are mapped into CLIP's shared space, where classification is performed via cosine similarity to associate expressions with their corresponding labels. 118
6.3 Confusion Matrices Across 5 Folds on DFEW. 137
6.4 Attention visualization of our proposed PE-CLIP model. The figure presents attention maps for *sadness* and *anger* (left to right), showing how focus evolves across different model configurations. Each section includes original frames (first row), baseline model (second row), model with ShA adapters (third row), and full PE-CLIP with ShA and TDA (fourth row). Warmer colors indicate stronger attention, with PE-CLIP progressively refining focus on key facial regions, enhancing representation refinement (via ShA) and temporal modeling (via TDA). 138

6.5 t-SNE Visualization of High-Level Feature Distributions on DFEW (fd1–5): baseline Model, PE-CLIP without adapters and prompt (Top row) vs. Proposed PE-CLIP model (Bottom row). 139

1 Illustration des défis de la reconnaissance des expressions faciales en contexte de conduite. 146
2 Workflow général d'un système de reconnaissance des expressions faciales, comprenant l'acquisition des données, le prétraitement, l'extraction de caractéristiques et la classification des émotions. 149
3 Exemples d'échantillons issus des jeux de données de reconnaissance des expressions faciales utilisés dans ce thèse. 152

List of Tables

2.1 Description of Related-FER Datasets. BEs=basic Emotions, CEs=compound emotions (combination of BEs), while a "N/A" in the table cells, indicates missing information (Part 1). 24
2.2 Description of FER-Related Datasets. BEs=basic Emotions, CEs=compound emotions (combination of BEs), while a "N/A" in the table cells, indicates missing information (Part 2). 25
2.3 Summary of general FER-related works (Part 1). 44
2.4 Summary of general FER-related works (Part 2). 45
2.5 Summary of FER-related works in Driving Environments (Part 1). 51
2.6 Summary of FER-related works in Driving Environments (Part 2). 52

3.1 Hyperparameter Settings for the DFER-GCViT model on the KMU-FED and FER2013 datasets. "–" indicates not applicable. 69
3.2 Comparison between the proposed method and state-of-the-art methods on the KMU-FED dataset. 70
3.3 Comparison between the proposed method and state-of-the-art methods on the FER2013 dataset. 71

4.1 Hyperparameter Settings for ShuffViT-DFER on the KMU-FED and KDEF datasets. "-" denotes not applicable. 90
4.2 Comparison between the proposed method and state-of-the-art and existing methods on the KMU-FED dataset. "NAN" indicates that the result for the corresponding setting was not reported in the original paper. 91
4.3 Comparison between the proposed method and state-of-the-art and existing methods on the KDEF dataset. 91
4.4 Results of the ablation study on KMU-FED dataset. "w" denotes "with", "w/o" denotes without, and "C" denotes our designed classifier. 92

5.1 Facial Expressions and Their Associated Descriptions Generated by ChatGPT-4. 103
5.2 Hyperparameter Settings for CLIVP-FER on the KMU-FED dataset. . . . 107
5.3 Obtained results using our CLIVP-FER model, compared with state-of-the-art methods on the KMU-FED dataset. 107
5.4 Cross-database performance evaluation of our proposed CLIVP-FER model. KMU-FED dataset corresponds to a driving environment, while FER2013 and RAF-DB datasets correspond to a non-driving environment. The training is conducted on KMU-FED, and the testing is on FER2013 and RAF-DB datasets. 109
5.5 Evaluating the impact of CLIVP-FER model's components on the KMU-FED dataset. The first row shows the results of using only image features. The second row shows the results of combining image and text features. The last row demonstrates the performance of our model utilizing both visual and textual features, along with an average pooling layer. 110

6.1 Facial Expressions and Their Associated Action Units (AUs) Descriptions for FER. 123
6.2 Hyperparameter Configurations for Training PE-CLIP on DFEW and FERV39K Datasets. 127
6.3 Comparison of PE-CLIP with state-of-the-art methods on the DFEW and FERV39K datasets. The table highlights the Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) achieved by each method, alongside their respective backbones and the number of tunable parameters (in millions). "-" denotes results were not reported in the original paper. . . 128
6.4 Accuracy Comparison Among Various Emotion Categories on DFEW. "-" denotes results were not reported in the original paper. 129
6.5 Performance Comparison of the Overall Framework Components on DFEW Dataset. 131
6.6 Impact of Different Adapters on PE-CLIP Performance on the DFEW Dataset ("TA": Textual Adapter, "ShA": Shared Adapter, "TDA": Temporal Dynamic Adapter). 132
6.7 Impact of dynamic scale mechanism in TDA. ("w/o": without, "DS": dynamic Scale). 133

6.8 Performance Comparison of Temporal Adapter Configurations within TDA on DFEW. 134
6.9 Influence of the reduction factor of adapters. 135
6.10 Performance comparison of key components on the DFEW dataset. 136

Acronyms

AAM	Active Appearance Models
ADAS	Advanced Driver Assistance Systems
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AUs	Action Units
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CS-Former	spatial transformer
CV	Computer Vision
DL	Deep Learning
DNNs	Deep Neural Networks
FACS	Facial Action Coding System
FER	Facial Expression Recognition
FFN	Feed-Forward Network
GRU	Gated Recurrent Units
Grad-CAM	Gradient-weighted Class Activation Mapping
HOG	Histogram of Oriented Gradients
ITM	Image-Text Matching
KNN	K-Nearest Neighbors
LBP	Local Binary Patterns
LPDP	Local Prominent Directional Pattern
LSTM	Long Short-Term Memory
MaPLe	Multi-modal Prompt Learning
MHSA	Multi-Head Self-Attention
MIM	Masked Image Modeling
ML	Machine Learning
MLM	Masked Language Modeling
MLP	Multilayer Perceptrons

MTCNN	Multi-task Cascaded Convolutional Network
NLP	Natural Language Processing
PAD	Pleasure-Arousal-Dominance
PCA	Principal Component Analysis
PEFT	Parameter-Efficient Fine-Tuning
RL	Reinforcement Learning
RNNs	Recurrent Neural Networks
SVM	Support Vector Machines
SGD	Stochastic Gradient Descent
ShA	Shared Adapter
TA	Textual Adapter
T-Former	temporal transformer
TDA	Temporal Dynamic Adapter
UAR	Unweighted Average Recall
V-A	Valence-Arousal
ViT	Vision Transformer
VLMs	Vision Language Models
VQA	Visual Question Answering
WAR	Weighted Average Recall

Academic Activities During the PhD

Throughout my doctoral research in France and Germany (2022–2025), I actively contributed to the academic community through scientific publications, teaching activities, and the supervision of undergraduate and master's theses.

Scientific Publications

The following research papers have been produced as part of this research:

Journals

- Saadi, I., I., Cunningham, D. W., Abdelmalik, T. A., Hadid, A., & El Hillali, Y. (2023). Driver's facial expression recognition: A comprehensive survey. Expert Systems with Applications, 122784. DOI: https://doi.org/10.1016/j.eswa.2023.122784 (Published)
- Saadi, I., Hadid, A., Cunningham, D. W., Taleb-Ahmed, A., & El Hillali, Y. (2025). PE-CLIP: A Parameter-Efficient Fine-Tuning of Vision Language Models for Dynamic Facial Expression Recognition. ACM Transactions on Multimedia Computing, Communications and Applications. (Published)

International Conferences and Workshops

- Saadi, I., Cunningham, D. W., Abdelmalik, T. A., Hadid, A., & El Hillali, Y. (2023, December). Driver's Facial Expression Recognition Using Global Context Vision Transformer. In 2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI) (pp. 1-8). IEEE. https://doi.org/10.1109/cvmi59935.2023.10464794

- Saadi, I., Cunningham, D. W., Abdelmalik, T. A., Hadid, A., & Hillali, Y. E. (2024). Shuffle Vision Transformer: Lightweight, Fast and Efficient Recognition of Driver Facial Expression. IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), (In print).

- Saadi, I., Hadid, A., Cunningham, D.W., Taleb-Ahmed, A., El Hillali, Y. (2025). Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment. Sensor-Based Activity Recognition and Artificial Intelligence. iWOAR 2024. Lecture Notes in Computer Science, vol 15357. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-80856-2_6

Teaching Activities

During my time as a research assistant at BTU, I had the opportunity to gain teaching experience by serving as a teaching assistant over the course of three semesters (2024–2025), contributing to the following courses:

- Research Module in Artificial Intelligence (Exercise)
- Oberseminar Graphische Systeme (Seminar)

Supervised Theses

Between 2024 and 2025, I had the opportunity to advise and co-advise several master's theses. These projects were carried out in collaboration with students working on topics related to artificial intelligence, computer vision, and facial expression recognition.

- Bansal Minal, "AI-based climate control system in cars", Master in AI, BTU University, 2025.
- Kakkar Vidhanshee, "A Resource-Efficient Multi-Modal Emotion Recognition System Using Advanced Deep Learning Model", Master in AI, BTU University, 2025.

- Duslapodu Giri Soundarya, "A Consistent Rule-Focused Approach to Multilingual Text Compliance Verification with mBERT Model", Master in AI, BTU University, 2025.
- Mehwish Farooq, "From static to dynamic facial expression Recognition", Master in AI, BTU University, 2025.
- Arshad Muhammad Faheem, "Automated Image Processing for Crack Detection and Coating in Raney Nickel-Coated Electrodes", Master in AI, BTU University, 2025.

1 Introduction

1.1 Context and Motivation

Facial expressions are a fundamental and natural means of conveying human emotions and intentions [1, 2]. As a key form of non-verbal communication, they play a crucial role in social interaction and the understanding of affective behavior. Psychological studies have shown that facial expressions communicate emotional information more effectively than other cues such as speech, body posture, or gestures [3]. It is estimated that verbal content contributes to only about one-third of emotional communication, while non-verbal signals, especially those conveyed through the face, account for the remaining two-thirds. This highlights the face as one of the most salient and consistent channels through which emotional states are expressed and interpreted [4]. Facial expressions have long attracted scientific interest. For example, Ekman and Friesen's cross-cultural study led to the identification of six basic emotions: happiness, sadness, anger, fear, surprise, and disgust, each associated with distinct facial muscle patterns and recognized across different cultures [5, 6]. Contempt was later added as a seventh basic emotion [7]. However, more recent findings in neuroscience and psychology argue that this model is culture-specific and not universal [8, 9].

In recent years, Facial Expression Recognition (FER) has gained significant attention in the fields of Artificial Intelligence (AI) and Computer Vision (CV). FER refers to the task of automatically identifying a person's emotional state from facial images or video sequences [2, 10]. It has a wide range of applications, including healthcare, education, and human–computer interaction. A particularly impactful application lies in the driving context, where recognizing a driver's emotional state can help prevent accidents and enhance road safety by supporting intelligent systems such as Advanced Driver Assistance Systems (ADAS) [11]. Facial expression recognition has made significant progress, starting with handcrafted features [12, 13, 14, 15] and advancing through deep learning approaches [16, 17, 18, 19], particularly those based on Convolutional Neural Network (CNN) and Vision

Transformer (ViT). Despite these advancements, several challenges persist. Many FER models still struggle under real-world conditions, where variations in lighting, head pose, occlusion, and expression intensity can significantly degrade performance. Furthermore, some of the most accurate methods, including multimodal approaches [20, 21, 22, 23, 24], are computationally expensive, limiting their deployment in resource-constrained environments. While lightweight FER models have been increasingly investigated to address this limitation [22, 25, 26], achieving an optimal trade-off between computational efficiency and recognition accuracy remains a key challenge [20, 23]. These ongoing challenges underscore the need for FER approaches that improve recognition performance under practical conditions while maintaining computational efficiency.

1.2 Research Objectives and Questions

Motivated by the limitations and challenges discussed in the previous section, the overarching objective of this thesis is to advance facial expression recognition by developing methods that improve recognition performance while ensuring computational efficiency under practical conditions, with particular emphasis on driving scenarios and dynamic, in-the-wild environments. To achieve this, the research is guided by the following specific objectives:

- Review and analyze the limitations of existing and state-of-the-art FER methods, including those specifically designed for driver FER systems.
- Develop deep learning models that enhance spatial and contextual representations for FER in challenging conditions, particularly within driving environments.
- Propose a lightweight and efficient FER architecture suitable for real-time deployment in resource-constrained settings, such as embedded automotive systems.
- Explore the integration of visual and textual information using Vision Language Models (VLMs) to improve semantic understanding and recognition performance in driver-focused FER.
- Extend the application of FER beyond driving scenarios to dynamic, in-the-wild environments by incorporating temporal modeling and efficient multimodal learning strategies.

- Conduct extensive experimental evaluations of the proposed methods in comparison with state-of-the-art approaches, and derive analytical insights to inform future research.

These objectives are addressed through the following research questions:

- **RQ1:** What are the main limitations of existing FER methods, including those applied in driving environments?
- **RQ2:** How can Vision Transformer-based architectures improve FER performance under challenging conditions such as illumination changes, pose variations, and occlusions?
- **RQ3:** What approaches can be used to develop lightweight FER models that maintain recognition accuracy while reducing computational cost, particularly in driving environments?
- **RQ4:** How can the integration of multimodal cues, particularly the combination of textual and visual information through VLMs, enhance semantic understanding and classification accuracy in FER?
- **RQ5:** How can vision language models be effectively adapted for dynamic, in-the-wild facial expression recognition using Parameter-Efficient Fine-Tuning (PEFT) techniques?

1.3 Research Contributions

Aligned with the research objectives and questions, this thesis presents a set of contributions aimed at enhancing facial expression recognition under challenging conditions. The key contributions are summarized as follows:

- **Comprehensive Review of FER:** A thorough overview of FER is presented, with a specific focus on its application in driving environments [27]. This includes a review of foundational concepts, application domains, public datasets, evaluation protocols, and commonly used methods, covering both traditional and deep learning-based approaches. A detailed analysis of FER methods, both general and driver-specific,

is conducted to identify key limitations that motivate the research contributions presented in this thesis.

- **Vision Transformer-Based FER for Driving Environments:** DFER-GCViT, a Global Context Vision Transformer architecture tailored for FER in driving scenarios, is proposed [28]. This model captures global facial features more effectively than conventional CNN-based approaches and addresses key challenges such as pose variation, illumination changes, and partial occlusions common in in-vehicle settings. Experimental validation on a driver-specific dataset demonstrates superior recognition performance compared to existing FER models.

- **Lightweight Hybrid CNN–ViT Architecture:** ShuffViT-DFER, a lightweight dual-stream architecture combining two pre-trained models, ShuffleNet V2 (CNN) and EfficientViT-M2 (ViT), is introduced [29]. This architecture is designed to balance accuracy and efficiency, making it suitable for FER in driving environments. Experiments conducted on benchmark datasets confirm the effectiveness of the proposed approach, achieving superior results compared to existing state-of-the-art methods.

- **Multimodal FER Using Vision-Language Models:** CLIVP-FER, a novel multimodal framework leveraging vision-language models (VLMs), specifically the CLIP model, is proposed to enhance FER in driving environments [30]. The approach integrates both visual and textual features to improve semantic understanding and reduce expression ambiguity. Experimental results on a benchmark driver FER dataset show that CLIVP-FER outperforms existing state-of-the-art methods, with additional cross-database evaluations highlighting its generalization capabilities.

- **Parameter-Efficient Dynamic FER:** PE-CLIP, a parameter-efficient fine-tuning framework that adapts the CLIP model for video-based FER in real-world settings, is proposed [31]. By incorporating adapters and multimodal prompt tuning, the model significantly reduces the number of trainable parameters while maintaining high accuracy. Evaluations on benchmark datasets demonstrate the effectiveness and efficiency of PE-CLIP, achieving competitive performance with fewer parameters.

Each contribution addresses a specific challenge in FER, ranging from robustness and efficiency to semantic understanding and adaptability. Collectively, they support the overarching objective of this thesis and demonstrate a coherent progression from constrained,

domain-specific settings to broader, in-the-wild environments that better reflect the real-world conditions.

1.4 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 provides the theoretical and technical background of FER, introducing key concepts, relevant terminology, and major application domains. It reviews publicly available datasets, evaluation metrics, and both traditional and deep learning techniques. The chapter concludes with a comprehensive literature review and a discussion of current limitations, including those specific to driving contexts. This chapter constitutes the first contribution of the thesis.

Chapter 3 presents the second contribution, which enhances FER performance in driving environments using a Vision Transformer-based architecture. It outlines the motivation, reviews related work, and details the proposed DFER-GCViT model. The chapter includes the experimental setup and performance evaluation.

Chapter 4 introduces the third contribution, which focuses on balancing recognition performance and computational efficiency in FER for driving environments. The chapter begins with a discussion of the motivation and a review of related work on lightweight FER models. It then presents the ShuffViT-DFER architecture, which combines CNN and ViT-based components. The experimental setup is described, followed by performance analysis and an ablation study that evaluates the contribution of each component.

Chapter 5 presents the fourth contribution, which explores a multimodal approach to FER by leveraging vision-language models. The chapter outlines the motivation and reviews related work in this area. It describes the proposed CLIVP-FER framework, which integrates visual and textual modalities. The experimental setup is outlined, and an analysis of the obtained results is presented, including cross-database evaluation and ablation analysis.

Chapter 6 introduces the fifth and final contribution, extending FER beyond the driving context to dynamic, in-the-wild scenarios. The chapter begins by outlining the motivation for this shift and reviewing related work on vision-language models and parameter-efficient

fine-tuning techniques. It presents the PE-CLIP framework, describing its architecture and key components. A detailed experimental evaluation is provided, including implementation details, performance comparisons, and an ablation study.

Chapter 7 concludes the thesis by summarizing the main contributions and findings. It discusses the limitations of the proposed work and suggests directions for future research.

2 Facial Expression Recognition: Background & Literature Review

2.1 Introduction

Facial expressions play a vital role in human social interaction, serving as one of the primary channels for conveying emotions and intentions. In recent years, FER has witnessed significant advancements, driven by the development of automated systems capable of recognizing expressions from images and videos. These systems leverage artificial intelligence techniques, including traditional machine learning and deep learning approaches, to achieve increasingly robust performance.

This chapter provides a comprehensive overview of facial expression recognition, with a particular focus on its application in driving contexts. Section 2.2 introduces the fundamental concepts and key terminology in FER, along with a discussion of its applications across various domains. It also describes the general pipeline of FER systems. Section 2.3 reviews publicly available datasets and outlines standard evaluation metrics and visualization-based interpretability techniques used to assess model performance. Section 2.4 presents the core traditional and deep learning techniques employed in FER tasks. Section 2.5 offers an in-depth literature review, summarizing existing methods in both general and driver-specific FER, and highlighting the main challenges encountered by current approaches. Finally, Section 2.6 concludes the chapter by summarizing the key findings.

2.2 Fundamentals of Facial Expression Recognition

Facial expressions are among the most natural and essential ways through which humans communicate emotions in daily life. Alongside voice, body posture, and gestures, facial expressions serve as a primary non-verbal channel for conveying affective information.

Psychological studies have long emphasized their crucial role. For instance, Mehrabian's research revealed that approximately 55% of emotional communication is transmitted through facial expressions, compared to 38% via speech characteristics and only 7% through spoken language [32]. This underscores the significant expressive power of the human face and its pivotal role in emotional understanding. Scientific interest in facial expressions has deep historical roots. Early beliefs in physiognomy, which held that character traits could be inferred from facial appearance, were eventually challenged by more scientific approaches [33, 34]. In the 17th century, John Bulwer produced one of the first systematic studies on facial movements [35], followed by Charles Bell's work linking facial expressions to sensory and motor functions [36]. A notable advancement came with Duchenne de Boulogne, who conducted pioneering experiments by electrically stimulating facial muscles to induce specific expressions, thereby systematically demonstrating the muscular basis of emotions. Examples of his work are illustrated in Figure 2.1. However, the scientific study of facial expressions was truly established by Charles Darwin. In his seminal work *The Expression of the Emotions in Man and Animals* (1872), Darwin argued that facial expressions are evolutionary adaptations shared across species [1].



Figure 2.1: Examples of facial expressions induced through electrical stimulation by Duchenne de Boulogne [36].

Building on Darwin's ideas, researchers such as Paul Ekman made foundational contributions to the field. In particular, he proposed that there are universal basic emotions, such as happiness, sadness, anger, fear, surprise, and disgust, that are consistently associated with specific facial muscle movements across different cultures [37, 38, 39]. His research firmly established facial expressions as reliable indicators of emotional states and laid the groundwork for automatic facial expression recognition systems [36]. However, findings

in neuroscience and psychology challenge the universality of these expressions, suggesting that emotional expressions may be influenced by cultural and individual differences [8, 9]. The importance of FER extends beyond theoretical interest. Studies have shown that facial expressions not only dominate interpersonal communication but also play a critical role in practical domains where understanding emotional cues enhances interaction, such as healthcare [40], education [41], marketing [42], and driving safety [11]. Indeed, facial expression recognition has become a key technology in the broader field of affective computing, which aims to equip machines with the ability to understand and respond to human emotions.

Given the strong connection between facial expressions and underlying emotions, FER systems aim to automatically detect and interpret these non-verbal cues from images or videos. FER typically focuses on recognizing emotions in two main forms: discrete emotions, also known as categorical emotions, as defined by Ekman (e.g., happy, sad), and dimensional emotion representations, in which emotional states are represented as points within a two- or three-dimensional space (e.g., valence and arousal levels) [43]. Additionally, many FER approaches analyze the activation patterns of specific facial muscles, known as Action Units (AUs), to provide a fine-grained understanding of facial behavior.

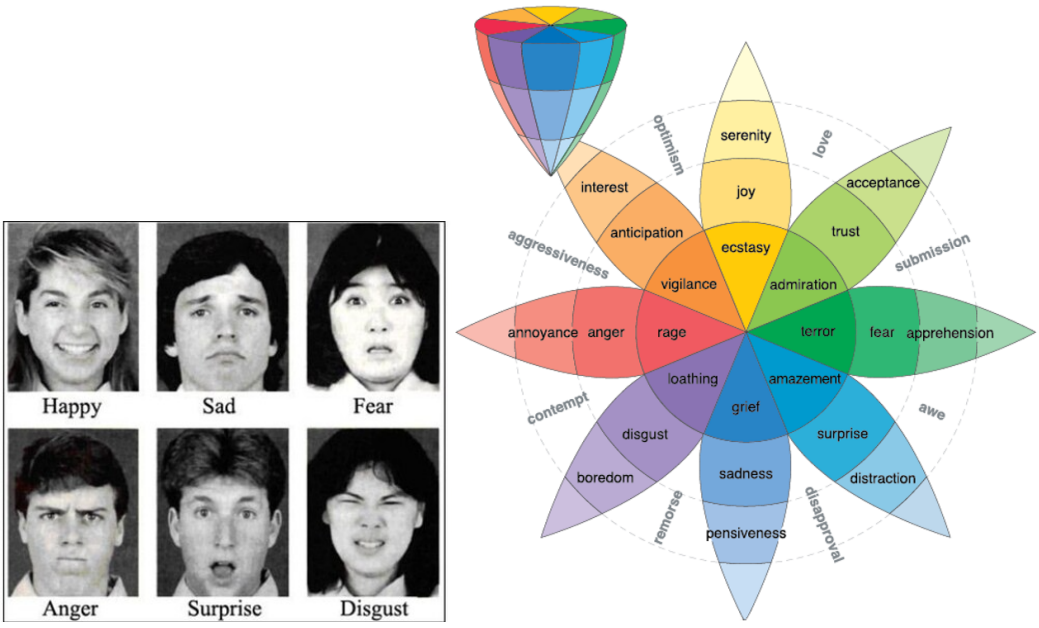
In this section, we provide an overview of these key concepts, which form the fundamental of facial expression recognition systems. We then explore the diverse applications of FER, with particular emphasis on its use in driving contexts. Finally, we outline the general workflow involved in developing a typical FER system.

2.2.1 Key Concepts and Terminologies

Facial Expression Recognition aims to automatically identify human emotions by analyzing facial cues. To understand the mechanisms behind FER, it is important to introduce the fundamental concepts that structure this field. These concepts relate to how emotions are modeled and interpreted from visual facial signals. In the following, we briefly present the main models and representations used in FER.

Discret Emotions: Refers to a set of basic emotional states expressed and identified through facial expressions. According to Ekman's theory [39], these emotions include anger, fear, disgust, happiness, sadness, and surprise, and are considered innate, not influenced by culture or language. Discrete emotions serve as the foundation for many

FER classification tasks, where the goal is to classify facial expressions into predefined emotion classes. Beyond Ekman's model, Plutchik's emotional wheel [44] is another widely adopted framework, organizing eight basic emotions (joy, trust, fear, surprise, sadness, anticipation, anger, and disgust) into a circular structure that captures the relationships among emotions as well as their intensity variations [3]. An illustration showing both models is provided in Figure 2.2. These frameworks have significantly influenced FER research by offering structured representations of emotional states.



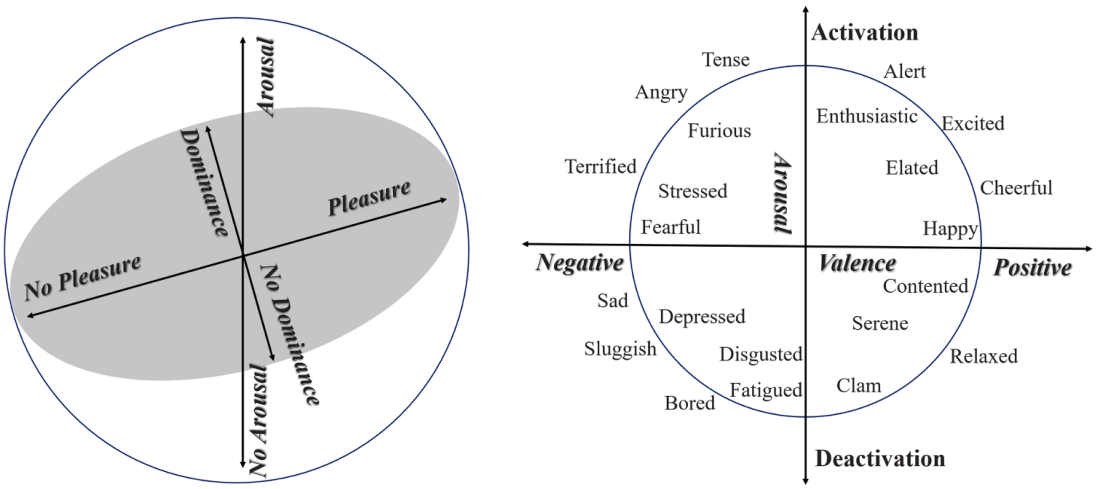
(a) Basic emotions.

(b) Plutchik's Wheel of Emotions.

Figure 2.2: Examples of discrete emotion models: Ekman's six basic emotions [45], Plutchik's wheel of emotions [3].

Dimensional Emotion: Unlike discrete emotion models, dimensional models represent emotions as points within a continuous, multidimensional space. One of the most widely recognized frameworks is the Pleasure-Arousal-Dominance (PAD) model [46], in which pleasure (or valence) reflects the positivity or negativity of an emotion, arousal indicates its intensity or activation level, and dominance describes the degree of control in a given situation. Among these dimensions, valence and arousal are the most commonly adopted in FER. Russell's Valence-Arousal (V-A) circumplex model [43] simplifies emotional representation into a two-dimensional space, where valence defines emotional polarity (from negative to positive) and arousal captures emotional intensity (from low to high).

This approach is particularly well-suited for FER regression tasks, where the objective is to estimate the degree of an emotion rather than classify it into discrete categories. An illustration of both the PAD and V-A models is provided in Figure 2.3.



(a) Pleasure-Arousal-Dominance (PAD) model. (b) Valence-Arousal Space of Emotions.

Figure 2.3: Dimensional emotion models. (a) PAD and (b) V-A models [3].

Action Units (AUs): represent standardized facial muscle movements associated with specific facial expressions. They are formally defined within the Facial Action Coding System (FACS), a widely adopted framework developed by Ekman and Friesen [47]. FACS provides a systematic method for decomposing facial expressions into individual muscle activations, allowing for precise analysis of facial dynamics. Each action unit corresponds to a distinct facial movement, such as AU1 (Inner Brow Raiser) or AU12 (Lip Corner Puller), as illustrated in Figure 2.4.

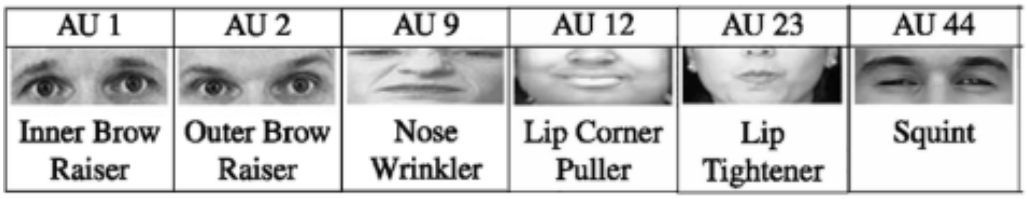


Figure 2.4: Illustration of lower and upper facial AUs as defined in the FACS. Adapted from [36].

2.2.2 Application Domains of FER

Facial expression recognition has emerged as a crucial technology, with a wide range of applications across various domains, reflecting its increasing importance in both research and industry. In the psychological and healthcare domains, FER is applied to monitor emotional states and assist in the early detection of mental health conditions such as depression or stress through wearable devices and mobile applications [48, 49]. It is also used to aid the diagnosis and treatment of neurological disorders like Parkinson's disease by analyzing facial dynamics [50], and to assess pain intensity in clinical conditions [51, 52]. In security and surveillance, FER helps identify abnormal or suspicious behaviors in high-risk environments [53]. In human-robot interaction and virtual assistants, FER enhances the ability of systems to recognize emotional states of users, enabling more natural and intuitive communication [54, 55]. In education, FER improves the learning experiences by detecting emotional cues such as frustration or disengagement, allowing adaptive learning platforms to personalize content and increase student engagement [56, 57]. In entertainment, FER has been applied to video games to enhance user immersion by adjusting elements like difficulty or music based on real-time emotional responses [36, 58]. Beyond entertainment, FER also plays a significant role in monitoring applications, particularly in driving contexts. Since emotions directly impact driver safety and comfort, detecting drivers' emotional states is essential for developing safer and more responsive transportation systems [11]. Given that a substantial part of this thesis focuses on facial expression recognition in the driving context, the following section provides a more detailed overview of this specific application domain.

FER in the Driving Context

In the context of driving, several FER systems have been increasingly deployed to enhance driver monitoring systems, supporting tasks such as detecting fatigue and drowsiness, identifying distraction, and recognizing facial expressions or emotional states. These systems are typically designed and implemented to integrate into advanced driver assistance systems or other monitoring systems and safety as embedded systems in vehicles. For instance, to detect driver fatigue and drowsiness, multiple studies have focused on analyzing facial features such as eye closure and facial movements to determine the driver's vigilance state [59, 60, 61]. Other studies have used FER to detect driver distraction by tracking

off-road eye movements and gaze direction [62]. Furthermore, FER has been used for driver emotion recognition, aiming to detect negative and positive emotions such as anger or happiness, which can affect driver behavior and lead to driving distractions [63, 64, 65]. Based on the numerous applications of FER in the driving context, driver emotion recognition has become an important research focus and is a central theme of this thesis. As discussed earlier, it involves analyzing facial expressions to determine the driver's emotional state, typically classified into one of the basic emotions identified by Ekman [39], as illustrated in Figure 2.5, where the captured driver's face is classified into a target emotion. This is particularly relevant in driving scenarios, where detecting negative emotions such as anger can help to prevent aggressive driving behavior and reduce the risk of accidents. Advanced driver assistance systems or other monitoring systems can use this information to issue timely alerts or take corrective actions, thereby improving road safety. This makes FER in driving a significant area of research for the automotive industry and related fields.



(a) Camera installed in a car. (b) FER for a driver.

Figure 2.5: Example of driver's facial expression recognition using a camera installed inside a car. The figure is only for illustration and is partially generated using OpenAI's ChatGPT-4.

It is crucial to emphasize that driver FER is a specialized subdomain of the broader FER field. The fundamental principles and techniques developed for general FER, including feature extraction, classification models, and evaluation methods (discussed in the following section), are largely applicable to driving scenarios. However, Driver's FER presents additional challenges arising from the unique conditions inside vehicles, such as varying

lighting conditions, potential occlusions (e.g., sunglasses or hands covering the face), head pose variations, and the need for real-time processing. These constraints highlight the link between general FER and its tailored adaptation for the driving context.

2.2.3 General workflow of FER systems

Facial expression recognition systems typically follow a structured pipeline composed of several key processing stages, each essential for accurately interpreting human emotions from facial data. Despite variations in application domains and environmental conditions, the core steps, pre-processing, feature extraction, and expression recognition (classification), remain fundamental across most FER architectures. The standard workflow is illustrated below in Figure 2.

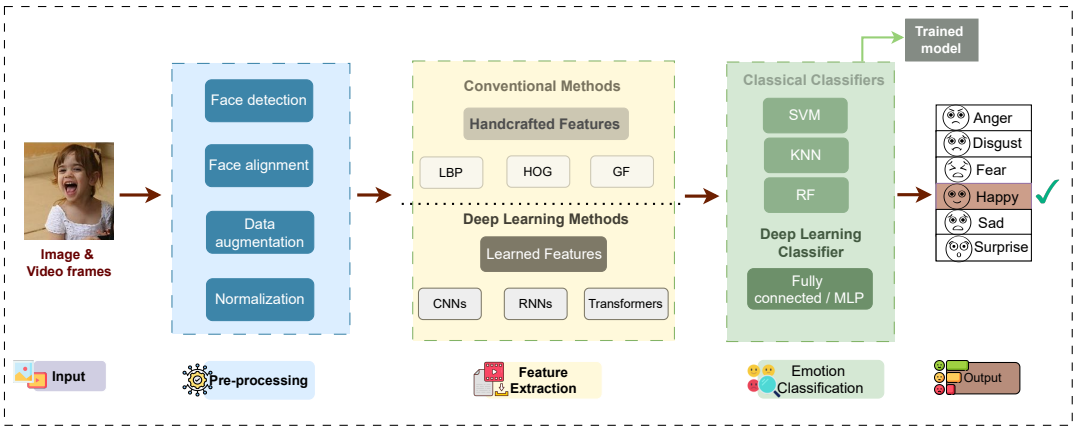


Figure 2.6: General workflow of a facial expression recognition (FER) system. The pipeline includes input data, pre-processing (face detection, alignment, augmentation, and normalization), feature extraction using either handcrafted or deep learning-based methods, and emotion classification using classical or deep learning classifiers.

In the following, we present a detailed explanation of each stage in the FER pipeline:

Pre-processing

Pre-processing is a critical initial stage in FER systems. It involves removing irrelevant information and standardizing facial input (images or videos) to facilitate the detection of relevant features. This step addresses various distortions caused by both environmental

and subject-related factors, such as background clutter, variable lighting, occlusions, and head pose variations [66]. The quality of pre-processing directly influences the accuracy of subsequent stages, including feature extraction and classification. The pre-processing pipeline typically involves the following steps:

- Face detection, which identifies and isolates facial regions from each frame. Traditional methods such as the Viola–Jones algorithm and Haar cascade classifier [67] offer fast detection, while more accurate modern approaches use deep learning-based models like the Multi-task Cascaded Convolutional Network (MTCNN), capable of detecting faces with high precision even under challenging conditions [68, 69].
- Face alignment, used to standardize facial geometry by locating key landmarks (e.g., eyes, nose, mouth) and correcting facial orientation using affine transformations or facial landmark detectors such as Dlib or MTCNN. This step ensures spatial consistency across input samples, which is particularly important in FER due to the localized nature of emotional cues.
- Normalization, which reduces variability caused by lighting conditions, head pose differences, and pixel intensity inconsistencies. Common techniques include illumination normalization, adjusting brightness and contrast using methods like histogram equalization [70], and pose elimination, which aligns facial features to a standard orientation. In addition, general feature scaling methods such as Z-score normalization, Min-Max scaling, and unit vector normalization are employed to standardize pixel values, enhancing numerical stability and model performance.
- Data augmentation, applied to artificially expand the training dataset and improve model generalization. Typical augmentation techniques include horizontal flipping, random cropping, rotation, brightness and contrast adjustments, and the addition of noise, ensuring that expression-related features remain intact [71].

Feature extraction

Feature extraction is the process of extracting relevant information from pre-processed facial data and converting it into a compact, discriminative numerical representation suitable for emotion classification. The quality of the extracted features plays a critical role in the overall performance of FER systems, as it directly affects the system's ability to recognize

and distinguish emotional patterns. Feature extraction methods are generally categorized into two broad groups: handcrafted features and learned features, depending on whether they are manually engineered or automatically learned by deep learning models.

In conventional FER systems, handcrafted features have been extensively used and are typically classified into two types: appearance-based and geometry-based. These features can also be categorized as global, which capture information from the entire facial region, or local, which focus on specific regions of interest, often corresponding to expressive areas such as the eyes, mouth, or eyebrows.

- Appearance features: represent texture and intensity patterns in the face. These features are typically extracted from either the entire facial region (global features) or selected areas of interest (local features). Common appearance descriptors include Local Binary Patterns (LBP) [12], which encode local texture by thresholding pixel neighborhoods, Histogram of Oriented Gradients (HOG) [13], which capture edge and gradient structure, and Gabor filters [72], which are effective for modeling spatial frequency and orientation. These methods have proven useful in capturing expression-related variations, especially in static RGB images. To account for the spatial importance of facial regions, researchers have proposed grid-based approaches where filters are applied locally. For example, Gabor filters can be applied to cells in a grid over the face, with the resulting feature maps encoded into statistical summaries [72]. Other studies, such as [73], used LBP histograms across varying block sizes and employed Principal Component Analysis (PCA) for dimensionality reduction. However, these global approaches may miss subtle, localized expression cues. To improve upon this, the authors in [74] extracted region-specific features by dividing the face into functional subregions (e.g., eyes, mouth), leading to enhanced recognition accuracy and lower feature dimensionality. In dynamic settings where temporal information is relevant, temporal extensions of appearance descriptors are employed. For example, LBP-TOP (Local Binary Patterns from Three Orthogonal Planes) extends the LBP operator into the spatio-temporal domain by applying it over three orthogonal planes [75]. Alternatively, [76] proposed extracting HOG features from localized facial regions identified by landmarks rather than the entire face, enhancing robustness to pose and scale variations. Although local appearance features are more selective, they require prior knowledge of the most expressive facial regions. As such, global descriptors are often preferred for simplicity [36].

- Geometric features: describe facial expressions by modeling the spatial relationships between key face landmarks, specific points that correspond to prominent facial features such as the eyes, eyebrows, nose, and mouth. These relationships include distances, angles, and shapes formed by points such as the eyes, eyebrows, nose, and mouth. While appearance descriptors capture surface-level changes (e.g., wrinkles or shading), geometric features are more directly associated with underlying facial muscle movements, often making them more interpretable in terms of action units. A common method involves computing Euclidean distances and angles between landmark points to form a feature vector. The authors [77], for example, extracted static and dynamic features from 52 landmarks and used machine learning-based classifiers to recognize expressions across frames. Other studies have integrated depth data for improved landmark tracking. For instance, Sujono and Gunawan [78] combined Kinect-based depth sensing with Active Appearance Models (AAM) and fuzzy logic for expression classification. To address shape instability in local regions, descriptors such as the Local Prominent Directional Pattern (LPDP) [79] have been developed, encoding directional cues around landmarks for more robust representation. Additionally, geometric features can be derived from low-level structural properties like corners, edges, or contours, using techniques such as corner detection or curvelet transforms to describe facial shape and symmetry at multiple scales [3, 71]. Although geometric features tend to be more robust to lighting variations and less data-intensive, their performance is highly dependent on accurate facial landmark detection, making them sensitive to occlusions, extreme head poses, or low-resolution input. As such, several FER systems combine both appearance-based and geometry-based features to leverage the strengths of each [80].

With the advancement of deep learning, learned features have become the dominant approach in facial expression recognition. Unlike handcrafted features, which rely on manually designed descriptors, learned features are automatically extracted from data using Deep Neural Networks (DNNs). These models employ hierarchical architectures composed of multiple nonlinear transformations to derive high-level, abstract representations [66]. Among the most widely used are CNNs [81, 82], which learn spatial features directly from raw pixel inputs, Recurrent Neural Networks (RNNs) [83], which are effective at modeling temporal dependencies across video frames, and hybrid models that combine both spatial and temporal learning [84, 85]. More recently, transformer-based architectures incorporating attention mechanisms have been introduced, allowing the model to dynamically focus on

the most expressive regions of the face [28, 86]. The architecture of these techniques is detailed in Section 2.4.

Classification

The final step in a FER system is to classify the visual input into one of the basic emotion categories. This can be achieved using traditional algorithms, deep learning models, or hybrid approaches. In traditional methods, feature extraction and classification are handled separately, with commonly used classifiers including SVMs [87], Random Forests [15], AdaBoost [14], and MLPs [88]. For instance, in [74], the authors used SVM to recognize the facial expressions from a single image frame based on the fusion of appearance and geometric features. In contrast, DL-based methods [9] perform FER in an end-to-end manner, where networks like CNN [89], VGG [90], and 3D CNNs [91] automatically extract features and classify expressions using fully connected layers. For example, a fully connected layer is used to classify expressions after extracting features using a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former) [86]. Alternatively, hybrid approaches use DL-feature extractors and apply conventional classifiers such as SVMs or Random Forests to the extracted deep features offering greater flexibility and potentially improved performance [92, 93], such as in [92] a hybrid approach developed by combining a CNN and SVM, where the CNN used to extract and refine visual features and SVM for the classification.

2.3 Datasets, Evaluation Metrics, and Interpretability in FER

Facial expression datasets form the cornerstone of training and evaluating FER methods. The development and validation of FER algorithms critically depend on the availability of diverse, annotated datasets that reflect varying conditions and expressions. These datasets support performance evaluation through quantitative metrics and interpretability tools, such as visualization techniques.

In this section, we present an overview of the most widely used FER datasets in the literature, covering both general-purpose collections and those tailored specifically to the driving context. We begin by outlining the distinguishing characteristics of these datasets. Following this, we review publicly available datasets that are frequently used

for facial expression classification, with a focus on those containing basic emotions as labels. Additionally, we summarize widely adopted quantitative evaluation metrics and introduce common visualization techniques employed in FER research. Figure 3 presents example images from various datasets used in this thesis. Tables 2.1 and 2.2 summarize the key characteristics of the datasets, including the number of subjects, number of samples, resolution, labeled emotions, collection conditions, elicitation methods, data modalities, and other relevant information.

2.3.1 Commonly Used FER Datasets

Various datasets have been collected and utilized in the literature to support the advancement of FER techniques, particularly those addressing facial expression classification tasks. These datasets differ in key characteristics as illustrated in Figure 2.7, and understanding these differences is essential for selecting appropriate data sources and evaluating model performance accurately.

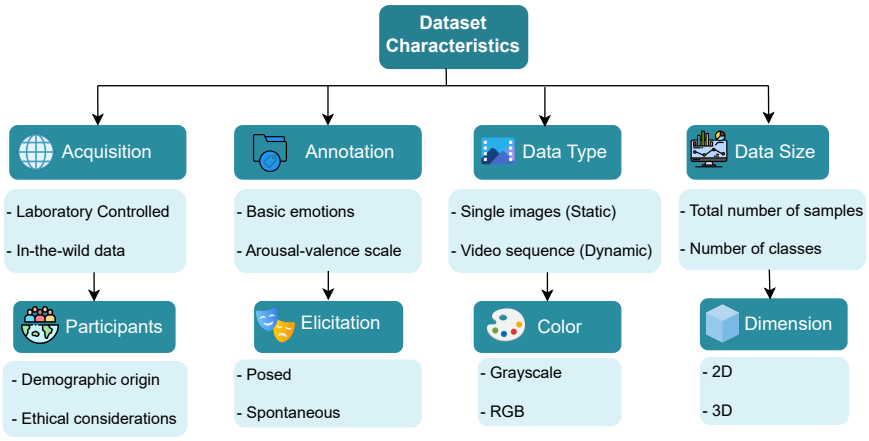


Figure 2.7: Overview of the facial expression dataset characteristics.

Some datasets were obtained in a laboratory-controlled environment where data was captured under restricted conditions with highly regulated lighting, camera position, and other factors to minimize variability. These datasets usually contain 2-dimensional (2D) RGB images with basic emotions, and the majority of the datasets used in FER research are collected in this type of environment, such as CK+ [94], MMI [95], and MPI Large [4]. In contrast, some datasets, such as RAF-DB [96] and FER2013 [97], were collected from the

wild environment in unconstrained settings closer to real-world conditions. For instance, from the web or movies, which better represent real-world conditions. These datasets are more challenging to use than those collected in controlled laboratory settings, but are essential for developing FER systems applicable to real-world scenarios. For the labeled approach, different methods are used to label emotions in the data, including discrete approaches such as basic emotions, dimensional approaches like Valence/Arousal, and the Action Units approach, as discussed in section 2.2.1, which can be combined in certain datasets. Furthermore, datasets differ in terms of data modality and size, for example, the number of images, video clips, or individual frames, the number of subjects involved in the data collection process, and the variability of subjects in terms of age, gender, and ethnicity (e.g. Asian, American, and Caucasian). In addition, data collection methods vary, including both posed and spontaneous expression elicitation.

In posed methods, participants are instructed to exhibit one of the basic expressions (happiness, sadness, anger, fear, disgust, or surprise). In contrast, spontaneous methods involve recording facial expressions naturally elicited in response to real-world situations, such as watching a movie or having a conversation [98]. By understanding these characteristics, researchers can better evaluate the strengths and limitations of different datasets and choose the most appropriate datasets for their specific research needs.

General FER Datasets

We first present the most widely used image-based and video-based datasets, including both lab-controlled and in-the-wild settings.

Lab-Controlled Datasets:

Extended Cohn-Kanade (CK+) [94]: is a widely used dataset for evaluating facial expression recognition systems. The dataset includes 593 video sequences and still images that feature seven different facial expressions, including six basic emotions and contempt. The images in the dataset have a resolution of 640 x 480 pixels or 640 x 490 pixels, with 8-bit precision grayscale values. CK+ sequences display a shift from a neutral expression to a peak expression, and the standard method for selecting data is to extract the first frame of each sequence and the last one to three frames with peak formation. Researchers have typically used the dataset with k-fold cross-validation experiments, where k values of 5, 8, or 10 are common choices.

MMI dataset (MMI) [95]: is a dataset comprising over 2900 high-quality videos and images of 75 subjects in a controlled environment. The sequences in this dataset are onset-apex-offset labeled, which means that each sequence starts with a neutral expression, peaks near the middle, and returns to a neutral expression. To conduct experiments, the most common approach is to utilize the first frame and three peak frames of each frontal sequence for 10-fold cross-validation. Additionally, the database has annotations for audio-visual laughter.

The Japanese Female Facial Expression(JAFFE) [99]: The JAFFE dataset consists of 213 images featuring seven facial expressions, six basic emotions, and one neutral expression, which were posed by ten Japanese females. The images are of size 256 x 256 pixels. However, the dataset poses a challenge as it offers limited examples per subject and expression, and all of the images are usually used for evaluation.

Karolinska Directed Emotional Faces(KDEF) [100]: The KDEF dataset consists of 4900 high-quality images of 70 individuals displaying seven facial expressions (including neutral) from five different angles. Each image has a resolution of 562 pixels $\times$ 762 pixels. To assess their methods' performance, researchers frequently employ the entire KDEF dataset, using either a train-test split or k-fold cross-validation.

The Large MPI Facial Expression database(MPI) [4]: The emotional and conversational facial expressions database has been verified and includes 55 expressions displayed by 19 German individuals, comprising ten females and nine males. The expressions are elicited using a method-acting protocol, resulting in both natural and clearly defined facial expressions. All facial expressions are captured from three different camera viewpoints, with three repetitions, two intensities, and a detailed frame annotation provided. The database can generate dynamic and static versions.

In-the-Wild Datasets:

Facial Expression Recognition 2013 dataset(FER2013) [97]: The FER2013 is a database of grayscale face images collected through the Google image search API. It includes 35,887 images, with 28,709 in the training set, 3,589 in the validation set, and 3,589 in the testing set. The images are resized to 48x48 pixels and mislabeled frames are removed. This dataset is primarily used for evaluating FER methods using the specific split of the dataset, although some researchers have used other train-test splits.

Real-world Affective Faces Database(RAF-DB) [96]: is a vast collection of facial images comprising 29,672 real-world samples sourced from the internet. Each image has

been independently labeled by around 40 annotators with seven classes of basic emotions and 12 classes of compound emotions, defined as combinations of two or more basic emotions. The dataset also provides automatic and accurate landmark locations. To evaluate the dataset, the 15,339 images from the basic emotion set are split into two groups of 12,271 training samples and 3,068 testing samples.

Affect from the InterNet(AffectNet) [21]: AffectNet is an extensive facial expression dataset consisting of more than one million images collected from diverse search engines using emotion-related keywords. the dataset features manually annotated labels for eight expressions, including neutral, happy, angry, sad, fear, surprise, disgust, and contempt, with the intensity of valence and arousal measured for each image. AffectNet is the largest dataset available that offers facial expressions in both categorical and dimensional models.

FERV39k (A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos) [101]: It's composed of 38,935 video clips labeled with seven classic expressions, including neutral, anger, contempt, disgust, fear, happiness, and sadness. The dataset covers 22 fine-grained scenes in four isolated scenarios, making it the first dynamic FER dataset with such a scenario-scene division and with 39K clips. Moreover, it has cross-domain supportability, which allows it to be used across multiple domains.

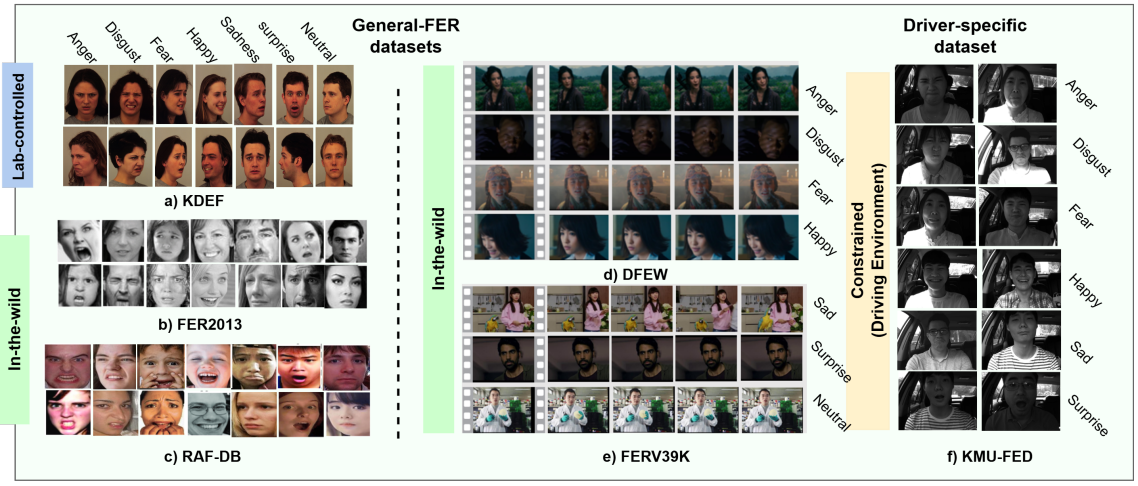


Figure 2.8: Some samples from FER datasets used in this thesis. The datasets include both lab-controlled (a: KDEF [100]) and in-the-wild general FER datasets, comprising image-based datasets (b: FER2013 [97], c: RAF-DB [96]) and video-based datasets (d: DFEW [102], e: FERV39K [101]), as well as a driver-specific dataset collected in a constrained driving environment (f: KMU-FED [103]). Expressions vary across datasets and include basic emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutral. Adapted from [2].

Dynamic Facial Expression in-the-Wild(DFEW) [102]: The DFEW is a facial expression video clip database consisting of 16,372 challenging videos taken from movies. The clips are affected by various interferences like lighting conditions, occlusions, and pose changes. The database has been labeled ten times independently by 12 expert annotators. The resulting database contains abundant and diverse annotations with large quantities of data.

Acted Facial Expressions in the Wild(AFEW) [104]: The Acted Facial Expressions in the Wild (AFEW) dataset is a collection of dynamic temporal facial expressions extracted from movies in close-to-real-world environments. It contains a wide range of subjects aged from 1-70 years, and each clip is annotated with attributes such as age, pose, gender, and expression. The dataset is labeled with seven categories of emotions and is divided into three mutually exclusive sets: Train (773 samples), Val (383 samples), and Test (653 samples) based on movie and actor.

Acted Facial Expressions in the Wild(MAFW) [105]: The MAFW dataset is a large-scale multimodal affective database comprising 10,045 video and audio clips captured in the wild. Each clip is labeled with one or more of 11 emotion categories and includes

Table 2.1: Description of Related-FER Datasets. BEs=basic Emotions, CEs=compound emotions (combination of BEs), while a "N/A" in the table cells, indicates missing information (Part 1).

Datasets	**Subjects**	**Samples / Resolution**	**Emotion Classes**	**Condition**	**Elicitation**	**Data Type**	**Specific Information**
CK+[94]	123 (Age between 18 and 50)	+10.000 images(593 different sequences) 640 × 490 pixels 640 × 480 pixels	6 BEs + neutral and contempt	Static life scenario	Posed & spontaneous	Video-based	A specific feature of the CK+ dataset is its diversity, as it includes both male and female subjects from different racial backgrounds.
MMI[95]	75 (age between 19 and 62)	Over 2,900 video sequences 720 × 576	6 BEs + neutral	Static life scenario	Posed	Video-based	The database is fully annotated with the presence of Action Units (AUs) in videos and partially coded on a frame-level basis.
JAFFE[99]	10	213 images 256 × 256	6 BEs + neutral	Static life scenario	Posed	Image-based	Gray images in the dataset display multiple facial expressions. Michael Lyons, Miyuki Kamachi, and Jiro Gyoba created the JAFFE dataset at Kyushu University.
KDEF[100]	70 (Age between 20 and 30)	4900 images 562 × 762	6 BEs + neutral	Static life scenario	Posed	Image-based	It was initially designed for application in the fields of psychological and medical research. Additionally, all images within the dataset are characterized by their clarity, lack of noise, and uniform off-white background.
MPI Large[4]	19	over 18800 samples of video sequences 768 × 576	55 expressions include 6 BEs,	Static life scenario	spontaneous	Video-based	It's includes both emotional and conversational expressions. -can be used to recognize the facial expression.
FER2013[97]	N/A	35,887 images 48x48 pixel	6 BEs+ neutral	collected from web	Posed & spontaneous	Image-based	-The dataset was first presented as part of the ICML 2013 Challenges in Representation Learning. Each sample in the dataset is designed to reflect real-world scenarios with significant variations in factors such as age and facial direction.
RAF-DB [96]	100	29,672 images	6 BEs + neutral + 12 CEs	Web collection	Posed & spontaneous	Image-based	Diverse samples: ages, ethnicities, lighting, occlusion, and head poses.
AffectNet[21]	450,000 (average age 33.01 years)	Approximately 1,000,000 facial images	6 BEs + neutral and contempt	collected from web	Posed & spontaneous	Image-based	Of the one million images, 450,000 have manually annotated labels for eight basic expressions.
FERV39k [101]	N/A	38,935 video clips 336×504 and 224×224	6 BEs + neutral	Collected from multi-scenario scenes	Posed & spontaneous	Video-based	Automatic video generation, high-quality labeling, and challenges in FER.

Table 2.2: Description of FER-Related Datasets. BEs=basic Emotions, CEs=compound emotions (combination of BEs), while a "N/A" in the table cells, indicates missing information (Part 2).

Datasets	**Subjects**	**Samples / Resolution**	**Emotion Classes**	**Condition**	**Elicitation**	**Data Type**	**Specific Information**
DFEW [102]	N/A	16,372 video clips	6 BEs + neutral	Over 1500 movies/TV shows	Posed & spontaneous	Video-based	Single-label discrete emotion annotations.
AFEW[104]	330(Age between 1 and 70)	1426 sequence videos	6 BEs+ neutral	collected from movies/TV	Posed & spontaneous	Video-based	The database was collected on the basis of Subtitles for Deaf and Hearing-impaired (SDH) and Closed Caption (CC).
MAFW [105]	N/A	10,045 clips high-definition (HD) quality	11 single 32 multiple	collected from Real-world content (reality TV, talk shows, news, variety shows)	spontaneous	Video-based	The dataset also provides bilingual descriptive texts (in English and Chinese) for each clip, capturing contextual details such as facial actions, body movements, and environmental cues—making it useful for both FER and emotion captioning tasks.
DEFE[106]	60 (Age between 19 and 56)	-164 video clips -640 × 480 and 1920 × 1080	Anger, happiness, and neutral (3 emotions)	Dynamic driving scenarios	Spontaneous	Video-based	The dataset enables the comparison between dynamic driving and static life scenarios and presents a challenging aspect of FER due to facial occlusions, such as glasses and hands.
DMD[107]	37 (Age between 18 and 50)	(25TB of raw data) (video recording format)	N/A	Dynamic driving scenarios (real and simulated)	Posed and spontaneous	Video-based	The DMD can potentially be used for tasks such as body pose estimation, action recognition, detection of hands-on-wheel, and estimation of gaze direction.
KMU-FED[103]	12	55 image sequences 1600 × 1200	6 BEs	Dynamic driving scenarios	Posed & spontaneous	Image-based	The most widely available dataset used for the driver's FER task.
ViVa-Face[108]	10	47515 frames 2704 x 1524	N/A	Dynamic driving scenarios	spontaneous	Image-based	-includes images of drivers captured under different illumination conditions and various gaze directions. This dataset was not originally intended for the purpose of recognizing driver expressions.
DrivFace[109]	4	606 images -640 × 480	N/A	Dynamic driving scenarios	spontaneous	Image-based	Consists of images collected over various days, each with three classes of gaze direction (looking right, frontal, and looking left). This dataset was not specifically designed for the purpose of recognizing driver expressions.

textual descriptions of the subject's affective behavior. To ensure label quality, annotations are refined using an expectation-maximization algorithm. MAFW is considered the first multi-modal in-the-wild dataset that includes both compound emotion annotations and emotion-related textual descriptions.

Driver-Specific FER Datasets

Compared to general-purpose FER datasets, driver-specific FER datasets remain relatively scarce. The availability of such datasets is limited, with only a few publicly accessible benchmarks tailored for FER in driving environments. Below, we review the most relevant datasets:

Keimyung University Facial Expression of Drivers (KMU-FED)[103]: The KMU-FED dataset consists of 55 image sequences of drivers captured in real-time using a Near-Infrared (NIR) camera mounted on either the dashboard or steering wheel. The dataset includes 12 subjects exhibiting various facial expressions while driving, with partial occlusions caused by hair or sunglasses, and different illumination conditions such as front, left, right, and backlight. The images have a pixel resolution of 1600 × 1200, making them an ideal resource for studying facial expression recognition in realistic driving conditions. As the dataset does not have a specific train and test split, researchers have used it with a k-fold cross-validation technique. This dataset forms a key component of our thesis, which explores FER in driving context.

Driver Monitoring Dataset(DMD)[107]: DMD dataset is a large-scale dataset comprising 41 hours of RGB, depth, and infrared (IR) video footage collected from 37 drivers in both real and simulated driving scenarios. Captured using a multi-camera setup, it includes annotations for driver distraction, fatigue, gaze-head pose, hand activity, and contextual data. Designed to support deep learning for Driver Monitoring Systems (DMS), DMD addresses the scarcity of comprehensive data for operational DMS tasks. It also includes spatio-temporal annotations for objects, actions, and events, and has been used to recognize driver facial expressions by annotating facial regions with emotional labels [110].

Vision for Intelligent Vehicles Applications Face challenge dataset(ViVa-Face)[108]: This dataset contains images of ten drivers captured in natural driving conditions with varying lighting, using a color video stream at 30 fps and a resolution of 2704×1524 pixels. Although not originally created for driver expression recognition, Verma et al. [111] manually labeled a subset of the data with custom emotion categories (Happy, Talking,

Makeup, Surprise, Neutral, Distracted) based on participant survey consensus, selecting only images with strong agreement on the assigned emotion.

Driver face dataset(DrivFace)[109]: The DrivFace dataset contains 606 facial images (640×480 resolution) collected from four drivers (two male, two female) with varying facial characteristics, such as glasses and beards. It includes annotations for face bounding boxes, facial key points, and gaze direction labels (three classes). Although originally unlabeled for expressions, similar to the ViVa-Face dataset, it was later manually annotated into four expression categories (happy, surprise, talking, neutral) based on survey feedback for experimental use.

A Spontaneous Driver Emotion Facial Expression dataset(DEFE)[106]: The DFEF dataset captures spontaneous driver emotions in response to video-audio stimuli. It includes frontal facial videos of 60 participants (13 females, 47 males), aged 19 to 56, recorded across various driving scenarios. After viewing emotion-eliciting clips, participants performed driving tasks and rated their emotional responses using both dimensional (arousal, valence, dominance) and discrete emotion methods. The dataset also contains demographic details and emotion intensity ratings, making it a valuable resource for developing emotion recognition systems aimed at enhancing driving safety.

2.3.2 Evaluation Metrics and Visualization Techniques

The evaluation of FER methods involves both quantitative metrics and visualization techniques, depending on the nature of the task (classification or regression) and the type of input data (static or dynamic). Since this thesis primarily focuses on classification-based FER, we emphasize evaluation metrics and tools commonly used for assessing categorical emotion recognition from still images and video sequences.

Quantitative Evaluation Metrics

For image-based FER, which treats facial expressions as static inputs, the most commonly used evaluation metrics are accuracy, precision, recall, and F1-score. These metrics are derived from the confusion matrix (defined in the next section) and are based on the following components:

- True Positives (TP): Correctly predicted instances of a given class.
- True Negatives (TN): Correctly predicted instances that do not belong to the given class.
- False Positives (FP): Incorrectly predicted instances as belonging to a class.
- False Negatives (FN): Instances of a class that the model failed to predict correctly.

The metrics are defined as follows:

- Accuracy: measures the overall correctness of the model by computing the proportion of correctly predicted samples out of the total predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

- Precision: indicates the proportion of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

- Recall: measures the proportion of correctly predicted positive observations to all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

- F1-score: provides a harmonic mean between precision and recall, offering a balanced measure, especially in the case of imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

For video-based FER, which involves analyzing sequences of frames, evaluation typically relies on metrics that account for both class imbalance and temporal consistency. Two primary metrics are commonly used:

- Unweighted Average Recall (UAR): calculates the mean recall across all classes without considering their frequency, ensuring fair evaluation across imbalanced

datasets.

$$\text{UAR} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{2.5}$$

where C is the number of emotion classes, and i indexes each class.

- Weighted Average Recall (WAR): is the standard recall weighted by class frequency, giving more influence to dominant classes.

$$\text{WAR} = \sum_{i=1}^{C} w_i \cdot \frac{TP_i}{TP_i + FN_i} \quad \text{with} \quad w_i = \frac{n_i}{\sum_{j=1}^{C} n_j} \tag{2.6}$$

where w_i is the normalized weight of class i, and n_i is the number of samples in class i.

Visualization Techniques

In addition to numerical performance metrics, visual interpretability techniques are increasingly used in FER research to gain deeper insights into model behavior and decision-making processes:

- Confusion matrix is a tabular representation that summarizes a model's prediction results, showing the number of correct and incorrect predictions for each class. It is particularly useful for identifying which classes the model tends to confuse with others [112].
- Heatmap attention visualization is a method used to interpret deep learning models by highlighting the regions of an input, typically an image, that most influence the model's decision. These regions are shown using a color-coded heatmap, where warmer colors (e.g., red, yellow) indicate stronger model attention or interest, and cooler colors (e.g., blue) indicate less relevance. This helps reveal what the model is focusing on during prediction [113].
- t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that maps high-dimensional data into 2D or 3D space, preserving local structure to reveal meaningful patterns and clusters for visualization. It improves on

earlier methods by reducing point crowding and highlighting relationships at multiple scales [114].

2.4 Modeling Techniques for FER

Facial expression recognition has undergone significant development, evolving from traditional handcrafted approaches to deep learning-based end-to-end systems, as discussed in Section 2.2.3. Early FER methods relied on classical machine learning techniques such as SVMs and Random Forests. Although these models performed well in controlled environments, they often struggled in more diverse, real-world conditions. With the rise of deep learning, a specialized class of machine learning, models such as CNNs, RNNs, and Transformers have demonstrated remarkable improvements by automatically learning expressive features from raw visual data. This section provides an overview of both conventional and Deep Learning (DL) approaches and concludes with a discussion on transfer learning.

2.4.1 Classical Machine Learning Models

Machine Learning (ML) is a branch of artificial intelligence that focuses on developing algorithms capable of making predictions or decisions based on data, without being explicitly programmed for each task. At its core, ML seeks to learn a function $f : X \rightarrow Y$ that maps input data X to an output$Y$, with the choice of function dependent on the learning algorithm used. ML techniques are typically grouped into three main categories based on the type of datasets used as experience: supervised learning, unsupervised learning, and reinforcement learning [115]. These categories are briefly defined as follows:

- Supervised Learning: is a machine learning approach where the model is trained on labeled data, meaning each input is paired with a known output. The goal is to learn a mapping function $f : X \rightarrow Y$ that accurately predicts outputs for new, unseen inputs. This learning task is typically divided into classification, when the output is discrete, and regression, when the output is continuous.
- Unsupervised learning: deals with unlabeled data, where the model is not given any target outputs. Instead, it learns by identifying patterns or structures within the

data on its own. Common tasks include clustering, which groups similar data points, or density estimation (association), which models the underlying distribution of the data.

- Reinforcement Learning (RL): involves an agent learning to make decisions by interacting with an environment. Instead of using labeled data, the agent receives feedback in the form of rewards or penalties based on its actions. Over time, it learns to choose actions that maximize cumulative rewards.

Additionally, semi-supervised learning combines aspects of both supervised and unsupervised methods by leveraging both labeled and unlabeled data [115]. Most machine learning techniques used in FER are based on supervised learning. Among these, the most widely adopted approaches include:

Support Vector Machines (SVM): SVM is a widely used algorithm for classification tasks, including FER. The core idea of SVM is to find an optimal hyperplane that separates different classes in a feature space with the maximum possible margin. This hyperplane is determined by the closest data points from each class, known as support vectors, which are critical to defining the decision boundary. While originally designed for linearly separable data, SVM can handle non-linear problems by using kernel functions to transform the input data into a higher-dimensional space where linear separation becomes possible. SVMs are well-suited for high-dimensional data and have been applied successfully in various domains, including image classification, speech recognition, text sentiment analysis, and medical diagnosis. Although SVMs can be less effective on large or noisy datasets and do not inherently yield probability estimates, they are still highly valued for their accuracy and reliability. Their ability to generalize well from limited data and handle high-dimensional input makes them a common choice in traditional FER systems [87, 116].

K-Nearest Neighbors (KNN): KNN is a non-parametric, instance-based learning algorithm commonly used for classification tasks, including facial expression recognition. It classifies an input sample based on the majority class of its k closest neighbors in the feature space, typically using distance metrics such as euclidean distance to measure similarity. KNN assumes that similar samples exist in close proximity and relies entirely on the training data during the prediction phase, classifying it as a lazy learner. Unlike parametric models, it does not make assumptions about the underlying data distribution, which contributes to its flexibility across different data types and modalities. Despite its simplicity, KNN can achieve competitive performance when the data is well-structured and

the feature space is meaningful. However, its prediction time increases with larger datasets, as it must compare each input with all training samples [116, 117].

Random Forest (RF): RF is an ensemble learning method that builds a collection of decision trees to improve classification or regression performance. Each tree in the forest is trained on a different random subset of the data using a technique called bootstrapping, where samples are selected with replacement. Additionally, at each node of a tree, a random subset of features is chosen to determine the best split, which helps reduce correlation among trees and improves model robustness. The final prediction is typically made by aggregating the outputs of all trees, often through majority voting in classification tasks. This approach helps to reduce overfitting, improve generalization, and increase accuracy compared to using a single decision tree. Random Forest has been applied effectively to a range of classification and regression tasks, demonstrating strong performance even when the data is incomplete or complex [116, 118].

2.4.2 Deep Learning Models

Deep learning is a subfield of machine learning that employs artificial neural networks composed of multiple computational layers to model complex patterns in data. Inspired by the architecture of the human brain, these networks consist of hierarchically organized layers of interconnected nodes, or neurons, that progressively extract abstract representations from the input. The term "deep" refers to the presence of multiple layers within the network, enabling the learning of high-level features and hierarchical structures in the data [119]. This depth facilitates the automatic integration of both feature extraction and classification within a single framework, eliminating the need for handcrafted features. Deep learning has demonstrated remarkable performance across a wide range of tasks [119, 120]. In FER, DL-based methods offer a unified approach by supporting end-to-end learning directly from visual input. The following subsections outline the most prominent DL-based techniques applied in FER, including convolutional, recurrent, transformer-based, and vision-language models.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks are one of the most widely used deep learning architectures for analyzing image data. In contrast to traditional techniques that depend on manually

engineered features, CNNs automatically learn hierarchical representations from raw input. This is achieved through a series of specialized layers designed to capture spatial and structural patterns, making CNNs highly effective for tasks such as image classification and object detection, and semantic segmentation [119]. Figure 2.9 illustrates a common CNN architecture, composed of multiple convolutional and pooling layers followed by fully connected layers.

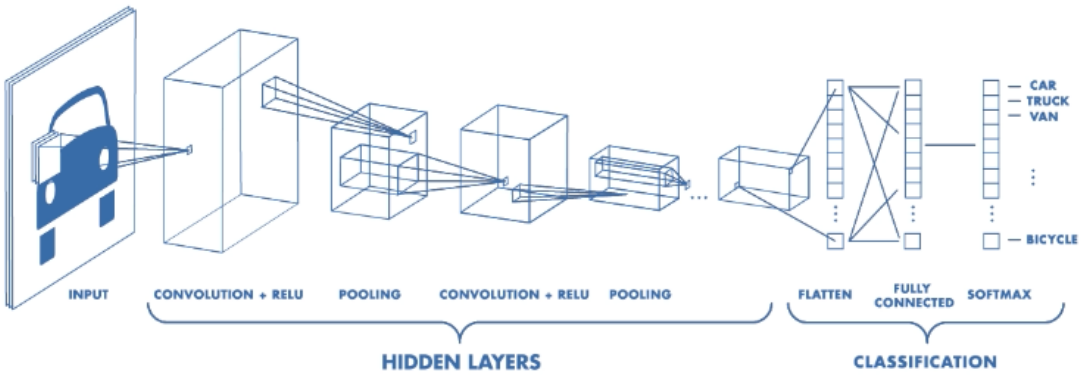


Figure 2.9: A representative CNN architecture illustrating a layered structure comprising convolutional, activation, and pooling operations, followed by fully connected layers used for classification [121].

A typical CNN includes the following core components:

- Convolutional Layer (Conv): is responsible for extracting features from the input image by applying a filter or kernel through a convolution operation, allowing it to capture specific features such as edges, textures, and shapes.
- Pooling Layer: is responsible for reducing the spatial dimensions or size of the feature maps generated by the convolutional layer by applying a down-sampling operation, which helps decrease computational complexity.
- Activation Function: introduces non-linearity into the model by applying a non-linear function, such as ReLU, to the output of the pooling layer, enabling the network to learn more complex patterns in the input data.

- Fully Connected Layer (FC): FC links every neuron from the previous layer to each neuron in the next, and is used to combine the extracted features in order to generate the final prediction.

Over the years, several CNN variants have been developed to improve the performance and generalization of the model. Well-known architectures include AlexNet [122], VGGNet [123], ResNet [124], Xception [125], MobileNet [126], and EfficientNet [127], each offering specific improvements suited to different visual recognition tasks. An overview of these architectures is presented in Section 4.4.1.

Recurrent Neural Networks (RNNs)

RNN is another class of neural networks specifically designed for processing sequential or time-series data. They differ from feedforward networks by incorporating feedback connections, which allow information from previous time steps to influence the current output. This internal memory enables RNNs to capture temporal dependencies, making them suitable for tasks where the order of data is important, such as speech recognition and natural language processing. However, standard RNNs often face challenges in learning long-range dependencies due to the vanishing gradient problem, where gradients become exceedingly small during backpropagation through time, making it difficult for the network to update weights effectively in earlier layers. To overcome this limitation, more advanced variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been introduced [119]. Figure 2.10 illustrates the structural differences among a standard RNN, LSTM, and GRU unit.

Long Short-Term Memory (LSTM): LSTM is a widely used variant of RNN, introduced by Hochreiter and Schmidhuber [128], to address the vanishing gradient problem commonly encountered in standard recurrent networks. LSTM units contain a memory cell capable of preserving information over long sequences, making them effective for learning long-term dependencies. The flow of information into and out of the cell is regulated by three gates: the "forget gate", which decides which parts of the previous state should be discarded, the "input gate", which determines what new information should be added, and the "output gate", which controls the output based on the current cell state. LSTM is considered one of the most effective forms of RNN [119].

Gated Recurrent Unit (GRU): GRU is another variant of RNN introduced by Cho et al. [129], designed to handle sequence data using gating mechanisms. It simplifies the LSTM architecture by using only two gates: the "reset gate", which controls how much past information to forget, and the "update gate", which decides how much of the past state should be retained. GRUs do not include an output gate, resulting in fewer parameters and reduced computational cost. This compact structure allows GRUs to learn long-term dependencies efficiently while maintaining performance comparable to LSTMs. GRUs are particularly effective in scenarios where model simplicity and training speed are important, and have shown good results on smaller or less frequent datasets [119].

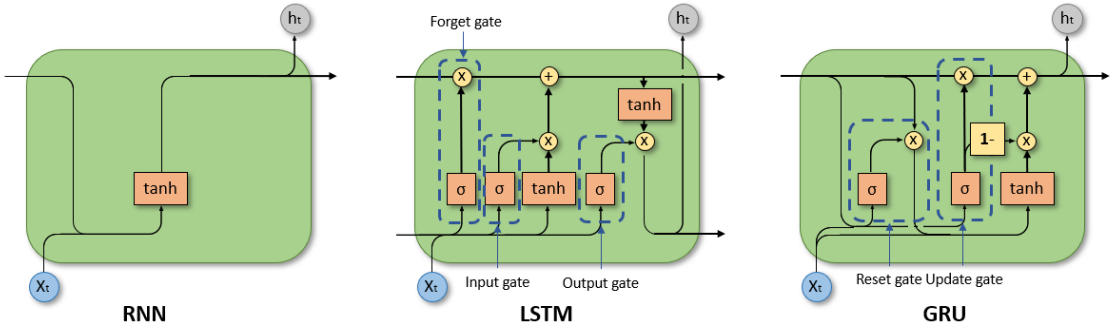


Figure 2.10: Illustration of RNN, LSTM, and GRU cells architecture [130]. RNN uses a simple recurrent unit with a tanh activation. LSTM introduces input, forget, and output gates, allowing better handling of long-term dependencies. GRU simplifies LSTM by combining input and forget gates into an update gate and using a reset gate. σ denotes the sigmoid function, tanh is the hyperbolic tangent, $\times$ and $+$ represents element-wise multiplication and addition, and $1-$ is subtraction from 1.

Vision Transformer (ViT) Network

The Vision Transformer [131] is a deep learning model that applies the transformer architecture [132], originally developed for Natural Language Processing (NLP), to visual tasks. Unlike CNNs, which extract features using local filters, ViTs rely on self-attention to capture global dependencies among different regions of an image. This design has achieved competitive performance in a range of computer vision tasks, including image classification, object detection, and segmentation. ViT is built upon the transformer model, which is particularly effective at handling sequential inputs. In ViT, instead of processing sequences of words, the input is a sequence of image patches, enabling the model

to interpret visual data using the same principles. Figure 2.11 illustrates the framework of the Vision Transformer.

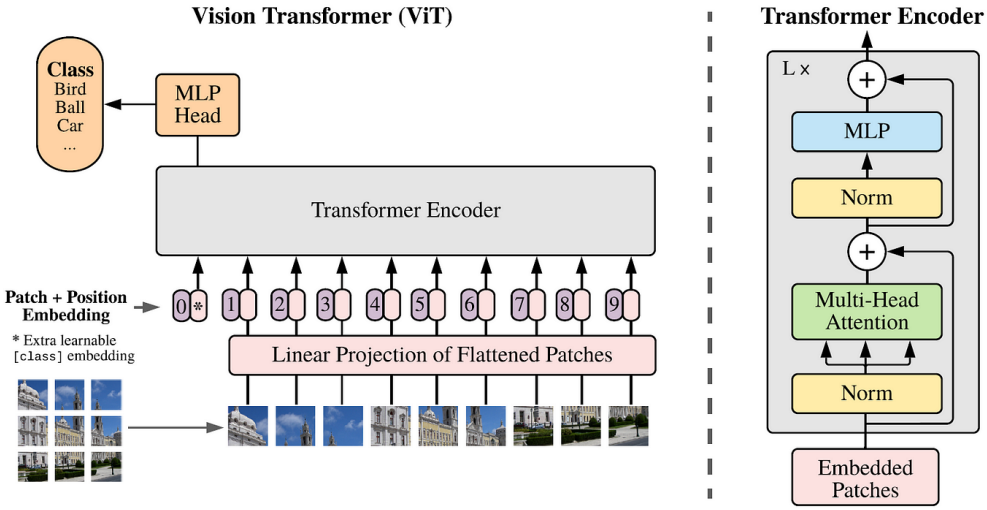


Figure 2.11: The architecture of vision transformer (ViT) [131]. An input image is split into fixed-size patches, linearly embedded, and combined with positional embeddings. A learnable class token is prepended to the sequence. The resulting embeddings are passed through L Transformer encoder layers, each composed of multi-head self-attention, layer normalization, and an Multilayer Perceptrons (MLP). The class token output is used for final classification.

The architecture of ViT comprises several key stages:

- Image Patching and Embedding: In the initial stage of the vision transformer, the input image of size $H \times W \times C$ is divided into non-overlapping patches of size $P \times P$. For example, a 224×224 image with 16×16 patches results in 196 patches. Each patch is flattened into a vector of size $P^2 \times C$, then linearly projected into a fixed-dimensional embedding space through a learnable transformation. This produces a sequence of $N = \frac{H}{P} \times \frac{W}{P}$ patch embeddings, which serve as the input tokens to the transformer.

- Positional Encoding: Since transformers do not inherently capture spatial order, positional encodings are added to the patch embeddings to retain the original structure of the image. These encodings provide information about the position of each patch, enabling the model to learn spatial relationships. In ViT, the positional embeddings

can be learned during training or predefined, with most ViT implementations using learnable positional encodings.

- Transformer Encoder Layers: The patch embeddings, enriched with positional information, are passed through a stack of transformer encoder layers. Each layer comprises two main components: Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN).

  1) Multi-Head Self-Attention (MHSA), this mechanism enables each patch to attend to all others in the sequence by computing weighted relationships. The attention scores are derived from learned query (Q), key (K), and value (V) projections using the formula:

  $$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.7}$$

  where d_k is the dimensionality of the keys.

  2) Feed-Forward Network (FFN), following attention, each patch embedding is passed through a position-wise FFN composed of two fully connected layers with a non-linear activation function (commonly GELU) applied between the two layers.

- MLP Head (Classification Head): The classification head in ViT plays a key role in producing predictions for image recognition tasks. A special classification token, known as the [CLS] token, is added at the beginning of the input sequence to gather information from all patches through the self-attention mechanism. After passing through the transformer layers, the final representation of the CLS token is used for classification. It is processed by an MLP, typically composed of one or two fully connected layers, followed by a softmax function to produce the final prediction.

ViT also has several variants developed to improve its accuracy and generalization. These include models such as DeiT [133], Swin Transformer [134], MobileViT [135], and CvT [136], each introducing different modifications to the original ViT architecture.

Vision Language Models (VLMs)

Vision-Language Models are deep learning architectures designed to learn joint representations from both visual and textual data. Unlike traditional unimodal models that process either images or text independently, VLMs integrate both modalities to enable multimodal

understanding and reasoning. This capability opens up a wide range of applications, such as Image Captioning, where the system generates a textual description of an image [137], Visual Question Answering (VQA), which involves answering natural language questions based on image content [138], and text-based image retrieval, which searches for the most relevant image given a textual query [139].

At their core, VLMs typically consist of three main components:

- Visual Encoder: Converts visual inputs (images or video frames) into meaningful representations (visual embeddings). Common backbones include CNNs and ViTs, which extract spatial and contextual features from visual data.
- Language Encoder: Maps input text (e.g., captions or prompts) into semantic vector representations (textual embeddings). Transformer-based models like BERT [140] and GPT [141] are commonly used for their effectiveness in understanding language structure and context.
- Cross-Modal Fusion Mechanism: To enable joint reasoning, visual and textual features are aligned or fused through mechanisms such as embedding projection and similarity learning (as in CLIP [141]) or through deep cross-modal attention (e.g., in models like UNITER [142] and BLIP [143]). This fusion allows the model to capture interactions between modalities and understand how language relates to visual content.

The success of VLMs depends largely on the training objectives used to align and fuse the modalities. These objectives are typically categorized as follows:

- Contrastive Learning: This strategy learns a shared embedding space by maximizing similarity between matching image-text pairs while minimizing it for mismatched pairs. It enables strong zero-shot transfer capabilities. Notable models include CLIP [141], ALIGN [144], and SigLIP [145], which employ dual encoders to calculate image-to-text and text-to-image similarities.
- Generative Learning: Generative training enables the model to generate new data in one modality conditioned on the other. For example, text-to-image generation produces images from textual descriptions, while image-to-text generation generates captions or descriptions from input images. This training strategy encourages deeper semantic understanding across modalities. Generative learning also involves masked prediction strategies such as Masked Image Modeling (MIM) and Masked Language

Modeling (MLM), where the model learns to infer missing parts of one modality based on the other. Notable models that use these techniques include BLIP [143] for text generation and FLAVA [146] for joint multimodal masking.

- Alignment Objectives: This training strategy enforces fine-grained alignment between visual and textual inputs by explicitly learning whether a given image-text pair matches. This includes Image-Text Matching (ITM), predicting whether an image and sentence pair correspond (e.g., VisualBERT [147], ALBEF [148]). And Region-word alignment, aligning specific image regions with textual phrases (e.g., GLIP [149], ViLBERT [150]). These objectives help the model build a more localized understanding of cross-modal content.

Through this unified architecture and diverse training strategies, VLMs can reason across modalities and transfer knowledge to a wide range of downstream tasks. These capabilities have made VLMs foundational to modern multimodal AI [151].

2.4.3 Transfer Learning Strategies

Transfer learning is a machine learning technique in which a model developed for one task is reused as the starting point for a different but related task. Rather than training a new model from scratch, transfer learning leverages the knowledge acquired from a large-scale source task and applies it to a target task with limited labeled data. This approach significantly reduces computational cost and training time while improving performance, particularly when data is scarce. Its effectiveness depends on the extent to which knowledge learned from the source task can be meaningfully transferred and adapted to the target domain [152, 153].

A common approach to applying transfer learning in deep learning is to reuse a pretrained model and adapt it to a new task. This can be done through two main strategies: feature extraction and fine-tuning (see Figure 2.12).

(1) In feature extraction, the pretrained model is used as a fixed feature extractor. All layers of the pretrained network are kept frozen, and only a new classification head is trained on top of the extracted features. This method is efficient when labeled data is scarce or computational resources are limited, as it avoids updating the entire model.

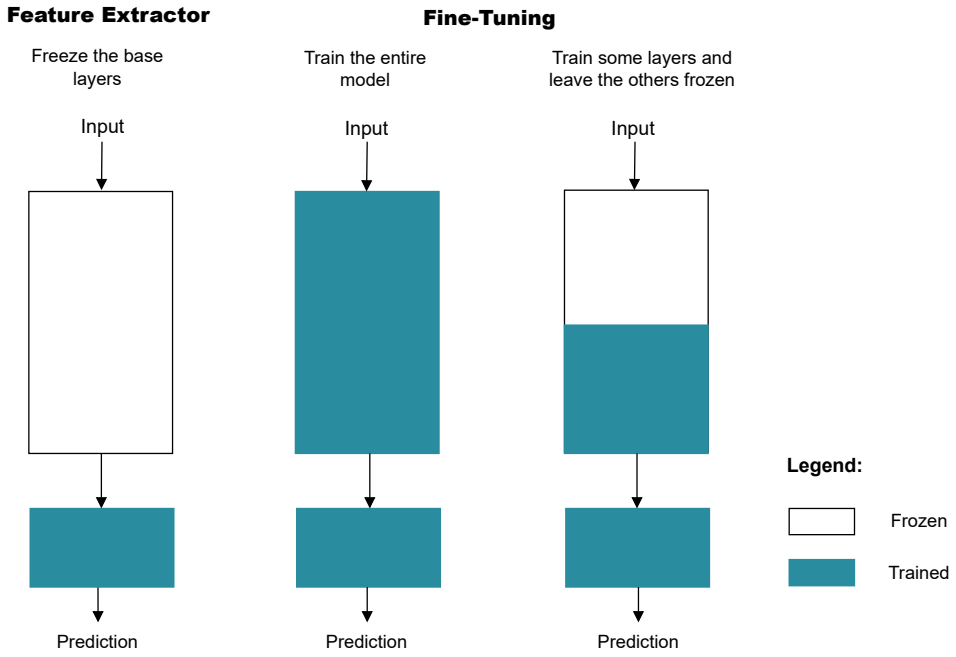


Figure 2.12: Deep learning-based transfer learning strategies. Adapted from [154].

In contrast, (2) fine-tuning involves updating some or all of the parameters of the pretrained model to better adapt to the target task. Fine-tuning strategies vary based on the task complexity, dataset size, and resource availability:

- Full fine-tuning, where all layers of the pretrained model are updated on the new task. This is suitable when a large labeled dataset and computational resources are available, allowing the model to learn both general and task-specific features effectively.
- Partial fine-tuning, where only a subset of the top layers is updated while the lower layers are kept frozen. This approach is often adopted when the dataset is moderately sized, allowing the model to retain general visual features while adapting high-level representations.

While fine-tuning is a powerful approach to adapt pretrained models to new tasks, it often involves updating a large number of parameters, which becomes computationally expensive, especially for large-scale models. To address this, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a more scalable alternative. Originally introduced in the field of natural

language processing to reduce the burden of full-model fine-tuning on large language models [155, 156], PEFT techniques such as Adapters, Prompt tuning have since been successfully adapted for computer vision and multimodal tasks [157, 158]. Rather than updating the entire model, these techniques freeze most pretrained weights and train only a small, task-specific subset of parameters. This significantly lowers memory usage and training time, while maintaining strong performance across diverse applications. As such, PEFT is gaining attention as an effective and practical solution for adapting large models in resource-constrained environments.

2.5 Review of Existing FER Methods

Facial expression recognition has attracted increasing research interest due to its broad range of applications and its significance in human-computer interaction. A growing body of literature has focused on developing effective FER systems using both traditional methods and advanced deep learning techniques, as previously discussed. In this section, we present a comprehensive review of related work, covering state-of-the-art approaches from early conventional methods to recent deep learning-based models. The review is organized to highlight advancements in general FER systems, as well as those specifically designed for driver FER. Additionally, we discuss the key challenges and limitations identified in the existing literature, which motivate ongoing research in this domain.

2.5.1 Approaches in General-Purpose FER

In this section, we review the papers in general-purpose FER, covering foundational handcrafted approaches, deep learning-based models, and recent trends that have shaped the field. A summary of the reviewed papers is presented in Tables 2.3 and 2.4.

A significant portion of existing research on expression recognition has relied on static images, largely because of the ease of data processing and the availability of annotated training and testing datasets [9]. Initial efforts were grounded in handcrafted feature extraction [159, 160, 161]. For instance, Yang et al. [162] proposed an image-based FER method using LBP for feature extraction and an SVM decision tree classifier. In this approach, facial images are first converted into LBP histograms to capture local texture features, then used to train a multi-class SVM arranged in a hierarchical decision tree to

enhance classification efficiency. Evaluated on the JAFFE dataset, the approach achieved an average accuracy of 92.45%, outperforming standard LBP and flat SVM models by leveraging a hierarchical classification structure. Dagher et al. [163] proposed a facial expression recognition approach based on a three-stage SVM framework to classify seven basic emotions. The system utilizes HOG for feature extraction and introduces a hierarchical SVM design to improve discrimination between visually similar expressions. The model was evaluated on three widely used image-based FER datasets: JAFFE, CK+, and RaFD, Using both Leave-One-Out and K-Fold cross-validation, the proposed method achieved improved classification accuracy compared to traditional single-stage SVM models, demonstrating robust performance across all datasets. Kulkarni and Kolhe [164] proposed and compared two facial expression recognition methods aimed at classifying happiness, surprise, and neutrality. The first method uses traditional image processing techniques, while the second employs Dlib for facial landmark detection. Both approaches were evaluated using SVM and MLP classifiers across three datasets: the 10k US Adult Faces, MUG Facial Expression, and a personal dataset. The Dlib-based method outperformed the image-processing-based approach, achieving an accuracy of over 96% compared to 91.5%.

With the emergence of deep learning, several methods have been proposed [10, 165, 166]. For CNN-based methods, Sun et al. [16] proposed an ROI-Attention Vectorized CNN model for static facial expression recognition, which incorporates attention mechanisms focused on predefined facial regions to enhance discriminative feature extraction. Inspired by Capsule Networks, the model generates vectorized features from key facial areas and reconstructs them through a decoder for robust expression classification. Evaluated on AffectNet, CK+, FER2013, and JAFFE datasets, the method achieved average accuracies of 70.0%, 87.2%, 72.5%, and 92.0%, respectively, which are higher compared to other existing methods. Özcan and Basturk [167], proposed a FER method that combines transfer learning with hyperparameter optimization to enhance classification performance. Their approach involves data preparation techniques and the utilization of a CNN architecture fine-tuned through transfer learning. The authors introduced a new dataset named ERUFER, comprising static images, and evaluated their method on both ERUFER and the widely used JAFFE dataset. The proposed model achieved a recognition accuracy of 92.56% on the ERUFER dataset and 99.53% on the JAFFE dataset, outperforming existing methods using 10-fold cross-validation.

On the other hand, some studies have proposed hybrid models that combine CNNs with recurrent layers to improve expression recognition. Jain et al. [17] proposed a hybrid

deep learning model that integrates CNNs with RNNs to enhance static FER. The CNN component is used for spatial feature extraction from static images, while the RNN captures dependencies across feature sequences to improve classification robustness. The model was trained and evaluated on the JAFFE and MMI datasets, achieving accuracies of 94.91% and 92.07%, respectively, highlighting the effectiveness of hybrid models, particularly in terms of reducing false detections and handling variations in expression.

Recently, transformer-based architectures have gained attention in FER for their ability to model global relationships and capture fine-grained attention across facial regions without relying on sequential processing [168]. Ma et al. [18] proposed a Visual Transformer framework with Attentional Selective Fusion (VTFF) for FER in unconstrained environments. The model combines handcrafted LBP and CNN features, which are fused using attention mechanisms before being passed through a transformer encoder to capture global context. Evaluated on RAF-DB, FERPlus, and AffectNet, VTFF achieved strong results with accuracies of 88.14%, 88.81%, and 61.85%, respectively, demonstrating its effectiveness in handling complex real-world expressions. Similarly, in an unconstrained environment, Li et al. [169] proposed FG-AGR, a fine-grained graph-based framework for facial expression recognition in the wild. The model combines adaptive region selection, visual transformers for local feature extraction, and graph reasoning to model inter-regional dependencies. Evaluated on RAF-DB, AffectNet-7, AffectNet-8, and FERPlus, FG-AGR achieved accuracies of 90.81%, 64.91%, 60.69%, and 91.09%, respectively, outperforming existing methods in challenging, real-world settings.

As research progressed, attention gradually shifted from static images to video-based approaches to capture the dynamics of facial expressions over time, an essential component often overlooked in static methods [9, 23]. This shift was further supported by the availability of large-scale, in-the-wild video datasets such as DFEW [102], MAFW [105], and FERV39k [101], which better represent real-world complexity, including variations in pose, lighting conditions, and occlusion. First, similarly to image-based FER, early methods for dynamic facial expression recognition relied on handcrafted spatiotemporal descriptors such as Local Binary Patterns on Three Orthogonal Planes (LBP-TOP), HOG, and optical flow [13, 175, 176]. These techniques were successfully applied to controlled datasets like CK+, MMI, and RAVDESS [177, 178]. Then, with the advent of deep learning and the availability of large-scale in-the-wild datasets [105], [102], [101], data-driven techniques have become the foundation for advancing DFER research. Recent deep learning-based DFER methods can be broadly categorized into three main folds. The first fold, 3D

Table 2.3: Summary of general FER-related works (Part 1).

Ref	Method	Datasets	Face detector	Data Split	Accuracy	Limits
[162]	SVM with HOG	JAFFE CK+ RaFD	N/A	LOOCV, K-Fold	92.6% 76.7% 97.0%	Relies on hand-crafted HOG features, which limits its ability to capture subtle or non-linear facial expressions under varied real-world conditions.
[163]	LBP + SVM Decision Tree	JAFFE	Viola–Jones	Leave-one-out K-fold (10 and 5)	92.45%	The lack of adaptability of shallow models in more challenging scenarios, especially when tested only on JAFFE, a constrained dataset.
[164]	SVM	-10k US Adult Faces MUG personal dataset	Dlib (in 2nd method)	(90,10)	91.5% (1st) 96% (2nd)	Emotion class limitation (only 3 emotions). Reproducibility issue due to use of personal datasets.
[16]	CNN with decoder	AffectNet CK+ Fer2013 JAFFE datasets	ERT and Dlib algorithm	N/A	70.0% 87.2% 72.5% 92.0%	Fixed ROIs limit adaptability to subtle or variant expressions. Not validated in unconstrained/real-world settings.
[167]	CNN	ERUFER JAFFE	Dlib_based	10-fold(JAFFE)	92.56% 99.53%	CNN trained on limited data may overfit and lack robustness.
[17]	Hybrid CNN-RNN	JAFFE MMI	N/A	(70,30)JAFFE + MMI	94.91% 92.07%	Temporal modeling applied on shallow datasets limits effectiveness. Lacks real-world evaluation.
[18]	VTFF	RAF-DB FERPlus AffectNet	MTCNN	official splits	88.14% 88.81% 61.85%	Handcrafted features limit representation of complex expressions. Transformer-based increases computational overhead.
[169]	FG-AGR	RAF-DB AffectNet-7 AffectNet-8 FERPlus	N/A	official splits	90.81% 64.91% 60.69% 91.09%	Multi-stage architecture introduces high computational complexity and memory load. Real-time efficiency remains unverified.

Table 2.4: Summary of general FER-related works (Part 2).

Ref	Method	Datasets	Face detector	Data Split	Accuracy	Limits
[170]	CNN-RNN	AFEW 6.0	Viola-Jones	official splits	59.02%	Relatively low accuracy. Limited modeling of long-term temporal dependencies.
[171]	M3DFEL	DFEW FERV39k	N/A	5 folds(DFEW) official split(FERV39K)	UAR: 56.10%, 35.94% WAR: 69.25%, 47.67%	High computational cost due to multi-scale and 3D feature modeling. Relies on a MIL strategy that requires dataset-specific fine-tuning.
[172]	MRAN	AFEW	DLib	N/A	49.01%	Relies on multimodal inputs (e.g., RGB, depth, thermal), limiting deployment flexibility. Limited generalization analysis.
[173]	Hybrid CNN-RNN	Enotional Wearable Dataset 2020	N/A	(80,10,10)	62.48%	Trained on a small, non-public dataset, limiting reproducibility. Emotion scope is narrow, affecting generalizability.
[174]	T-MEP	DFEW MAFW AFEW	RetinaFace	5 folds (MAFW, DFEW)	UAR: 57.16%, 39.37%, 50.22% WAR: 68.85%, 52.85%, 52.96%	Requires audio-visual inputs. Limited performance and high computational cost.
[86]	Former-DFER	DFEW AFEW	RetinaFace	10-fold(JAFFE)	UAR: 53.69%, 47.42% WAR: 65.70%, 50.92%	Transformer-based design increases model complexity. Limited training data constrains robustness.
[20]	IAL	DFEW FERV39K	N/A	5 folds (DFEW) official split (FERV39K)	UAR: 55.71%, 35.82% WAR: 69.24%, 48.54%	Primarily addresses intensity variation, expression discriminability remains limited. Limited performance.

CNN-based Methods, involves leveraging 3D CNNs to jointly capture spatial and temporal information [179]. Early work, such as Fan et al. [170] employed 3D-CNNs to model video appearances and motions. Particularly, the C3D network captures spatiotemporal features directly by applying 3D convolutions over short video segments. Evaluated on the AFEW 6.0 dataset, the approach achieved an accuracy of 59.02%, highlighting the benefit of 3D CNNs in modeling temporal expression dynamics. Similarly, Wang et al. [171] proposed M3DFEL, a framework for dynamic FER that addresses label ambiguity and temporal imbalance using a Multi-Instance Learning strategy. The model integrates 3D CNNs for short-term motion modeling and a dynamic aggregation module for long-term dependencies. Evaluated on DFEW and FERV39K, it outperformed previous methods, demonstrating strong performance under real-world conditions.

The second fold is CNN-RNN-based Methods, which typically utilize CNNs to extract spatial facial features from each frame, followed by RNNs to model temporal dependencies across sequences. Commonly used CNN backbones include VGG [123] and ResNet [124], while RNNs (LSTM [128] and GRU [180]) are widely used for temporal modeling. For instance, Lee and Kim [172] introduced a multi-modal recurrent attention network which integrates color, depth, and thermal modalities to capture comprehensive facial cues. The model employs recurrent attention mechanisms to focus on salient spatiotemporal features across modalities, enhancing the recognition of subtle expressions. Recently, Manalu and Rifai [173] developed a hybrid CNN-RNN architecture using Time Distributed layers for sequential video processing. Three models were evaluated on a custom dataset: a custom CNN-RNN, MobileNetV2-RNN, and InceptionV3-RNN. The InceptionV3-RNN model achieved the highest accuracy of 66%, demonstrating the effectiveness of the proposed method.

The third and emerging fold, transformer-based Methods, has gained attention due to the transformative success of transformer models [131]. Transformers have been utilized as spatial, temporal, or spatiotemporal feature extractors for Dynamic FER. For example, Zhao and Liu [86] introduced Former-DFER, which comprises convolutional spatial transformers (CS-Former) to extract robust spatial features and a temporal transformer (T-Former) to model temporal dependencies across frames. Evaluated on the DFEW and AFEW datasets, Former-DFER achieved state-of-the-art performance. Meanwhile, Wang et al. [174] proposed a Transformer-Based Multimodal Emotional Perception (T-MEP) framework that fuses visual and audio cues to enhance emotion perception. Designed for real-world scenarios, the model addresses challenges like occlusion and pose variation. Evaluated on DFEW,

MAFW, and AFEW, it achieved state-of-the-art results, demonstrating the effectiveness of multimodal fusion in DFER. Additionally, Li et al. [20] proposed an Intensity-Aware Loss (IAL) to improve dynamic FER by giving more attention to low-intensity expressions, which are often misclassified. Combined with a Global Convolution-Attention module, the method achieved state-of-the-art results on DFEW and FERV39k.

More recently, with the emergence of Vision-Language Models, such as CLIP model [141] (more details in section 5.3), their adaptability across a range of tasks [181, 182, 183] has led researchers to explore their utility in FER task to develop more robust FER systems by aligning visual features with semantic information [23, 184, 185, 186, 187, 188].

2.5.2 FER in Driving environments

Regarding FER methods in driving environments, as discussed in Section 2.2.2, recognizing drivers' facial expressions plays a crucial role in enhancing road safety. Consequently, several approaches have been proposed to accurately and efficiently detect facial expressions under challenging conditions such as variable illumination, occlusions (e.g., sunglasses or hands), and head pose variations. Tables 2.5 and 2.6 provide a summary of these FER methods.

An early conventional-based approach is presented in [103], where the authors introduced a facial expression recognition method tailored for real-time embedded systems in intelligent vehicles. This method utilizes geometric features and a hierarchical Weighted Random Forest (WRF) classifier. Facial landmarks are first detected using DLib, from which geometric features are extracted and then classified by the hierarchical WRF. The system was evaluated on the CK+, MMI, and KMU-FED datasets, achieving accuracies of 92.6%, 76.7%, and 97.0%, respectively. Notably, the approach does not rely on GPU acceleration yet achieves performance comparable to or better than state-of-the-art methods, while maintaining low memory and computational requirements, highlighting its suitability for resource-constrained environments. Building on this work, Jeong et al. [65] proposed a lightweight multilayer random forest (LMRF) model for fast facial expression recognition. Similar to the previous method, facial landmarks are detected using DLib, and geometric features based on distance ratios and angles are extracted. These features are then classified using a two-layer LMRF with a few trees. Despite its compact structure, the model achieved accuracies of 93.4%, 77.4%, and 95.1% on the CK+, MMI, and KMU-FED

datasets, respectively. It also outperformed state-of-the-art model compression techniques, offering a balance between computational efficiency and accuracy. Azman et al. [189] proposed a real-time driver facial expression recognition system focused on detecting anger to help mitigate road rage and reduce accident risk. The system uses the Viola-Jones Haar feature algorithm for frontal face detection and an SVM for classification. Trained on the JAFEE dataset, the model achieved an average accuracy of 97%. On the other hand, Patil et al. [26] introduced a method that combines geometric and texture-based features through the fusion of local LBP and facial landmarks. Facial landmarks are extracted using OpenCV DLib, while Haar features are used for face detection. Emotion classification is performed using an SVM classifier. This method achieved an accuracy of 86.7% on the CK+ dataset.

With the rise of deep learning, particularly the use of CNNs for either feature extraction or classification, several studies have demonstrated considerable potential. For instance, to balance performance and efficiency, recent works have explored lightweight deep architectures, such as the approach by Verma et al. [111], in which they proposed a real-time FER framework for drivers that contains face detection using the Viola & Jones algorithm with feature extraction via pre-trained AlexNet and VGG16 models. The extracted features are fused and classified using Grassmann graph embedding-based discriminant analysis while considering variations in lighting and driver movement. Evaluated on several datasets (VIVA FACE, DriveFace, JAFFE, MMI, CK+), the system achieved high accuracies of 96.95%, 94.6%, 98.8%, 98.8%, and 97.4% on each dataset, respectively. Outperforming several existing methods. In another study [190], a lightweight SIC-NET+RC-NET dual-stream CNN model was proposed to handle varying vehicle speeds during driving scenarios. The SIC-NET reduces the model parameters, while the RC-NET improves the residual convolutional network channel. After preprocessing the input images, the system fuses the features extracted by the SIC-NET and RC-NET for recognition, with SIC-NET used when the vehicle speed is over 60km/h and SIC-NET+RC-NRT used when it's 60km/h or less. The FER2013 dataset was used to evaluate the proposed system, and the achieved accuracy was 69.65%, indicating robustness and effective recognition performance, with the potential to enhance driver safety and reduce property damage.

Zaman et al. [191] introduced a method for recognizing driver emotions without requiring the driver to perform any additional efforts. Their approach combined an improved Faster R-CNN face detector, where the base layers were replaced by a custom 11-layer CNN block for better feature learning, with a NasNet Large CNN model for facial expression recognition

via transfer learning. The system was evaluated on five benchmark datasets: JAFEE, CK+, FER-2013, AffectNet, and a custom dataset, achieving accuracies of 98.48%, 99.73%, 99.95%, 95.28%, and 99.15%, respectively. These results demonstrate high precision and robustness in diverse facial recognition tasks relevant to driver monitoring. Similarly, Sahoo et al. [192] developed a driver monitoring system using dashboard camera images analyzed through deep CNNs. They designed two custom models, a 6-layer and a 10-layer CNN, and compared them to a pre-trained VGG16 model using the FER2013 dataset, which includes challenging real-world conditions such as occlusions and poor image quality. After applying preprocessing techniques like image scaling and data augmentation, the 10-layer CNN achieved the best performance with 68.34% accuracy, followed by the 6-layer CNN at 66.67% and VGG16 at 63.68%. These results suggest that custom-designed CNNs can outperform pre-trained models when adapted for in-vehicle emotion recognition. Expanding upon the driver's emotional state, Tauqeer et al. [193] proposed a unified framework for detecting both emotional and behavioral distractions during driving. They trained a CNN and a VGG16 model separately on the State Farm Distracted Driver dataset (behavioral states) and the FER2013 dataset (facial expressions). The CNN achieved 99% accuracy on the distracted driver dataset and 50% on FER2013, while VGG16 reached 86% and 38% on the respective datasets. Although VGG16 required less training time, the CNN outperformed it in classification accuracy across both domains, emphasizing the effectiveness of CNNs in real-time distraction detection systems.

Building on the challenges of facial expression recognition under occlusion, Sudha et al. [194] proposed the Driver Facial Expression Emotion Recognition (DFEER) method, specifically designed to handle semi-facial occlusions. The approach involves computing optical flow between occluded image frames using the Farneback method, followed by reconstruction of the flow using a Parallel Multi-Verse Optimizer (PMVO) trained on optical flow data. This reconstructed flow aims to restore blocked facial regions and is then passed to a Very Deep Convolutional Network (VGGNet) for expression classification. The method was evaluated using the CK+ and KMU-FED datasets and compared with other recognition models based on metrics such as precision, recall, F1-score, and accuracy. DFEER achieved high performance, with accuracy rates of 95.92% on CK+ and 94.78% on KMU-FED. In a related effort to enhance road safety through real-time driver state monitoring, Shang et al. [195] introduced a non-invasive and efficient system for detecting driver fatigue and emotional state. Unlike approaches that rely solely on single-frame facial features, this method processes video sequences using DLib to identify facial landmarks and regions from

CSI camera input. A dual-threshold fatigue detection model was proposed, incorporating metrics such as head posture, eye closure frequency and duration, and yawning frequency. For emotion recognition, the authors designed a lightweight RMXception convolutional neural network. These fatigue and emotion indicators were then fused using time-series data to form a comprehensive evaluation of the driver's condition. The RMXception model was trained on the FER2013 dataset and achieved an emotion recognition accuracy of 73.32%.

To further improve FER, Jain et al. [196] proposed a novel system called Squirrel Search Optimization with Deep Learning Enabled Facial Emotion Recognition (SSO-DLFER). This system comprises two main stages: face detection and emotion classification. The RetinaNet model was employed for accurate face detection, while emotion recognition utilized a NASNet-LarGRU classifier. Additionally, system performance was enhanced through hyperparameter tuning using the Squirrel Search Optimization (SSO) algorithm, which is inspired by the foraging behavior of flying squirrels. The proposed SSO-DLFER framework was evaluated on the KDEF and KMU-FED datasets, achieving high accuracy rates of 99.69% and 99.50%, respectively. While the system demonstrated state-of-the-art results, the authors suggested that recognition performance could be further improved by adopting ensemble learning strategies.

Varma et al. [110] introduced a hybrid deep learning model designed to capture spatio-temporal patterns in video sequences. Their method employed a Convolutional Bidirectional Long Short-Term Memory (CBiLSTM) network to learn spatial and temporal dependencies simultaneously for emotion classification. This model was evaluated on the KMU-FED and DMD datasets, achieving an F1-score of 0.958 on KMU-FED and an average F1-score of 0.745 on the DMD dataset.

Additionally, Sukhavasi et al. [197] proposed a hybrid approach to improve emotion recognition accuracy under occlusions, illumination changes, and pose variations. Their method consists of three stages: preprocessing with AdaBoost and Haar features for face detection, Gabor and LBP-based feature extraction, and final classification using a CNN combined with an SVM. The model achieved notable accuracy rates of 84.41% on FER-2013, 95.05% on CK+, 98.57% on KDEF, and 98.64% on KMU-FED. In a related work, the same authors [198] introduced two DNN-based approaches to address similar challenges. The first employs Viola-Jones and the KLT algorithm used to track and extract feature points from the detected face, followed by a modified deep network with batch normalization and

Table 2.5: Summary of FER-related works in Driving Environments (Part 1).

Ref	Method	Datasets	Face detector	Data Split	Accuracy	Limits
[103]	Hierarchical WRF	CK+ MMI KMU-FED	DLib	(5,10)folds	92.6% 76.7% 97.0%	Geometric-only features limit expression nuance capture. Struggles with occlusions, pose variation, and demographic diversity.
[65]	LMRF	KMU-FED CK+ MMI	DLib	K-fold	95.1% 93.4% 77.4%	Lacks scalability analysis for deeper models and real-time in-vehicle deployment strategy. Evaluated on limited datasets.
[189]	SVM	JAFEE	Viola and Jones	5-folds	0.976	Detects only "anger" class. Haar-based face detection struggles with non-frontal face orientations.
[26]	SVM	CK+	Haar feature	(80,20)	86.7%	Limited to five-class emotion recognition on a single lab-controlled dataset (CK+). No evaluation on mixed emotions or multimodal inputs reduces practical relevance.
[111]	AlexNet+VGG16 +GGDA	DrivFace ViVaFace MMI JAFEE CK+.	Viola and Jones	Cross subject validation	96.95% 94.6% 98.8% 98.8% 97.4%	Sensitive to hyperparameter tuning and high-dimensional features. Reliance on manually annotated, limited datasets.
[190]	lightweight SIC-NET+RC-NET	FER2013	N/A	Specific split(FER2013)	69.65%	Lightweight design may trade off representational capacity, affecting subtle expression recognition. Tested only on FER2013; generalization to in-vehicle contexts remains unclear.
[191]	NasNet-Large	JAFFE -CK+ FER-2013 AffectNet custom dataset	Improved faster R-CNN	(70,30)	98.48% 99.73% 99.95% 95.28% 99.15%	Lacks testing in varied driving conditions. The custom dataset is not fully representative. Generalizability across diverse populations is unverified.
[192]	Custom CNN VGG16	FER2013	N/A	(90,10,10) (80,20)	68.34%	FER2013 accuracy is insufficient for real use. Lacks validation on real-world driving datasets.

Table 2.6: Summary of FER-related works in Driving Environments (Part 2).

Ref	Method	Datasets	Face detector	Data Split	Accuracy	Limits
[193]	CNN VGG16	FER2013	N/A	Specific split(FER2013)	50% 38%	Focuses solely on driver distractions. Ignores external factors and lacks comparison with related work.
[194]	VGGNet	KMU-FED CK+	N/A	K-fold	95.92% 94.78%	Handling semi-occlusions only. Limited performance in complex occlusion scenarios.
[195]	DeriskNet(Deep CNN)	KDEF	N/A	(70,30)	Precision, Recall, and F1-score for each class	-Deep CNN-based design not evaluated for efficiency or real-time performance. The system lacks validation under driving variability.
[196]	SSO-DLFER(GRU)	-KDEF -KMU-FED.	RetinaNet	(70,30)	99.69% 99.50%	Does not implement ensemble learning, limiting the recognition performance of the SSO-DLFER technique.
[110]	CBiLSTM	KMU-FED subset DMD Custom experimental dataset	FAN	LOOCV	0.958(F1-score) 0.745(F1-score) 0.414(F1-score)	Suffers from class imbalance and is limited to categorical emotions. Lacks architectural enhancements like attention and hierarchical classification.
[197]	CNN+SVM	FER2013 -CK+ KDEF KMU-FED.	Haar features cascade	(80,20)	84.41% 95.05% 98.57% -98.64%	Unable to detect high-intensity emotions on challenging terrains such as rugged mountainous regions and steep valleys. Less robust to occlusions and expression variations in masked or low-light conditions.
[198]	DCNN	CK+ FER2013 KDEF KMU-FED	Viola and Jones MTCNN	N/A	96.15%¸93.41% 84.58%¸83.68% 99.18%¸98.47% 99.09%¸98.18%	DCNN lacks spatial attention, reducing sensitivity to key expression areas. Struggles with facial distortions caused by external conditions (e.g., terrain, partial occlusion).

leaky ReLU. The second uses MTCNN for face detection and alignment and a deep neural network for classification. On CK+, FER-2013, KDEF, and KMU-FED datasets, the two approaches achieved accuracy rates of 93.41%, 83.68%, 98.47%, 98.18% and 96.15%, 84.58%, 99.18%, 99.09%, respectively.

2.5.3 Limitations and Open Challenges in FER

Over the past decade, numerous FER methods and datasets have been proposed, leading to significant progress in the field. However, several challenges and limitations remain, particularly concerning the quality and diversity of available datasets, and the effectiveness, efficiency, and robustness of recognition models under challenging environments. Addressing these issues is crucial for developing more robust and generalizable FER systems.

FER Dataset-Related Challenges

One of the major challenges in FER research is (1) **the limited size and lack of diversity in available datasets**: many commonly used datasets contain a small number of samples and are often collected under controlled conditions or from demographically homogeneous populations. Characteristics such as age, gender, and ethnicity are frequently underrepresented or unlabeled, which limits the dataset's ability to support models that generalize well to real-world [9]. This results in inherent bias in FER models and restricts their robustness across diverse populations.

Another limitation is (2) **the limited availability of large-scale, video-based FER datasets**: although recent datasets like DFEW, FERV39k, and MAFW have made progress in this field, their availability is still limited compared to static image-based datasets, such as CK+, FER2013, and RAF-DB. This scarcity limits the development of temporal models such as 3D CNNs or video-based transformers [86, 171, 174], which require rich temporal data to learn meaningful spatio-temporal features. The same challenge persists in the context of *driver-specific* FER datasets, which are scarce, and as a result, most FER systems rely on general-purpose FER datasets such as CK+ and FER2013 for training. This lack of Driver FER datasets presents a significant challenge in the field, as evidenced by previous studies that primarily utilized the KMU-FED dataset as a suitable driver-specific dataset. This highlights that creating FER datasets with facial expressions of drivers is a

complex and time-consuming process, involving significant effort in terms of quantity and quality.

A further challenge is (3) **the inconsistency in annotation and labeling**: the subjective nature of emotion interpretation and the absence of standardized labeling protocols often lead to significant variability across datasets. For instance, similar expressions may be labeled differently depending on the dataset or annotators, as seen in VivaFace, DriveFace, and DMD [111]. This variability complicates performance evaluation and cross-dataset comparisons. Additionally, many datasets vary in the number and types of expressions included, some exclude key expression classes such as neutral, which can affect the model's ability to distinguish between subtle expression transitions.

Lastly, (4) **class imbalance** remains a critical issue in FER datasets, where the distribution of samples across different classes is not equal. For instance, neutral or happy expressions are dominant, while less commonly expressed expressions like "disgust" or "fear" are significantly underrepresented. This results in a skewed dataset with a significantly larger number of samples in one class compared to the others, as seen in AffectNet and RAF-DB, which leads to biases in model training toward majority classes, reducing accuracy on minority emotions.

FER Method-Related Challenges

A major challenge in FER is (1) **the limited performance and generalization of existing models**. Although many approaches demonstrate high accuracy under lab-controlled environments [167, 191], including driver-specific models evaluated on structured datasets such as KMU-FED [191, 196, 197], their effectiveness often declines in real-world settings where variations in illumination, head pose changes, and occlusion are common challenges. This reflects a limited ability to generalize to different visual conditions. For instance, the MRAN model [172], though effective in structured scenarios using multimodal inputs, has not been validated under real-world conditions involving spontaneous expressions and visual noise. Similarly, the ROI-attention model proposed by Sun et al. [16] relies on predefined facial regions and assumes consistent alignment, which limits its robustness in dynamic or uncontrolled environments. Moreover, even recent video-based FER models, which leverage temporal information for better emotion recognition, often report suboptimal accuracy when evaluated in-the-wild environments. Despite improvements, performance on

challenging datasets like DFEW or AFEW typically remains lower than many image-based methods [86, 174, 199]. Another practical limitation is (2) **the high computational complexity and resource demands**. Complex architectures, such as transformer-based [20, 86], or CLIP-based [23, 184], while achieving best performance, require substantial memory and processing time due to large model sizes. This limits their suitability for real-time or embedded applications, particularly in driving environments, where computational resources are constrained and low-latency processing is critical. Although some lightweight alternatives have been proposed, they often achieve lower accuracy [26, 65, 190]. Developing robust FER systems that balance efficiency and performance remains an open challenge.

Additionally, (3) **temporal modeling** remains a challenge in existing FER systems. While many recent works attempt to capture how expressions change over time in video sequences, modeling these dynamics accurately and efficiently is still challenging, which limits recognition performance. Some methods rely on recurrent networks, 3D convolutional models, transformer-based, or more recently, CLIP-based architectures [86, 170, 172, 188], but they often struggle to capture subtle expression transitions and long-range temporal dependencies.

In response to the outlined challenges, this thesis investigates methods to improve the performance, efficiency, and robustness of FER systems under real-world and variable conditions. At the model level, we explore architectures that improve spatial representation, enabling more effective learning of facial features under diverse visual scenarios. To address computational limitations, particularly for deployment on embedded systems, we emphasize lightweight designs that strike a balance between accuracy and efficiency. Furthermore, we examine temporal modeling to capture dynamic expression patterns in video-based inputs. These directions collectively aim to overcome key limitations in current FER approaches and contribute to the development of more robust and adaptable systems.

2.6 Conclusion

In this chapter, we provided a comprehensive overview of facial expression recognition, covering its key concepts, diverse applications, and particular importance in the driving context, along with a review of the general FER system pipeline. We also outlined commonly used datasets, evaluation metrics, and interpretability techniques, and discussed foundational traditional and deep learning methods. Additionally, a detailed literature

review was conducted to identify key limitations and open challenges that motivate the need for improved FER approaches.

In the next chapter, we introduce the first proposed method, which focuses on improving facial expression recognition performance within the driving environment.

3 A Global Context Vision Transformer for Driver Facial Expression Recognition

3.1 Introduction

As discussed in the previous chapter, facial expression recognition in driving environments plays a crucial role in enhancing road safety and supporting Advanced Driver Assistance Systems (ADAS). However, despite recent progress, existing FER methods, particularly those based on CNNs, often fall short in effectively capturing global contextual information, limiting their performance under realistic driving conditions, which involve variations in illumination, head pose, and partial occlusions. In this chapter, we present a contribution that aims to improve FER performance by introducing a novel approach leveraging a Global Context Vision Transformer model, called DFER-GCViT. The proposed model effectively captures global facial features, significantly improving recognition accuracy on the driver-specific dataset compared to CNN-based methods. Experimental validation demonstrates the robustness and effectiveness of the proposed model.

This chapter is organized as follows: Section 3.2 outlines the motivation for our contribution. Section 3.3 reviews related work in driver FER and identifies the research gap addressed in this chapter. Section 3.4 presents an overview of the baseline GC-ViT model. Section 3.5 details the proposed DFER-GCViT architecture. Section 3.6 describes the datasets, experimental setup, and discusses the results. Finally, Section 3.7 concludes the chapter by summarizing the key findings.

3.2 Motivation

With the rapid advancements in technology, a variety of techniques have emerged to improve road safety and mitigate the risks associated with human factors during driving.

One critical aspect in this domain is the recognition of the driver's facial expressions, which plays a crucial role in enabling monitoring systems such as the ADAS to understand the driver's emotional state and distraction. This understanding allows ADAS to adapt its responses, providing appropriate assistance or alerts when necessary. By accurately capturing and interpreting the facial expressions of drivers, systems can gain profound insights into the driver's emotional state, thus enabling enhanced driver safety, comfort, and overall driving experience.

However, despite the significant progress in facial expression recognition generally [9], most of the proposed models seem to mainly work well under controlled environments (eg. shopping malls) but suffer in dealing with real-world scenarios such as the driving environment (See Fig. 1) which involves additional challenges seen in changing lighting conditions, occlusions, and pose variations.



Figure 3.1: Facial expression recognition in driving environments presents additional challenges, including varying lighting conditions, occlusions, pose variations, and motion blur [200].

Previous works in the literature of driver's FER have adopted different approaches, including conventional methods and deep neural networks [26, 200]. Most of the proposed methods are mainly based on CNN variants, which have achieved relative success, but they often struggle to capture the subtle and variable facial cues encountered under the challenging conditions of the driving environment.

Recently, Vision Transformers have been introduced as an alternative to overcome certain CNN limitations. However, the standard ViT architecture, as detailed in Section 2.4.2, also presents several challenges, such as (1) the loss of important local details, which are crucial for recognizing subtle facial expressions, (2) the lack of inductive biases, such as locality and translation equivariance, that help guide feature learning, and (3) high computational costs that limit real-time applicability. These limitations motivate the introduction of a novel architecture based on a Global Context Vision Transformer (GC-ViT), which aims to enhance FER accuracy by capturing both global and local spatial dependencies and rich

semantic relationships across facial regions. By emphasizing global context, the model can better handle the complex and variable conditions of driving scenarios, where occlusion or lighting variations demand a broader understanding of the entire facial context.

3.3 Related Work

In this section, we discuss the most relevant recent advances in driver facial expression recognition, with particular emphasis on ViT-based approaches and their potential to enhance driving performance and road safety.

Researchers have proposed various techniques to accurately identify and classify facial expressions or emotions exhibited by drivers. While existing CNN- and RNN-based methods have achieved promising performance, CNNs possess inherent limitations in modeling global contextual information in facial images, which is crucial for accurate expression recognition. Due to their limited receptive fields, CNNs often struggle to distinguish subtle changes in facial expressions and to capture global spatial dependencies among facial features. To partially address these challenges, Vision Transformer (ViT) architectures [197] have demonstrated promising results in various computer vision tasks compared to earlier methods. Nonetheless, research exploring the application of ViTs to facial expression recognition remains relatively limited.

The authors in [19] proposed a modified ViT architecture, called Squeeze ViT, designed to capture both global and local facial features. Their method achieved state-of-the-art performance on both lab-controlled and in-the-wild FER datasets. On the other hand, [201] introduced the Patch Attention Convolutional Vision Transformer (PACVT) to address the challenge of facial occlusions by combining convolutional feature extraction with patch-based attention and Vision Transformer modules to selectively focus on unoccluded, discriminative facial regions. Experimental results demonstrated superior performance on both occluded and cross-dataset FER benchmarks. In [202], the TransFER model was proposed, which utilizes ViT to model the relationships between local facial regions. By enhancing patch-level interactions and addressing redundancy in attention heads, TransFER achieved state-of-the-art performance across multiple FER benchmarks.

In the context of driver FER, only two works have been reported so far. The first work by [203] proposed using the original ViT architecture to capture global attention, along

with a parallel imaging framework based on the StarGAN network for data augmentation. Experiments on CK+ and KMU-FED datasets demonstrated the effectiveness of this approach in improving FER performance under challenging conditions. The second work [204] introduces a multi-modal Vision Transformer, ViT-DD, to detect driver distraction, utilizing a semi-supervised learning approach. Experimental results showed that ViT-DD outperformed existing state-of-the-art methods, achieving notable improvements in distraction detection accuracy.

Motivated by the challenges outlined in Section 3.2 and aiming to address the limitations of existing approaches, we propose a novel Vision Transformer-based architecture for driver facial expression recognition. Specifically, the proposed approach builds upon the global context vision transformer model and leverages transfer learning to mitigate overfitting, a critical concern given the typically limited availability of training data in this domain. This is particularly important for ViT models, which generally require large-scale datasets for effective training. The main contributions of this work are summarized as follows:

- We investigate the use of Global Context Vision Transformers for facial expression recognition, with a specific focus on the driving environment. To the best of our knowledge, this is the first study that explores the potential of GC-ViT in this domain.
- We propose a novel approach called DFER-GCViT, which is based on transfer learning and extends the classifier head of the GC-ViT architecture. Our approach aims to enhance the model's performance in accurately recognizing driver facial expressions.
- We tackle the limitations faced by ViT models in the context of facial expression recognition, as well as the challenges posed by real-world scenarios through the inherent advantages of the GC-ViT architecture.
- We conduct a comparison with state-of-the-art methods on two datasets: the Keimyung University Facial Expression of Drivers (KMU-FED) dataset and the Facial Expression Recognition 2013 (FER2013) dataset.

3.4 Overview of Global Context Vision Transformer (GC-ViT)

Before introducing the proposed methodology, we provide an overview of the GC-ViT, which serves as the foundation for our approach.

GC-ViT Architecture

The GC-ViT was introduced by Hatamizadeh et al. [205] in 2022. It incorporates several key components that work together to enable effective feature extraction and classification. GC-ViT can be considered as a hierarchical ViT framework, similar in design to the Swin Transformer [134], but it introduces unique architectural elements, including a global query generator, a global self-attention module, and a downsampling operator (as illustrated in Figure 3.2). These elements distinguish GC-ViT from both Swin Transformer and the original ViT.

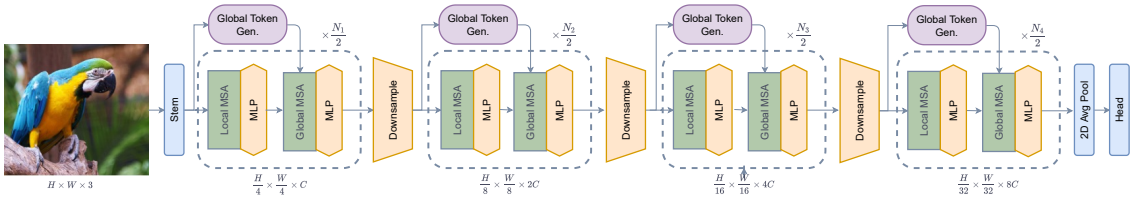


Figure 3.2: The GCViT architecture [205], showing the hierarchical structure with local and global multi-head self-attention blocks.

The GC-ViT architecture begins with the Stem/PatchEmbed layer, responsible for dividing the input images of size $H \times W \times 3$ into patches and converting them into embeddings. These patches are then processed through a series of GC-ViT steps, also known as levels, which are responsible for feature extraction. Each level consists of alternating local and global multi-headed self-attention (MHSA) blocks. These blocks apply attention mechanisms and project features into specific dimensions to enable both localized and global contextual learning. The global attention mechanism leverages a Global Token Generator to extract global features. The calculation of the global query tokens can be expressed as follows:

$$X^i = F - \text{MBConv}(X^{i-1}) \tag{3.1}$$

$$X^i = \text{MaxPool}(X^i) \tag{3.2}$$

Here, X represents the feature maps, and i denotes the stage ($i \in \{1, 2, 3, 4\}$). *F-MBConv*

block refers to the Fused-MBConv block, while MaxPool represents the max pooling layer in the generator. the Fused-MBConv block is similar to the one used in EfficientNetV2 [206] with certain extensions as below:

$$\begin{aligned} \hat{X} &= \text{DW-Conv}_{3\times3}(X), \\ \hat{X} &= \text{GELU}(\hat{X}), \\ \hat{X} &= \text{SE}(\hat{X}), \\ X &= \text{Conv}_{1\times1}(\hat{X}) + X \end{aligned} \tag{3.3}$$

SE, $GELU$, and $DW - Conv_{3\times3}$ represent the Squeeze and Excitation block, the Gaussian Error Linear Unit, and the 3×3 depth-wise convolution, respectively.

Furthermore, the Global MHSA module utilizes the extracted global query tokens to interact with the key and value local features, unlike Local MHSA, which focuses on learning key, value, and query local features. The computation of the global self-attention query, key, and value features can be described as follows:

$$\mathbf{Q_g} \in \mathbb{R}^{B^* \times C \times h \times w} := [\mathbf{q_g}, ..., \mathbf{q_g}], \mathbf{q_g} \in \mathbb{R}^{B^\times C \times h \times w} \tag{3.4}$$

$$\mathbf{q_g} \in \mathbb{R}^{B^* \times N \times C} \overset{reshape}{\leftarrow} \mathbf{Q_g} \in \mathbb{R}^{B^* \times C \times h \times w} \tag{3.5}$$

$$\mathbf{k}, \mathbf{v} = \mathbf{g}(\mathbf{x}) \in \mathbb{R}^{B^* \times N \times C} \tag{3.6}$$

where q_g represents the global attention query with dimensions $B \times C \times h \times w$, where B, C, h, and w denote batch size, embedding dimension, local window height, and width, respectively. $B^* = \text{B} \times \text{N}$, where N represents the number of local windows. Moreover, the self-attention module itself is calculated as:

$$Attention(q_g, k, v) = \text{Softmax}\left(\frac{q_g k}{\sqrt{d}} + b\right) v \tag{3.7}$$

Here, the scaling factor is denoted as d, and the learnable relative position bias term is represented by b.

After the attention blocks, Downsampling is applied to reduce the spatial dimensions of the feature maps while preserving critical information. This is achieved by using a convolutional downsampling module, which is similar to the global token generator, except it uses a CNN instead of MaxPooling. Following the final attention stage, a Pooling operation is applied to convert the feature maps from $N \times 2D$ to $N \times D$, where N represents the number of tokens, and D is the embedding dimension. Finally, a head maps the features for classification.

3.5 Proposed Methodology: DFER-GCViT

3.5.1 Overall Architecture

The proposed methodology, DFER-GCViT, leverages the base GC-ViT model for driver facial expression recognition and comprises three main stages. First, a pre-processing step is performed, which involves face detection using Multi-task Cascaded Convolutional Networks (MTCNN) [68]. The detected faces are then cropped, and data augmentation is applied to enhance the dataset. Second, the preprocessed images are passed through the DFER-GCViT model, where hierarchical features are extracted and subsequently used for facial expression classification. Figure 3.3 illustrates the architecture of our proposed model. The feature extraction and classification process is explained in detail in Figure 3.4. The step-by-step process of the DFER-GCViT methodology is summarized in Algorithm 1.

Algorithm 1 Driver Facial Expression Algorithm using Transfer Learning DFER-GCViT

1: Split the dataset into training and testing sets.
2: Apply appropriate pre-processing techniques to the dataset.
3: Build the model architecture using the pre-trained GC-ViT network parameters.
4: Fine-tune the proposed DFER-GCViT model on the training data.
5: Extract expressive features using the DFER-GCViT model.
6: Classify the expressions using the modified classifier based on the extracted features.
7: Evaluate the performance of the DFER-GCViT model on the testing set.

3.5.2 Pre-processing Module

As the initial step, the pre-processing module prepares the facial image inputs to ensure they are properly formatted and standardized for processing by the model. First, we apply

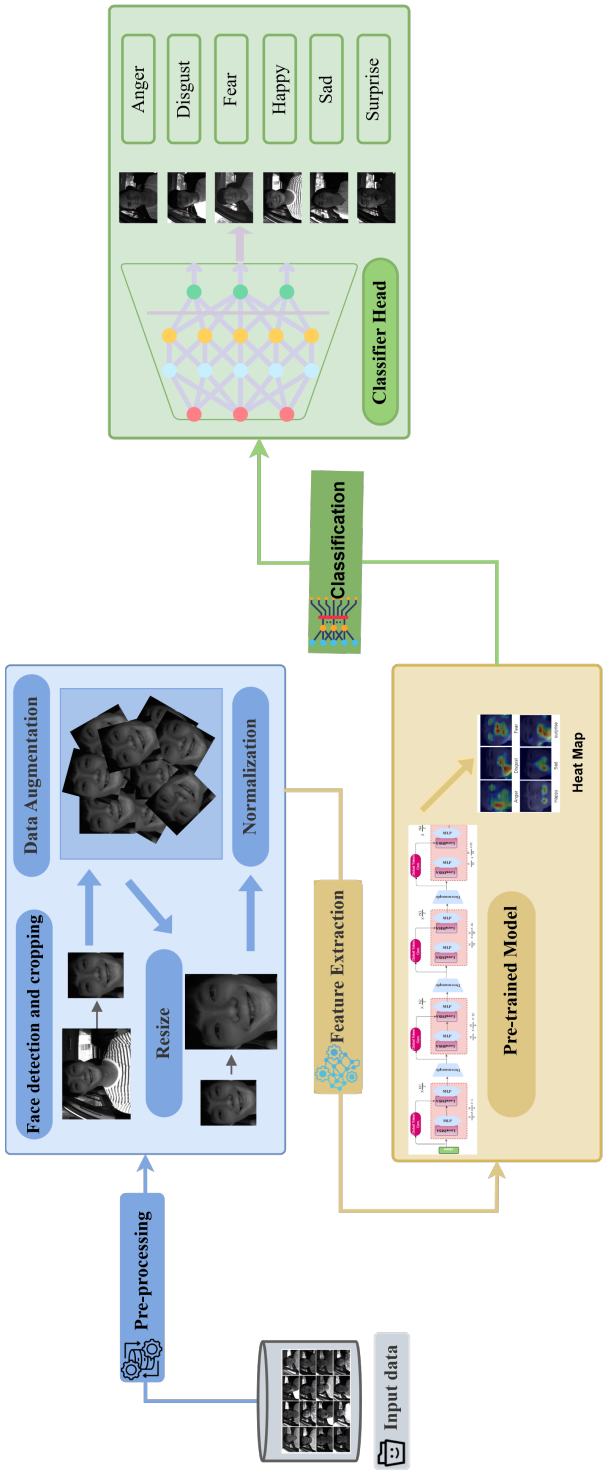


Figure 3.3: Overall workflow of the proposed DFER-GCViT methodology. The process begins with the application of a series of preprocessing techniques to the input facial images, including face detection, cropping, resizing, normalization, and data augmentation. The preprocessed images are then passed through the pre-trained GCViT model for feature extraction, producing both feature maps and attention heatmaps. Finally, the extracted features are input into a classification head to predict facial expression classes.

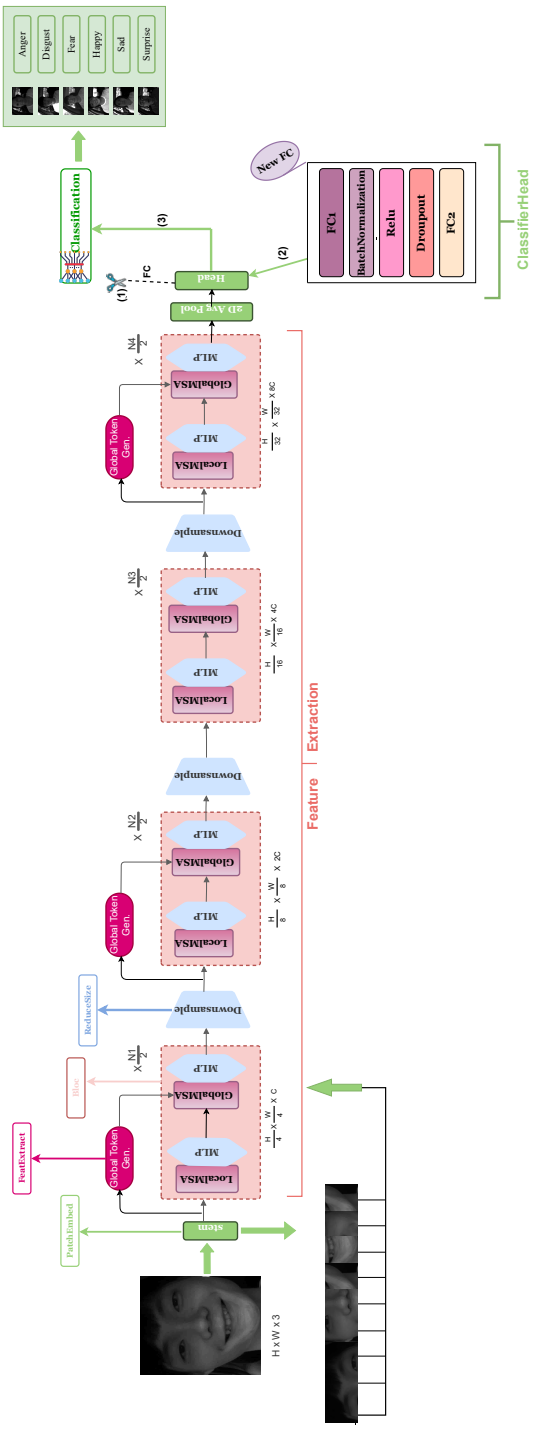


Figure 3.4: Detailed view of the feature extraction and classification modules in the GCViT-based model for driver facial expression recognition. The figure illustrates the removal of the original fully connected (FC) layer, its replacement with a new FC layer in the enhanced classifier head, and the subsequent classification of driver facial expressions.

face detection using the MTCNN face detector to detect and localize the faces in the images. The detected faces are then cropped to focus on the facial region of interest. To increase the diversity and amount of training data, we apply various data augmentation techniques as geometric transformations. These techniques include Random Horizontal Flip, Random Rotation, Random Affine, and Random Perspective. These transformations introduce variations in facial orientation, pose, and perspective, helping the model to generalize better to different facial expressions and head orientations. Additionally, we filter the Gaussian blur to reduce noise and smooth out any imperfections in the data. This helps improve the overall quality of the images and enhances the model's ability to capture meaningful facial features. After these augmentations, we resize the images to a standard input dimension of 224×224 pixels. This ensures consistency in the input size for our DFER-GCViT model. Furthermore, we standardize the pixel values across the images to help improve the convergence of the model during training and ensure that the features have a consistent scale and range.

3.5.3 Feature Extraction Module

The GC-ViT architecture consists of two primary components: feature extraction and classification. In the proposed model, we employ transfer learning by leveraging the feature extraction component to obtain informative representations from the preprocessed input images. To accomplish this, we utilize the pre-trained weights and architecture of GC-ViT, allowing us to extract patterns and representations from preprocessed images with dimensions of $224 \times 224 \times 3$. Through this process, the images are transformed into a set of feature maps that encode hierarchical representations, capturing both local and global contextual information of facial expressions at each stage of the GC-ViT (as explained in section 3.4). The extracted features provide a rich representation of the input images, enabling subsequent classification tasks. The overall workflow is depicted in Figure 3.4, providing a visual representation of the feature extraction process.

3.5.4 Driver Facial Expression Recognition Module

The driver facial expression recognition module based on GC-ViT is a crucial component responsible for identifying and categorizing facial expressions exhibited by drivers. In our approach, we modify the classifier head to enhance performance. Specifically, instead of

using a traditional fully connected (FC) layer, we have replaced it with a sequential module consisting of two linear layers, a batch normalization layer, a ReLU activation layer, and a dropout layer. This modification, illustrated in the ClassifierHead part of Figure 3.4, is designed to further improve the model's ability to refine and classify expressive features extracted from the driver's face.

The first linear layer (FC1) performs a linear transformation of the input, which allows the model to learn more expressive features from the data. The subsequent batch normalization layer normalizes the activations, leading to improved stability and faster convergence during training. The introduction of the ReLU activation function introduces non-linearity, allowing the model to capture complex feature relationships and enhance its discriminative power. Additionally, the dropout layer is applied during training to mitigate overfitting by randomly dropping out a fraction of the activations, promoting better generalization to unseen data. The output of the second linear layer (FC2) serves as the final prediction for the facial expression recognition, providing the corresponding classification results. This modified configuration of the ClassifierHead not only enhances the discrimination and generalization capabilities of the GC-ViT model but also allows the model to be fine-tuned for the specific nature of facial expression recognition. By leveraging the structure of the sequential module, the model is expected to better refine high-level expressive features extracted by the backbone and improve overall classification performance.

3.6 Experimental Analysis

We evaluated the performance of the proposed method using two publicly available benchmark datasets: KMU-FED (Keimyung University Facial Expression of Drivers) [103] and FER-2013 (Facial Expression Recognition 2013) [97]. This section outlines the evaluation strategy, provides details on the experimental setup, and presents a comparative analysis of the results obtained.

3.6.1 Dataset Description

The KMU-FED dataset [103], as previously described in section 2.3.1, contains facial expression images of drivers categorized into six emotion classes. To ensure robust eval-

uation and fair performance comparison, we applied 10-fold cross-validation during our experiments.

The FER-2013 dataset [97], also introduced in section 2.3.1, provides a large set of grayscale facial images from unconstrained environments (In-the-wild), containing seven expression classes. For model training and evaluation, we used the original data splits: training, public test, and private test.

3.6.2 Implementation Setup

The proposed DFER-GCViT model was implemented using the PyTorch open-source deep learning framework and executed on an NVIDIA Quadro RTX 5000 GPU with 16 GB of RAM. The DFER-GCViT model is based on the GCViT-Base variant, a pre-trained version of the global context vision transformer architecture, which is used for both feature extraction and classification.

For the KMU-FED dataset, the training process utilized a batch size of 16. The Adaptive Moment Estimation (Adam) optimizer was employed with a fixed learning rate of 0.0001. The cross-entropy loss function was utilized, the model was trained for 60 epochs, and a 10-fold cross-validation scheme was applied. For the FER2013 dataset, a batch size of 16 was used. The Stochastic Gradient Descent (SGD) optimizer was employed with a momentum value of 0.9 and a learning rate decay of 0.002. Similar to the KMU-FED dataset, the cross-entropy loss function was used, and the model was trained for a total of 130 epochs. A summary of these implementation details is provided in Table 3.1.

3.6.3 Results and Discussion

To evaluate the effectiveness of our proposed DFER-GCViT method, we conducted comparative experiments on the KMU-FED and FER2013 datasets, comparing their performance against state-of-the-art methods.

Results on KMU-FED dataset

On the KMU-FED dataset, we conducted a comparative analysis between our proposed DFER-GCViT model and three other methods to demonstrate the effectiveness of our

Table 3.1: Hyperparameter Settings for the DFER-GCViT model on the KMU-FED and FER2013 datasets. "–" indicates not applicable.

Hyperparameters	**KMU-FED dataset**	**FER2013 dataset**
Optimizer	Adam	SGD
Loss Function	Cross-Entropy	Cross-Entropy
Batch size	16	16
Learning rate	0.0001	0.002
Epochs	60	130
Folds	10	–

modified architecture in improving accuracy. The comparison includes the following methods: (1) a hierarchical weighted random forest (hierarchical WRF) classifier that used geometric features [103], (2) a connected convolutional neural network (CCNN) that combines low-level and high-level features [69], and (3) a modified version of SqueezeNet 1.1, where the last layer has been adjusted [207].

Our DFER-GCViT model achieved the highest performance, with an average accuracy of 98.2%, outperforming all compared state-of-the-art methods. Specifically, the hierarchical WRF achieved an accuracy of 94.7%, resulting in a 3.5% difference compared to our model, likely due to its reliance on handcrafted geometric features, which may not capture subtle appearance-based variations effectively. The CCNN attained 97.3%, demonstrating competitive performance, though its CNN-based architecture lacks the global contextual features. The modified SqueezeNet 1.1 recorded a significantly lower accuracy of 83.43%, reflecting the limitations of lightweight CNNs in handling complex expression variations. A summary of these results is presented in Table 3.2. Additionally, we compared our proposed model with the original GC-ViT (base version) using the same dataset. In this baseline, only the number of output classes was adapted to match the KMU-FED dataset, without modifying the classifier head. This model achieved an average accuracy of 97.6%, while our proposed DFER-GCViT model, with a modified classifier head, achieved 98.2%. The 0.6% improvement highlights the effectiveness of the new classifier head in better utilizing facial features for classification.

Table 3.2: Comparison between the proposed method and state-of-the-art methods on the KMU-FED dataset.

Methods	**Accuracy**
hierarchical WRF [103]	94.7%
CCNN [69]	97.3%
modified SqueezeNet 1.1 [207]	83.4%
DFER-GCViT (Our method)	**98.2%**

Results on FER2013 dataset

In the evaluation on the FER2013 dataset, our proposed DFER-GCViT method was compared against several state-of-the-art methods, including (1) a deep FER model integrating both CNNs and the ensemble deep learning algorithm [208], (2) an adapted VGGNet configuration consisting of four convolutional stages and three fully connected layers [209], and (3) a modified version of SqueezeNet 1.1 [207]. Remarkably, our DFER-GCViT achieved an accuracy of 73.78%, outperforming all compared methods. Specifically, it outperformed deep FER by 1.87% and the modified SqueezeNet by an impressive 12.69%. Although our model achieved only slightly higher accuracy than the adapted VGGNet by 0.5%, the results were still interesting. The results are summarized in Table 3.3. Additionally, we conducted a similar comparison on the KMU-FED dataset with the original GC-ViT method. Our proposed method exhibited superior performance, outperforming the original GC-ViT method by 1% (achieving an accuracy of 73.8% compared to the original method's accuracy of 72.8%). These findings highlight the effectiveness of the proposed method in accurately recognizing facial expressions on the FER2013 dataset.

We observed that the accuracy on the FER2013 dataset is lower in comparison to the accuracy achieved on the KMU-FED dataset. This disparity can be attributed to the challenges present within the FER2013 dataset, such as the presence of corrupted images, noisy images, and low-resolution images.

Table 3.3: Comparison between the proposed method and state-of-the-art methods on the FER2013 dataset.

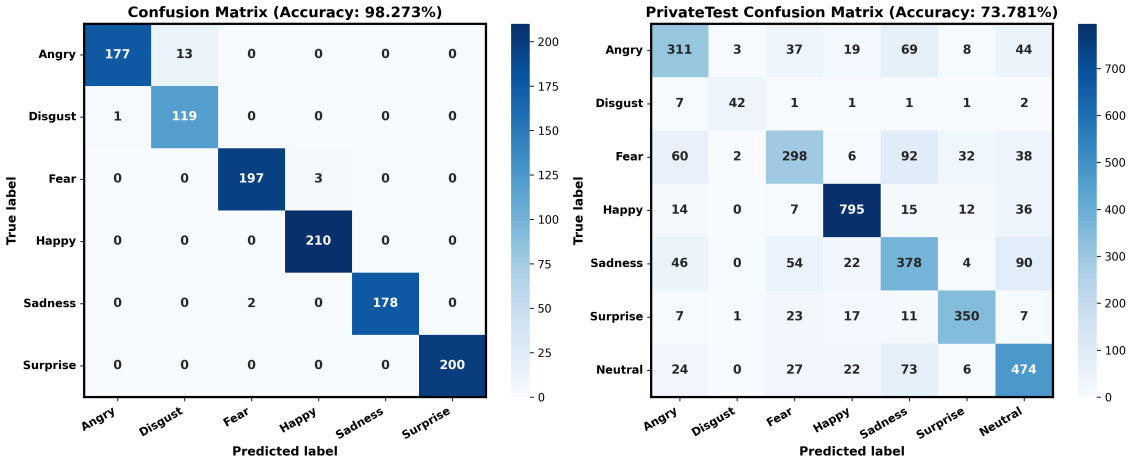
Methods	**Accuracy**
deep FER[208]	71.91%
VGG [209]	73.28%
modified SqueezeNet 1.1 [207]	61.09%
DFER-GCViT (Our method)	**73.78%**

Visualization-Based Analysis

To further evaluate the performance of the proposed DFER-GCViT model, we conducted a visualization-based interpretability analysis using the KMU-FED and FER-2013 datasets. First, we examine the confusion matrices shown in Figure 3.5, which quantitatively illustrate the model's ability to distinguish between different facial expressions. Next, we analyze the visualization heatmaps presented in Figure 3.6, which provide qualitative insight into the model's decision-making process by highlighting the facial regions most influential in classification.

Figure 3.5 (a) presents the confusion matrix for the KMU-FED dataset, illustrating how well the model distinguishes between different facial expressions.

The expressions of happy and surprise achieve the highest accuracy, with no instances of misclassification. Conversely, the expressions of fear and sadness exhibit a relatively small number of misclassifications. However, the model struggles to accurately distinguish between the angry and disgust expressions, resulting in a significant number of misclassifications for the angry. This difficulty can be attributed to the similarities in facial features and muscle movements associated with anger and disgust expressions. Moving on to the FER2013 dataset, as depicted in Figure 3.5 (b), the expressions of happy and neutral demonstrate the highest performance, with accurate classification. The model also appears confident in recognizing the angry and sad classes. However, similar to the observations from the KMU-FED dataset, these expressions may share visual similarities with other expressions, making it more challenging for the model to distinguish them accurately.



(a) KMU-FED (b) FER2013

Figure 3.5: Confusion matrices of the KMU-FED dataset and FER2013 dataset.

To gain insights into the model's reasoning and decision-making process, we employed the Gradient-weighted Class Activation Mapping (Grad-CAM) [210] technique to visualize the attention regions of the DFER-GCViT model. Grad-CAM generates heatmaps that highlight the areas of the input image that contribute most to the model's classification decisions. We focused on visualizing the layers before the classification layer, specifically the layers in the block of stage 4 from the GC-ViT architecture (see Figure 3.4), as these layers capture high-level features and semantic information. The colors in the heatmap represent the importance of different regions of the face. Warmer colors, such as red and yellow, indicate the most relevant regions during the classification process, while cooler colors like blue and green represent less important regions.

By analyzing the heatmaps, we can observe how the model localizes and focuses on specific regions of the face for different facial expressions. In the case of the KMU-FED dataset, as shown in Figure 3.6 (a), the model exhibits concentration on discriminative features for the target expressions. Most expressions focus primarily on the mouth region, highlighting its important role in distinguishing between these facial expressions. Additionally, specific expressions show certain features that contribute more to the classification. However, for the anger class, we observe that the model focuses on the wrong region, neglecting other important features. This emphasizes the lower performance of the model in accurately classifying the anger expression, as indicated by the confusion matrix. Similarly, with the FER2013 dataset depicted in Figure 3.6 (b), the model's focus varies for each specific



(a) KMU-FED (b) FER2013

Figure 3.6: Heatmap visualization of DFER-GCViT using Grad-CAM technique for facial expressions on KMU-FED and FER2013 datasets.

expression, highlighting the regions such as eyes and eyebrows that contribute the most to the classification decision. Understanding these patterns provides valuable insights into the visual cues that drive the model's decision-making process and can aid in the interpretation and improvement of the facial expression recognition system.

3.7 Conclusion

In this chapter, we have explored the application of GC-ViT for facial expression recognition by introducing a novel transfer learning-based method, DFER-GCViT. Built upon the global context vision transformer to leverage its ability to capture both local and global features using a self-attention mechanism. Our approach focused on modifying the classifier head of the original GC-ViT by adding additional layers to accurately recognize facial expressions and emotions exhibited by drivers. The proposed method demonstrated competitive performance compared to state-of-the-art approaches. These results demonstrate the effectiveness of our approach in improving accuracy, allowing for accurate recognition of facial expressions despite the inherent challenges in realistic scenarios, such as occlusions, head pose variations, and varying illumination conditions. This research highlights the potential of vision transformers in driver facial expression recognition, offering a novel perspective compared to previous methods.

While the method shows strong performance, further refinement is needed to enhance its

practical applicability. In the next chapter, we introduce a more efficient and lightweight architecture that aims to improve both accuracy and computational efficiency by combining the strengths of vision transformers and convolutional neural networks.

4 Shuffle Vision Transformer: Lightweight, Fast and Efficient Recognition of Driver Facial Expression

4.1 Introduction

Existing methods for facial expression recognition, particularly in driving environments, have shown promising results but often rely on computationally intensive models, as previously noted, limiting their deployment in resource-constrained environments. In this chapter, we address this limitation by introducing a lightweight, transfer learning-based dual architecture named ShuffViT-DFER, designed to balance efficiency and accuracy. This is achieved by harnessing the strengths of two lightweight and efficient models, CNN and ViT. By effectively integrating features from both models, the proposed architecture enhances recognition performance while maintaining low computational cost. Experiments conducted on two publicly available datasets, KMU-FED and KDEF, demonstrate that the method achieves competitive performance compared to existing state-of-the-art approaches, indicating its potential for efficient facial expression recognition.

The motivations behind this contribution are outlined in Section 4.2. Section 4.3 reviews related work and positions our approach within the existing literature. An overview of the selected pre-trained models is provided in Section 4.4, followed by a detailed description of the proposed architecture in Section 4.5. Section 4.6 presents the datasets used, experimental setup, and performance analysis. Finally, Section 4.7 summarizes the main findings of the chapter.

4.2 Motivation

Facial expression recognition in the driving environment has gained increasing attention due to its potential to enhance road safety, as emphasized in the prior chapters. In

this context, several research works have focused on the development of methods for the recognition of drivers' facial expressions, such as [26, 192, 193, 211]. However, these methods often struggle to achieve both high accuracy and computational efficiency, especially in challenging scenarios involving non-frontal head positions, occlusions, and variation in lighting conditions.

While some approaches, especially deep learning-based models (e.g., [212], [213]) demonstrate superior performance compared to conventional techniques (e.g., [26], [189]), they typically require large amounts of data and significant computational resources, making them less suitable for real-time applications. Nevertheless, some recent studies have attempted to address the computational challenge by introducing lightweight models designed to operate efficiently with low-latency processing at the cost of lower performance, as noted in [65, 195]. These challenges motivate the need for approaches that balance accuracy with computational efficiency, particularly for deployment on embedded systems. A detailed review of related work is provided in the following section.

4.3 Related Work

In this section, we review existing works that focus on using pretrained deep learning-based architectures for facial expression recognition. For instance, the study in [214] introduced a modified ResNet18 network with an enhanced feature attention (EFA) module to extract rich features from facial expression images. It employed a joint discriminative correlation alignment (JDCA) loss to align feature distributions between single-driver (SD) and two-driver (TD) images while leveraging label information for driver facial expression recognition. RAF-DB, FEAR2013, AffectNet, and KMU-FED datasets were used to train the network for best performance. In [215], the authors used transfer learning with pre-trained CNN architectures such as AlexNet, SqueezeNet, and VGG19 to perform facial emotion recognition tasks. The models were evaluated on various benchmark datasets with real-time in-vehicle challenges, including FER2013, JAFFE, KDEF, CK+, SFEW, and KMU-FED. Results indicated that the pre-trained VGG19 model generally beat AlexNet and SqueezeNet, achieving the highest accuracy of 99.7% on the KMU-FED dataset. The other models competed with state-of-the-art methods and obtained similar performance across benchmark databases.

On the other hand, in [216] they evaluated their approach on the custom dataset under various real-world scenarios, using AlexNet and VGG16 for feature extraction and GGDA for emotion recognition. To detect the driver's road rage, a pre-trained VGG16 model was used to extract features, and the classification of emotions under the challenges of multi-pose and varying illumination conditions proposed in [217] achieved an accuracy of 78.57%, 61.98%, and 94.27% on the RAF-DB, AffectNet, and KMU-FED datasets. They used OpenCV 3.3, a deep neural network (DNN) for face detection. A real-time framework for stress detection was presented by [69]. It consists of three modules: face detection using MTCNN, a proposed connected convolutional network (CCNN) that combines low-level and high-level features for the facial expression module, and the last module is stress detection. The proposed facial expression model achieved comparable performance with an accuracy of 93.6% on the CK+ dataset, 81.3% on Oulu-CASIA dataset, and 97.3% on KMU-FED. The authors of [200] proposed a new FER approach called FERDERnet, which was performed on a dataset collected from on-road drivers. The approach consists of three modules: the face detection module (FD) to detect the driver's face, the augmentation-based resampling module (ABR) to reduce training dataset imbalance and improve the model's generalization, and the emotion recognition module (ER) to classify on-road driver emotions. The paper conducted different experiments with some baseline networks as backbones, such as Googlenet, Resnet50, InceptionV3, and Xception, to evaluate the proposed approach on the on-road driver facial expression dataset. The proposed model with the Xception backbone outperformed the other backbones and achieved a high recognition accuracy of 0.966.

In summary, while prior studies have made progress in driver facial expression recognition using pretrained CNN architectures, several critical limitations remain. As noted in the previous section, most approaches rely on large and computationally demanding models such as VGG19, ResNet50, and InceptionV3, which are not well-suited for real-time or embedded automotive systems. Furthermore, many of these methods overlook the importance of capturing global features, a critical factor for recognizing subtle emotional expressions under challenging conditions. In addition, few works explicitly address model efficiency, which is essential for FER in constrained environments. In contrast to existing methods, this chapter presents a novel transfer learning-based approach that balances recognition performance with computational efficiency by leveraging lightweight, resource-efficient models. This makes the proposed method particularly well-suited for deployment on embedded systems. The main contributions of our work are summarized as follows:

- We introduce ShuffViT-DFER, a novel lightweight and efficient approach for driver's facial expression recognition.
- We leverage the strengths of ShuffleNet V2 [25] and EfficientViT [218] architectures, exploiting features from both models with a new classification scheme, achieving accurate and fast recognition.
- We demonstrate that dual-architecture transfer learning allows for efficient capture of subtle facial cues and expressions with limited data while maintaining low-latency processing suitable for real-time performance.
- We explore the use of the Efficient ViT approach in drivers' facial expression recognition and application of the Grid Search technique to optimize the model's hyperparameters for improved performance.
- We conduct extensive experiments and evaluations on two benchmarking and publicly available databases, obtaining interesting performance compared to the state-of-the-art methods.

4.4 Pre-trained Models for FER: An Overview

As discussed in Chapter 2 and earlier in this chapter, deep learning techniques, particularly CNNs and ViTs, have significantly advanced facial expression recognition, consistently outperforming conventional handcrafted approaches [26, 103]. However, deploying FER systems in resource-constrained environments remains challenging due to the high computational and memory demands of these models. This section provides a brief overview of widely used pre-trained CNN and ViT models, with a focus on those selected for the proposed methodology.

4.4.1 Pre-trained CNN Architectures

CNNs (previously detailed in Section 2.4), are among the most widely used neural network architectures for image classification tasks, including facial expression recognition. Several CNN architectures have been pre-trained on large-scale datasets such as ImageNet [219], and these pre-trained models have become the standard choice for transfer learning, particularly

in scenarios where labeled FER data is limited. Utilizing pre-trained CNNs enhances feature extraction capabilities and contributes to more accurate and efficient classification. Below, we provide an overview of commonly used pre-trained CNN architectures in FER, including several lightweight models suitable for deployment in resource-constrained environments.

AlexNet [122] is one of the pioneering deep convolutional neural networks that significantly advanced image classification in 2012. It consists of five convolutional layers followed by three fully connected layers and uses ReLU activations, max pooling, dropout, and data augmentation to improve performance and reduce overfitting. Trained on the ImageNet dataset using two GPUs, as it is considered computationally intensive, AlexNet achieved state-of-the-art results and set a new benchmark for deep learning in computer vision.

VGG [123] VGGNet is a deep convolutional neural network architecture proposed by Oxford's Visual Geometry Group in 2014. It is characterized by a simple and consistent architecture, using small 3×3 convolutional filters throughout the network. VGG improves performance by significantly increasing the depth of the network, typically using 16 or 19 layers. The model was originally trained on multiple GPUs, and its pre-trained weights have become a widely used baseline for feature extraction in various computer vision tasks. Despite its efficiency, VGGNet contains around 138 million parameters, making its deployment relatively large and computationally intensive.

ResNet [124] Residual Neural Network is a deep convolutional neural network architecture that introduced the concept of residual connections or skip connections, along with features of heavy batch normalization. These connections help address the degradation problem often observed in very deep networks (e.g., in RNNs) by enabling the model to learn identity mappings, allowing gradients to pass through layers more effectively without vanishing or exploding. As a result, ResNet enables the successful training of very deep architectures while maintaining lower complexity compared to models like VGGNet in 2015. Several variants of ResNet exist, differing in the number of layers, including ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152.

Xception [125] developed by Google in 2017, is a deep convolutional neural network architecture known for its innovative approach to feature extraction. It is based on depthwise separable convolutions, which decompose standard convolutions into two distinct steps: a depthwise convolution, where spatial filtering is applied independently to each input channel, and a pointwise convolution (1×1) that combines these filtered outputs across channels. This unique structure reduces the number of parameters while preserving

model performance. It comprises 36 convolutional layers, grouped into 14 distinct modules, with linear residual connections added around each module except for the first and last. Xception has achieved impressive results on various benchmark tasks.

MobileNet [126] is a family of convolutional neural network architectures optimized for efficient performance on mobile and edge devices. First introduced by Google in 2017, MobileNet models focus on reducing computational complexity without compromising performance. They utilized depthwise separable convolutions, which reduce the number of parameters and computational cost while maintaining competitive accuracy. Over time, the MobileNet family has evolved through multiple versions, including MobileNetV1, V2, and V3, each offering further advancements in speed and efficiency. MobileNetV1, in particular, demonstrates its effectiveness despite its lightweight design by achieving a top-5 classification accuracy of 89.9% on the ImageNet dataset.

SqueezeNet [220] is introduced in 2016, is a CNN architecture that achieves AlexNet-level accuracy while significantly reducing the number of parameters. Its lightweight architecture is designed to reduce model size while maintaining high performance. The core component of SqueezeNet is called the "fire module", which consists of two key layers: a squeeze layer and an expand layer. The squeeze layer uses 1×1 convolutions to compress the input and reduce dimensionality, while the expand layer applies a combination of 1×1 and 3×3 filters to increase channel depth and extract richer feature representations. Thanks to its design, SqueezeNet has been widely adopted in various domains, particularly for deployment on mobile and embedded devices where efficiency is critical.

EfficientNet [127] is introduced by Google in 2019, is a family of CNN architectures that achieve high performance on image classification tasks while maintaining a small number of parameters and computations. The key innovation of EfficientNet is a compound scaling method that uniformly scales the depth, width, and resolution of the network to optimize the trade-off between accuracy and computational resources. The EfficientNet variants start with a baseline model (EfficientNet-B0) and then expand into deeper variants (B1 to B7), each offering improved accuracy with significantly fewer parameters, making it suitable for resource-constrained applications.

ShuffleNet [221] is introduced by Megvii (Face++) in 2017, is an efficient CNN architecture specifically designed for mobile and low-power devices. It focuses on achieving a balance between high accuracy and low computational cost. The key innovations in ShuffleNet incorporate pointwise group convolutions and a novel operation called channel shuffle.

Pointwise group convolutions reduce computational complexity by splitting the input channels into groups, but they can limit information flow between groups. To address this, the channel shuffle operation is introduced to enable cross-group communication by reordering the channels, thus reducing computation cost while maintaining accuracy. It effectively achieves a lower top-1 error than the MobileNet model on ImageNet classification and achieves approximately a 13× actual speedup compared to AlexNet while maintaining comparable accuracy.

4.4.2 Pre-trained Vision Transformers architectures

While CNNs have long been dominant in computer vision, vision transformers (detailed in Section 2.4) have recently emerged as a powerful alternative. They have shown competitive or even superior performance to CNNs in tasks such as image classification, leading to the development of several transformer-based architectures [222].

DeiT [133] The Data-efficient Image Transformer (DeiT) is an improved vision transformer model introduced in 2021, designed to achieve high performance on image classification tasks while requiring significantly less training data than conventional vision transformers. Its architecture is based on ViT, with several modifications that enhance its computational efficiency and make it more suitable for data-limited scenarios. DeiT employs a distillation-based training approach, where the model learns to mimic the behavior of a larger, more accurate teacher model. Additionally, it incorporates data augmentation techniques to increase the diversity and quality of the training data. These enhancements allow DeiT to maintain the strengths of transformer-based models with low computational cost.

MobileViT [135] is a lightweight, general-purpose vision transformer introduced in 2022, designed for mobile devices. It offers a unique approach to global processing of information using transformers. Standard ViT models require more parameters and exhibit substandard optimizability due to their disregard of spatial inductive bias. MobileViT aims to learn global representations using transformers as convolutions, allowing for implicit incorporation of convolution-like properties, simple training recipes, and easy integration with downstream architectures. This approach is sensitive to L2 regularization and relies on extensive data augmentation strategies to mitigate overfitting.

LVT [223] The Lite Vision Transformer (LVT) is a lightweight, compact architecture designed for mobile use in visual recognition tasks. It incorporates Convolutional Self-

Attention (CSA) and Recursive Atrous Self-Attention (RASA) to improve feature representation at both low and high network levels. CSA enriches low-level features while maintaining spatial awareness and computational efficiency. RASA captures multi-scale contextual relationships, boosting representational capacity with minimal parameter overhead. LVT outperforms previous lightweight models on benchmark tasks, making it an effective transformer backbone for real-time and resource-constrained applications.

EfficientViT [218] is a new high-speed vision transformer model introduced in 2023 to address inefficient memory operations in current models. The model uses a single memory-bound Multi-Head Self-Attention network between efficient FFN layers, improving memory efficiency and overall channel communication. It also addresses computational redundancy by introducing a cascaded group attention module that feeds attention heads with different splits of the full feature, saving computational costs and improving attention diversity. EfficientViT outperforms existing efficient models for computer vision while striking a good trade-off between accuracy and speed. The EfficientViT model family includes 6 variants that differ in depth and width scales, with a set number of attention heads assigned to each stage, starting from the baseline model, EfficientViT-M0.

4.4.3 The Selected Pre-trained Models for Driver's FER

In this work, we selected two lightweight and efficient architectures: ShuffleNet V2 [25], a CNN-based model, and EfficientViT-M2 [218], a ViT-based model. These architectures were chosen for their complementary strengths in enhancing feature representation while maintaining low computational cost. While other lightweight models, such as MobileNet [224] and MobileViT [225] have also been explored in the literature, ShuffleNet V2 and EfficientViT-M2 offered a favorable balance between accuracy and efficiency for our target application. Together, they enable effective capture of both local and global facial features, including subtle expression cues that are crucial for accurate recognition. In the following section, we describe how these architectures are utilized in our proposed approach. A brief overview of each architecture is first provided below.

ShuffleNet V2

ShuffleNet V2 [25] is designed to address the issues of the original ShuffleNet, particularly those related to computational efficiency and network design balance. It introduces a more

balanced approach to achieving high accuracy while maintaining low computational cost. The design is based on four key guidelines:

- G1) Equal channel width minimizes memory access cost (MAC).
- G2) Excessive group convolutions increase MAC.
- G3) Network fragmentation reduces the degree of parallelism.
- G4) Element-wise operations are non-negligible performance.

As depicted in Figure 4.1, the first two blocks on the left illustrate the basic and spatial downsampling units of ShuffleNet V1 [221], while the last two blocks on the right represent the basic unit and spatial downsampling unit in ShuffleNet V2. To follow the guidelines, ShuffleNet V2 incorporates several architectural modifications.

First, a channel split operation is applied at the beginning of each unit. The input feature map with c channels is divided into two branches with $c - c'$ and c' channels, respectively. According to guideline G3, one branch is retained as an identity mapping, while the other undergoes transformation through three convolutional layers, each with equal input and output channels, in line with G1. The two 1×1 pointwise convolutions in this branch are no longer group-wise, partially conforming to G2, since the initial split already introduces channel separation. After convolution, the two branches are concatenated, restoring the original number of channels (G1). A channel shuffle operation, similar to that in ShuffleNet V1, is then applied to enable feature interaction across the branches. Following this, the next unit begins. Element-wise operations, such as ReLU and depthwise convolutions, are applied only within the convolutional branch. Additionally, the three operations, concatenation, channel shuffle, and channel split, are combined into a single element-wise operation, as described in G4. Finally, in the spatial downsampling unit (the last block), the channel split operation is omitted, and the number of output channels is doubled.

EfficientViT-M2

As discussed earlier, EfficientViT [218] is a new family of fast and memory-efficient vision transformers designed to address several limitations of traditional transformer architectures. These limitations include high computational complexity, memory inefficiency in transformer

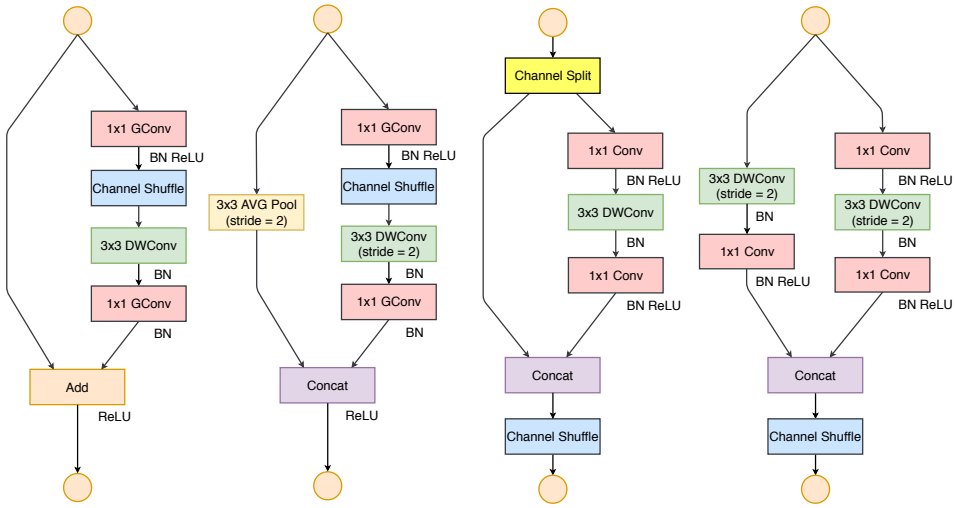


Figure 4.1: Core Architectural Units of ShuffleNet V1 and ShuffleNet V2 [25].

operations, and computational redundancy across attention heads. To address these challenges, EfficientViT introduces several architectural innovations:

- (1) Sandwich Layout, the model employs a new sandwich layout for efficient memory block construction, reducing memory usage and enabling efficient channel communication through more memory-efficient feed-forward networks. This design also introduces an extra interaction token layer, enhancing model capacity and introducing inductive bias of local structural information.

- (2) Cascaded Group Attention (CGA), the MHSA layers face redundancy issues, making computations inefficient. To address this, a new attention module, CGA, is proposed. Inspired by group convolutions in efficient CNNs, it splits full features across heads, reducing computation and improving efficiency. The model continues to improve accuracy and capacity by encouraging layers to learn projections on features with richer information.

- (3) Parameter Reallocation, The EfficientViT framework improves parameter efficiency by reallocating network parameters by expanding critical module widths and shrinking non-important ones. The model uses Taylor analysis to set small channel dimensions or maintain input dimensions. The feed-forward network's expansion ratio is reduced to 2 for parameter redundancy. The model also removes redundant parameters in unimportant modules.

The overall architecture of EfficientViT is illustrated in Figure 4.2. The network begins with an overlapping patch embedding that converts 16×16 image patches into tokens of dimension C_1. The architecture is composed of three main stages, each consisting of stacked EfficientViT blocks. At every subsampling layer, the number of tokens is reduced by a factor of 4, corresponding to a 2x reduction in spatial resolution. For this purpose, EfficientViT introduces a subsample block, which also contains a sandwich layout. However, the self-attention layer is replaced with an inverted residual block to help preserve information during subsampling. Throughout the model, Batch Normalization (BN) is used instead of Layer Normalization (LN), as BN can be integrated into preceding convolutional or linear layers, offering better runtime efficiency. Additionally, ReLU is utilized as the activation function.

The EfficientViT-M2 variant is used in our proposed method. It shares the same architecture as the base model but differs in depth, width, and the number of attention heads across the three stages, with configurations of $\{128, 192, 224\}$, $\{1, 2, 3\}$, and $\{4, 3, 2\}$, respectively.

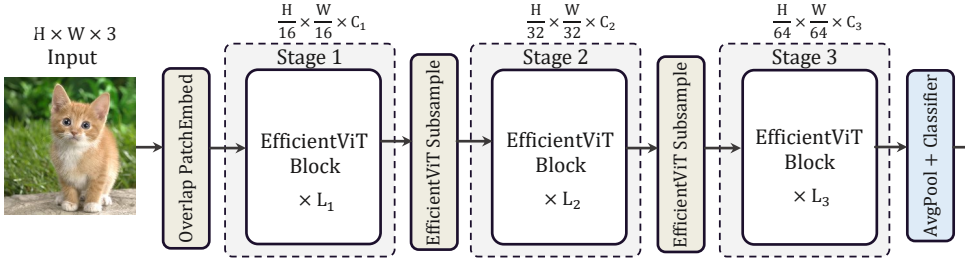


Figure 4.2: The architecture of EfficientViT [218].

4.5 Proposed Model: ShuffViT-DFER

4.5.1 Overall architecture

The pipeline of our proposed architecture is illustrated in Figure 4.3. The input to the model consists of a cropped facial image, detected using MTCNN [68]. Subsequently, the detected faces undergo data augmentation, such as horizontal flipping, rotation, and scaling, to enhance the diversity and size of the training set and improve model generalization. Feature extraction is then performed using the two pretrained models, ShuffleNet V2 [25]

and EfficientViT-M2 [218], to overcome limited training data and benefit from the complementary strengths of CNNs and ViTs. The extracted features from the two models are fused and passed to a classifier to enable accurate expression recognition.

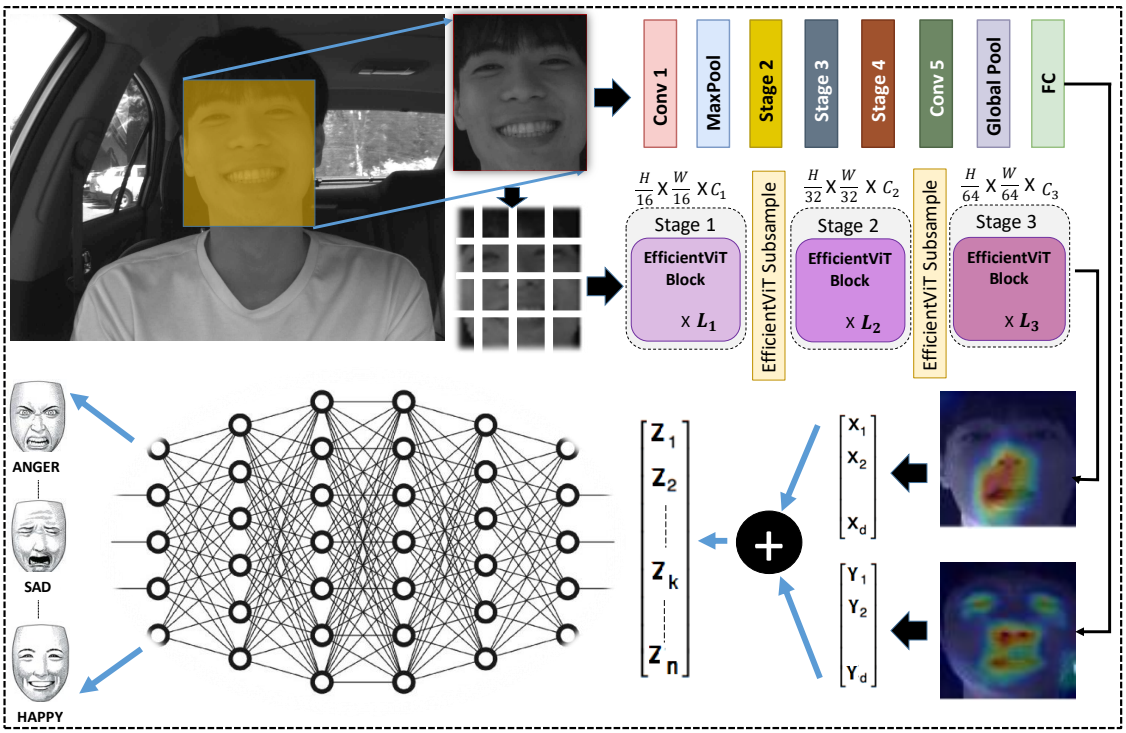


Figure 4.3: The workflow of our proposed model for driver facial expression recognition. The pipeline begins with face detection using the MTCNN, followed by data augmentation techniques applied to the cropped face. The preprocessed images are then processed in parallel by two pretrained models, ShuffleNet-V2 and EfficientViT-M2. The extracted features are fused and passed through a custom classifier to predict the driver's facial expression accurately.

4.5.2 Pre-processing

Several image preprocessing techniques were used to improve the training data for effective feature extraction and classification. The face was first detected using MTCNN, then cropped and resized to 224×224 pixels. In addition, diverse transformations were used as part of data augmentation to alleviate overfitting issues related to small FER datasets. These transformations included Random Horizontal Flip, Random Rotation, ColorJitter, Random Affine, and Gaussian Blur, which introduced variations in appearance and position,

enhancing the dataset's variety. Additionally, pixel value normalization was performed to maintain a consistent scale for features.

4.5.3 Lightweight and efficient feature extraction

We extract features using two parallel branches based on ShuffleNet V2 and EfficientViT-M2 models, employing transfer learning to address the limited data constraints and harness the strengths of both models for effective feature extraction, thereby enhancing classification accuracy. As detailed in Section 4.4, ShuffleNet V2 is a lightweight CNN architecture, it is known for its high computational efficiency and remarkable accuracy through channel splitting, pointwise group convolution, and channel shuffling mechanisms. However, despite its impressive efficiency, its shallow architecture limits its ability to capture complex hierarchical features compared to deeper networks, which are needed for distinguishing subtle facial expressions. The output features from ShuffleNet V2 can be represented as:

$$X = (X_1, X_2, \ldots, X_d). \tag{4.1}$$

where $X \in \mathbb{R}^{B \times d}$ denotes the extracted feature matrix, B is the batch size, and d as the feature dimension after global average pooling.

To complement these features and capture both local and global representations, we consider the EfficientViT-M2, a High-Speed ViT family. Unlike CNNs that process the input locally, Vision Transformers operate on non-overlapping image patches, as illustrated in 4.3. Additionally, as also discussed in Section 4.4, EfficientViT-M2 is effective in encoding both local and global information by combining efficient depthwise convolutions and attention mechanisms, enabling it to capture more complex features while utilizing limited computational resources. Its outputs can be represented as:

$$Y = (Y_1, Y_2, \ldots, Y_d). \tag{4.2}$$

Here, $Y \in \mathbb{R}^{B \times d}$ represents the transformer-based features extracted from the same input.

We perform a fusion of features to leverage the strengths of both architectures by concatenating the outputs from ShuffleNet V2 and EfficientViT-M2, forming a single feature vector. As a result, we obtain:

$$Z = (Z_1, Z_2, \ldots, Z_k, \ldots, Z_n) \tag{4.3}$$

which can be expressed as:

$$Z = X \oplus Y. \tag{4.4}$$

where $\oplus$ denotes the concatenation operation along the feature dimension.

This fused representation is then passed to the classification module to predict the driver's facial expression.

4.5.4 Classification

For the final expression recognition phase, we designed a custom classifier tailored explicitly for accurate and robust emotion prediction. Unlike standard classification heads, our custom classifier emphasizes optimized learning of subtle differences between emotion classes, particularly relevant in challenging driving scenarios. Specifically, our classifier consists of three fully connected layers (FC) that sequentially transform high-dimensional facial features into discriminative representations suitable for emotion classification. Between these linear transformations, we applied two batch normalization layers to stabilize the feature distributions, ensuring faster and more reliable training convergence. To effectively capture complex and nonlinear relationships within facial expressions, each batch normalization layer is followed by a Rectified Linear Unit (ReLU) activation function. Furthermore, to mitigate overfitting and enhance the classifier's generalization capability, two dropout layers with a dropout probability of 0.5 were strategically placed after the activation functions, randomly dropping connections and ensuring that the classifier does not overly rely on specific neurons. The output from the final fully connected layer produces logits corresponding directly to emotion classes.

4.6 Experimental Analysis

In this section, we present the experimental evaluation of our proposed method. First, we briefly describe how we used the datasets for experimentation. Next, we describe the experimental setup, including implementation details and training process. We then present the obtained results, provide a comparison with state-of-the-art methods, and conclude with an ablation study to assess the contributions of individual components within the model.

4.6.1 Experimental Data

To evaluate the effectiveness of our proposed method in driving scenarios, we first used the same dataset utilized in the previous chapter to assess our proposed approach, the KMU-FED dataset (see Section 2.3.1). For this dataset, to ensure fair comparison and consistency with existing studies, we applied different data splitting strategies, including 10-fold and 5-fold cross-validation, as well as an 80%-20% train-test split.

In addition to the KMU-FED dataset, we also used the Karolinska Directed Emotional Face (KDEF) dataset (see Section 2.3.1) to further evaluate the model's performance. For experimental consistency and comparability with prior work, we adopted an 80%-20% train-test split for this dataset as well.

4.6.2 Experimental Setup

We utilized the Grid Search technique [226] to determine the optimal hyperparameters for both datasets. Grid Search is an exhaustive search strategy that systematically evaluates all predefined combinations of hyperparameter values to identify the configuration that yields the best model performance. This approach helps save time compared to manual tuning.

Based on the selected hyperparameters, for the KMU-FED dataset, we used a batch size of 128 and optimized the model using the Adam optimizer with a fixed learning rate of 0.001. The training was conducted over 90 epochs using the cross-entropy loss function. For the KDEF dataset, we utilized a smaller batch size of 32, a reduced learning rate of 0.0001, and extended the training duration to 400 epochs. Similar to the KMU-FED setup, the

Adam optimizer and cross-entropy loss were employed. The selected hyperparameters for both datasets are summarized in Table 4.1.

Table 4.1: Hyperparameter Settings for ShuffViT-DFER on the KMU-FED and KDEF datasets. "-" denotes not applicable.

Hyperparameters	**KMU-FED dataset**	**KDEF dataset**
Optimizer	Adam	Adam
Loss Function	Cross-Entropy	Cross-Entropy
Batch size	128	32
Learning rate	0.001	0.0001
Epochs	90	400
Folds	10 , 5	–

4.6.3 Results and Discussion

Comparison with State-of-the-Art

The results of our proposed method, along with a comparison to state-of-the-art and existing approaches, are summarized in Tables 4.2 and 4.3 across both datasets.

On the KMU-FED dataset, our proposed ShuffViT-DFER method achieved an average accuracy of 97.3% under the 10-fold cross-validation setting, which is comparable with (CCNN [69]) and outperforms other models such as Hierarchical WRF (94.7%) and SqueezeNet (83.4%). Under the 5-fold cross-validation, our method achieved 95.6%, surpassing all other reported methods, including LMRF (95.1%), FTDRF (93.6%), and d-RFs (91.2%). For the 80%-20% train-test split, our method achieved an excellent accuracy of 100%, significantly outperforming SqueezeNet (95.83%) and CNN+SVM (98.64%).

On the KDEF dataset, our proposed ShuffViT-DFER model achieved an accuracy of 92.44%, which notably surpasses all the other compared methods. Specifically, it outperformed EfficientNet-FER(88.17%), RBFNN(88.80%), and MPCNN(86.90%), as well as other efficient models such as SqueezeNet(86.86%) and Efficient-swishNet DCNN(85.50%). This performance margin of approximately 3.6% to 7% improvement over the best competing

Table 4.2: Comparison between the proposed method and state-of-the-art and existing methods on the KMU-FED dataset. "NAN" indicates that the result for the corresponding setting was not reported in the original paper.

Methods	**Accuracy (10 Fold)**	**Accuracy (5 Fold)**	**Accuracy (80%-20%)**
Hierarchical WRF [103]	94.7%	NAN	NAN
CCNN [69]	97.3%	NAN	NAN
LMRF [65]	NAN	95.1%	NAN
d-RFs [65]	NAN	91.2%	NAN
FTDRF [65]	NAN	93.6%	NAN
DF [65]	NAN	90.5%	NAN
SqueezeNet [207]	83.4%	82.7%	95.83%
CNN+SVM [197]	NAN	NAN	98.64%
ShuffViT-DFER (Our method)	**97.3%**	**95.6%**	**100.00%**

Table 4.3: Comparison between the proposed method and state-of-the-art and existing methods on the KDEF dataset.

Methods	**Accuracy**
MPCNN [227]	86.90%
SquezeeNet [207]	86.86%
RBFNN [228]	88.80%
Efficient-swishNet DCNN [229]	85.50%
EfficientNet_FER [230]	88.17%
ShuffViT-DFER (Our method)	**92.44%**

models highlights the strength of our architecture in capturing expressive features across varying head poses. This robustness stems from the effective integration of local and global representations through the dual-architecture design.

In summary, these results assess the validity of our proposed architecture when it comes to the recognition of facial expressions under realistic challenges.

Ablation Analysis

To better gain insight into the performance of our model in terms of speed efficiency and accuracy, we carried out an ablation analysis by comparing the performance of each individual module in our architecture before and after fusion. We considered several factors to ensure a fair and comprehensive comparison, including the number of parameters, the processing time per image, and classification metrics, such as accuracy, precision, recall, and the F1-score, which are presented in section 2.3.2. The results are presented in Table 4.4..

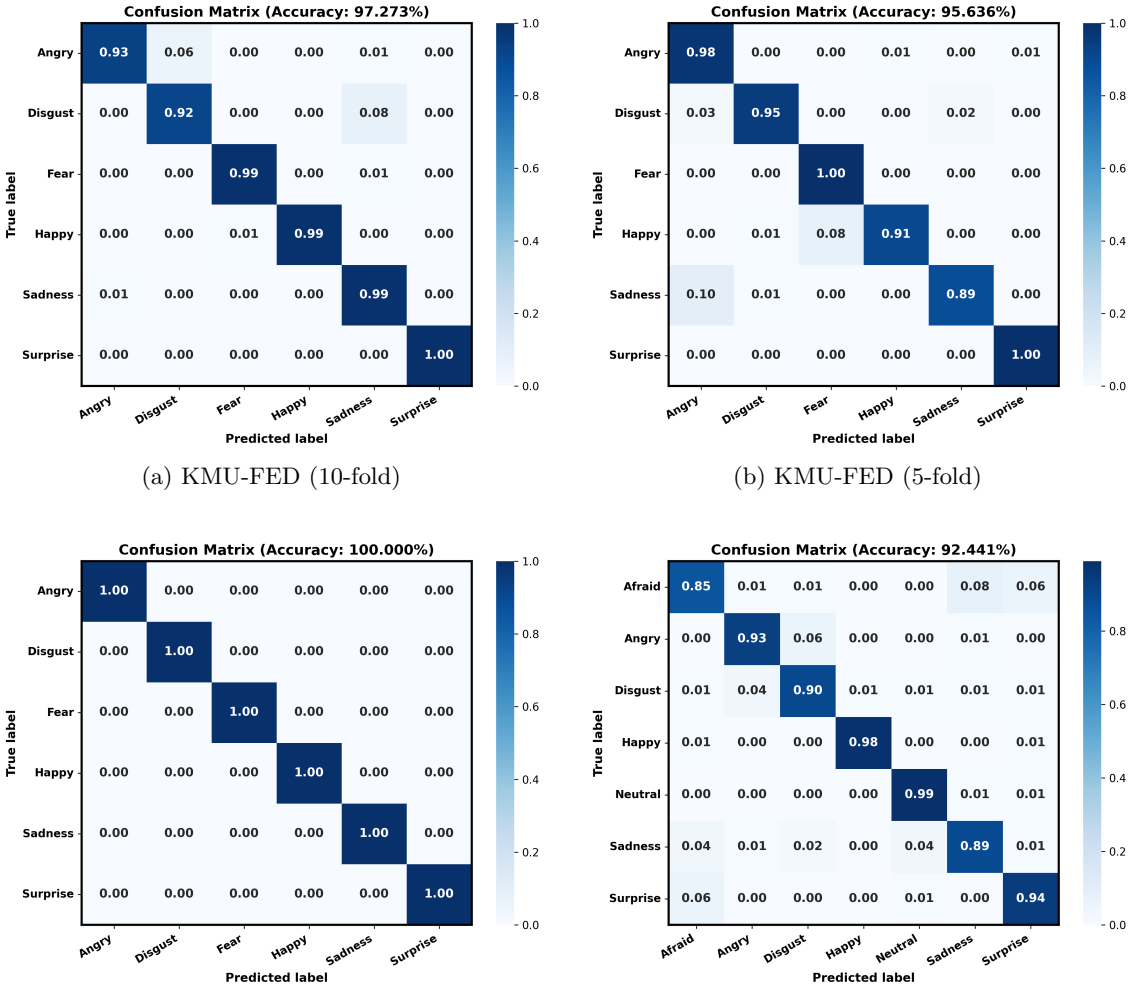
Table 4.4: Results of the ablation study on KMU-FED dataset. "w" denotes "with", "w/o" denotes without, and "C" denotes our designed classifier.

Methods	Params(M)	Processing Time(ms)	Accuracy	Precision	Recall	F1-score
ShuffleNet V2 w/o C	2.3M	**1.03 ms**	93.455%	0.94	0.93	0.93
ShuffleNet V2 w C	**2.0M**	1.11 ms	90.455%	0.90	0.91	0.90
EfficientViT-M2 w/o C	4.2M	2.80 ms	86.273%	0.86	0.86	0.86
EfficientViT-M2 w C	4.7M	2.70 ms	95.727%	0.96	0.96	0.96
ShuffViT-DFER (Ours)	5.9M	3.30 ms	**97.273%**	**0.97**	**0.97**	**0.97**

As expected, the proposed model has slightly more parameters and requires longer processing time than the individual modules, while still maintaining relatively low computational cost and providing the best performance overall. It achieves an average accuracy of 97% and a processing time of 3.3 ms per image (after face detection and cropping). ShuffleNet V2, whether used with its original classifier or the proposed classifier, has fewer parameters and is therefore more suitable for real-time applications, but this comes at the expense of reduced accuracy. EfficientViT-M2 offers improved performance compared to ShuffleNet V2, but it is still less accurate than our proposed architecture. These results demonstrate that our method effectively combines the strengths of ShuffleNet V2 and EfficientViT-M2, along with the proposed classifier head, to improve overall performance. Moreover, it achieves a strong balance between computational efficiency and classification accuracy

Visualization-Based Analysis

To further evaluate the recognition performance of our proposed model, the results in terms of confusion matrices on the KMU-FED and KDEF datasets are depicted in Figure 4.4, showing a detailed overview of the model's accuracy across different facial expressions.



(a) KMU-FED (10-fold)

(b) KMU-FED (5-fold)

(c) KMU-FED (80%–20%)

(d) KDEF

Figure 4.4: Confusion matrices of the proposed model for the KMU-FED and KDEF datasets.

In the case of the KMU-FED dataset, we conducted experiments with varying data splits to investigate the generalization ability of our model. The results showed good performance for most expression categories across all split settings. When using 10-fold cross-validation (Figure 4.4 (a)), the model performed very well at identifying expressions such as surprise, sadness, happiness, and fear. However, its performance decreased slightly when distinguishing between angry and disgust expressions, due to the visual similarities in their appearance. Under the 5-fold cross-validation scenario (Figure 4.4 (b)), the model also showed good performance, successfully achieving perfect classification for fear and

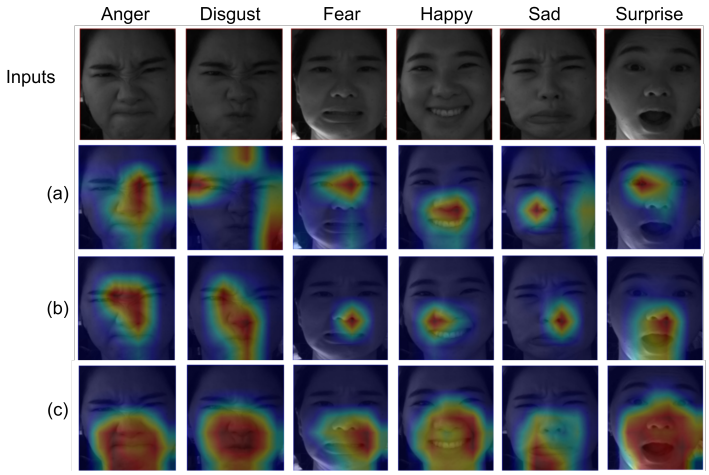


Figure 4.5: XGrad-CAM attention maps for different facial expressions, generated using (a) ShuffleNet V2, (b) EfficientViT-M2, and (c) the proposed ShuffViT-DFER model on the KMU-FED dataset.

surprise with no error rate. Moreover, (Figure 4.4 (c)), our model demonstrated excellent classification across all expression categories while using an 80%–20% split, reinforcing the effectiveness of our proposed model on the KMU-FED dataset.

For the KDEF dataset (Figure 4.4 (d)), our model also performed well, particularly in recognizing neutral and happy expressions. However, a slight decline in accuracy was observed for the afraid class, indicating that it can be challenging to distinguish afraid from visually similar expressions, such as sad or surprised.

To evaluate if our proposed method effectively learns the relevant facial features and to show its robustness, we applied the XGrad-CAM [231] visualization technique. This analysis was conducted on our model as well as two baseline methods, ShuffleNet V2 and EfficientViT-M2, using images of the same person exhibiting different facial expressions from the KMU-FED dataset. The resulting attention heatmaps are presented in Figure 4.5. These visualizations illustrate how each model attends to different facial regions when predicting specific expressions. Comparing the outcomes, it can be observed that ShuffleNet V2 for some emotions, such as happy and anger, the model focuses on key regions like the mouth and brow area, respectively. However, its attention is primarily focused on broader, less specific facial regions. EfficientViT-M2 shows more structured and localized attention compared to ShuffleNet V2. It frequently highlights the mouth and eyes, especially for disgust, anger, and surprise. However, its attention tends to be similar across some

expressions, such as fear and sadness, which may reduce its ability to clearly distinguish between them. In contrast, the proposed method, ShuffViT-DFER, demonstrates focused and consistent attention on the lower facial region, particularly around the mouth and chin, areas that play a key role in distinguishing expressions such as happiness, disgust, and surprise. While the attention maps do not vary drastically between emotions, this stable focus suggests that the model is effectively leveraging high-information regions to make accurate predictions. This behavior aligns with the model's strong classification performance, supporting its effectiveness in both accuracy and interpretability.

4.7 Conclusion

This chapter introduced ShuffViT-DFER, an efficient and fast method for driver facial expression recognition. The proposed approach leverages transfer learning by integrating two lightweight pre-trained models, ShuffleNet V2 and EfficientViT-M2, within a specialized classification scheme. Extensive experiments are conducted on two distinct and challenging datasets that simulate real-world challenges. The obtained results demonstrated the effectiveness of the proposed approach in improving accuracy while maintaining low processing time, which is essential in resource-constrained environments, such as driving environments.

Building on the insights gained in this Chapter and Chapter 3, which demonstrated the ability of deep learning-based models relying solely on visual input to accurately and efficiently recognize facial expressions. In the next chapter, we expand the investigation to multimodal facial expression recognition. Specifically, we explore how combining visual cues and textual information through Vision-Language Models (VLMs) can enhance the semantic understanding and classification of facial expressions in driving environment.

5 Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment

5.1 Introduction

This chapter expands our investigation from vision-only-based architectures to multimodal methods, specifically Vision-Language Models, to further enhance FER performance in driving environments. We are witnessing an increasing interest in VLMs as reflected in the impressive results across a large spectrum of tasks. In this context, we introduce in this chapter a novel architecture that exploits the capabilities of VLMs for facial expression recognition, particularly in the driving environment, to enhance road safety. We present an approach called CLIVP-FER, which uses the Contrastive Language-Image Pretraining (CLIP) model and combines both visual and textual data to overcome the environmental challenges and ambiguities in facial expression interpretation. The proposed approach is thoroughly evaluated on a benchmark driving dataset, KMU-FED. The experiments showed superior performance compared to state-of-the-art methods. Cross-database evaluation is also provided, showing good generalization abilities.

Section 5.2 outlines the motivation behind this contribution and reviews relevant work on facial expression recognition. Section 5.3 introduces the baseline model, CLIP, providing an overview of its architecture. Section 5.4 presents our approach in detail, covering each stage from data pre-processing to feature extraction using CLIP, followed by emotion classification. Section 5.5 describes the experimental datasets and setup, and reports the obtained results alongside comparisons with state-of-the-art methods. This section also includes a cross-database evaluation and an ablation study to further analyze the performance and robustness of the proposed architecture. Finally, section 5.6 summarizes the key findings.

5.2 Motivation and Related Work

Accurate and efficient facial expression recognition in driving contexts can significantly enhance road safety by enabling advanced driver assistance systems to respond appropriately to a driver's emotional state, as discussed in previous chapters. A variety of approaches have been proposed for driver FER, ranging from handcrafted feature-based techniques to deep learning methods [26, 28, 65, 191, 196]. These systems, including those developed in Chapters 3 and 4, primarily rely on visual input to detect facial cues and infer emotional states. However, purely visual systems are often limited by challenging real-world conditions such as variable lighting, occlusions, and head pose variations. Additionally, many existing models are computationally intensive, making them unsuitable for deployment in driving environments.

Recently, several studies have begun exploring the application of VLMs for facial expression recognition, showing promising results [23, 24, 232, 233]. For example, Li et al. [24] proposed a CLIP-based framework for dynamic and static facial expression recognition, incorporating fine-grained text descriptors for each expression. Similarly, Zhao and Patras [23] focused on dynamic facial expression recognition, incorporating temporal modeling and learnable context to capture fine-grained temporal features. Foteinopoulou and Patras [232] addressed zero-shot classification challenges in dynamic facial expression recognition by using sample-level text descriptions for natural language supervision. A quantitative assessment of GPT-4V's performance in General Emotion Recognition (GER) tasks is provided in [233].

Although these studies demonstrate the potential of VLMs in facial expression recognition, they have not yet been explored in the context of driving environments, which present distinct challenges such as constrained head movement, frequent occlusions, and abrupt lighting changes. Furthermore, these approaches often rely on detailed textual descriptors and computationally demanding architectures, which limit their feasibility for real-time deployment in embedded automotive systems.

Motivated by these recent developments and the previous limitations, we propose a novel architecture that leverages the capabilities of CLIP model [141] for driver FER. Our method combines both visual and textual modalities to address the environmental challenges and reduce the ambiguity commonly associated with facial expression interpretation in driving conditions. In addition, we apply average pooling to the features extracted by CLIP to

reduce the dimensionality and highlight the salient information, thereby reducing the computational cost and improving the robustness of the classifier. The contributions of this work are described as follows:

- We introduce CLIVP-FER, an innovative approach for extracting features from both images and text using the CLIP model, along with a MLP classifier for accurate classification. This method effectively leverages multimodal data to provide a detailed understanding of the driver's facial expressions.
- Our methodology harnesses the CLIP model solely as a feature extractor, without additional training. This strategy enables us to leverage the strengths of a pre-trained model, significantly reducing computational costs.
- We apply average pooling to the features extracted by CLIP to reduce the dimensionality of the features and highlight salient information, thereby improving the performance of the model.
- We conduct a comprehensive evaluation of the proposed approach, which includes assessing its generalization capabilities, speed efficiency, and performing an ablation study.
- We compare our approach to the state-of-the-art and existing methods on a benchmark dataset of a driving environment, resulting in significant performance improvements.

5.3 Contrastive Language-Image Pre-Training (CLIP): An overview

Contrastive Language-Image Pretraining, CLIP [141], is a multi-modal deep learning model developed by OpenAI in 2021 that learns to align visual and textual information within a shared embedding space. Trained on a large-scale dataset of image–text pairs using a contrastive learning objective (detailed in Section 2.4.2), to align relevant images with their corresponding textual descriptions while distinguishing unrelated pairs. As a result, both images and texts are projected into a shared high-dimensional embedding space, where similar inputs are located close to each other. Additionally, CLIP demonstrates strong zero-shot capabilities, allowing it to perform effectively on downstream visual tasks without requiring task-specific fine-tuning.

As depicted in Figure 5.1, CLIP is built on a dual-encoder architecture to map images and text into a shared embedding space. The model consists of two independently trained encoders: a vision encoder for processing visual input and a language encoder for textual input.

- **Image Encoder:** This component is responsible for extracting salient features from the input image. It converts the visual input into a high-dimensional embedding vector. The image encoder is typically implemented using architectures such as CNNs, like ResNet-based models (RN50, RN101, RN50x4, RN50x16, RN50x64) or Vision Transformers models (ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14@336px).
- **Text Encoder:** The text encoder processes the textual input, such as a class name or descriptive caption, and converts it into a corresponding high-dimensional vector. This encoder is based on a standard Transformer architecture.
- **Shared Embedding Space:** Both encoders project their respective inputs into a common embedding space, enabling CLIP to directly compare visual and textual representations and learn the semantic relationships between them.

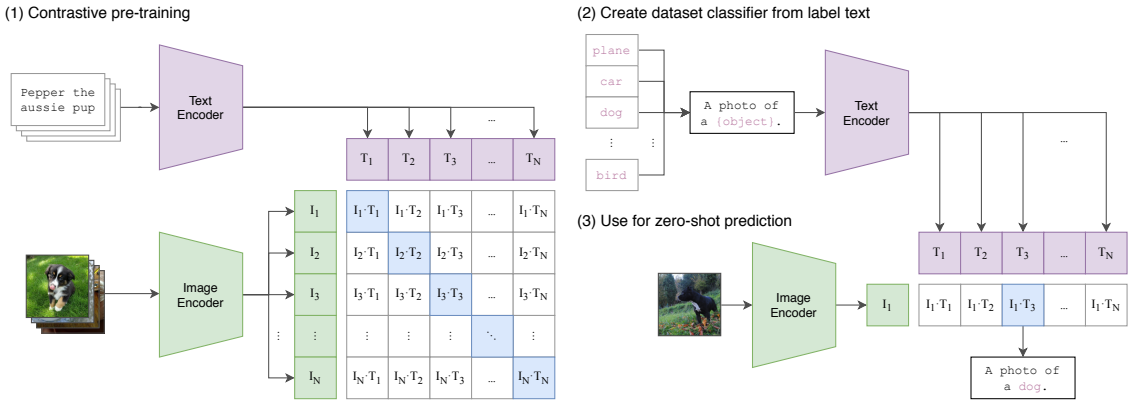


Figure 5.1: The Overall CLIP workflow [141].

The overall CLIP workflow, as illustrated in Figure 5.1, can be broken down into three main steps:

(1) Contrastive Pre-training: Involves training the model using a large-scale collection of 400 million image and text pairs from the internet. Each image is paired with a corresponding textual description, while additional mismatched pairs are also included. Throughout this

process, the model maximizes the similarity of correctly matched pairs while minimizing that of unrelated ones. This enables the model to develop a shared representation space that captures the relationship between visual and textual information.

(2) Dataset Classifier Creation from Label Text: For each image, a set of textual descriptions is generated, consisting of one semantically correct description and multiple incorrect alternatives. This setup produces both positive pairs (correct image–text matches) and negative pairs (mismatches). The textual inputs are then processed by the text encoder to generate class-specific embedding vectors. At this stage, the model is optimized using a contrastive loss function, which plays a key role in training. This loss penalizes incorrect matches between images and texts while rewarding correct associations within the shared embedding space. As a result, the model learns to capture the semantic similarity between visual and textual inputs more effectively.

(3) Zero-shot Prediction: After the pre-training step, CLIP leverages learned representations for inference using a zero-shot classification strategy. Particularly, the trained text encoder now functions as a zero-shot classifier. When a new image is provided, CLIP can generate predictions without any task-specific fine-tuning by processing the image through the image encoder and comparing it against predefined textual class descriptions via the dataset classifier. CLIP computes the cosine similarity between the image embedding and each of the text embeddings to determine the most relevant label. The model has been trained to maximize the similarity of correctly matched pairs while minimizing the similarity of mismatched ones. Through this process, CLIP constructs a multimodal embedding space where semantically aligned images and texts are positioned close together. The final prediction corresponds to the text label with the highest similarity (logit) score.

In the context of this chapter, we leverage only the dual-encoder architecture of CLIP, a vision encoder and a text encoder, for extracting high-level semantic features from visual and textual inputs, respectively. The following section presents a detailed description of the pipeline of our proposed method.

5.4 Proposed Method: CLIVP-FER

5.4.1 Overall Architecture

Our proposed architecture, called CLIVP-FER, is illustrated in Figure 5.2. The inputs of the system consist of a face image and a tokenized text descriptor. Feature extraction is independently conducted on each input utilizing CLIP encoders. Subsequently, a two-stage average pooling layer (AvgPool1d) is applied to the image features and to the concatenated image-text features in order to enhance the feature saliency and decrease the dimensionality. These refined features are then fed into a MLP classifier for determining the facial expression of the driver.

5.4.2 Data Pre-processing

For the visual data, and to enhance the model's robustness and prevent overfitting, which is particularly crucial for small FER datasets, facial images are detected and cropped using MTCNN detector [68], then a series of transformations is applied to the input images, including random horizontal flip and random rotation, hence introducing variability. Additionally, resizing and normalization are applied to the images.

The textual data in this study consists of descriptive captions for each expression class, providing detailed characterizations of the corresponding emotions. These captions were generated using the ChatGPT-4 language model, enabling more informative textual inputs rather than using standard class names. As illustrated in Table 5.1, each class is paired with a descriptive caption.

To generate these descriptions, we used the following prompt structure: *"What are the facial features of the {class name} emotion?"* For instance, for the "Happy" class, the prompt was: *"What are the facial features of the Happy emotion?"*. The generated response typically included detailed descriptions of facial movements involving key features such as the eyes, mouth, and cheeks. We then refined these responses by prepending the phrase "A facial expression with" to ensure consistency. For example, the final caption for the Happy class became: *"A facial expression with wide, bright eyes, raised cheeks, a broad smile showing teeth, and relaxed eyebrows."*. This strategy, inspired by the method

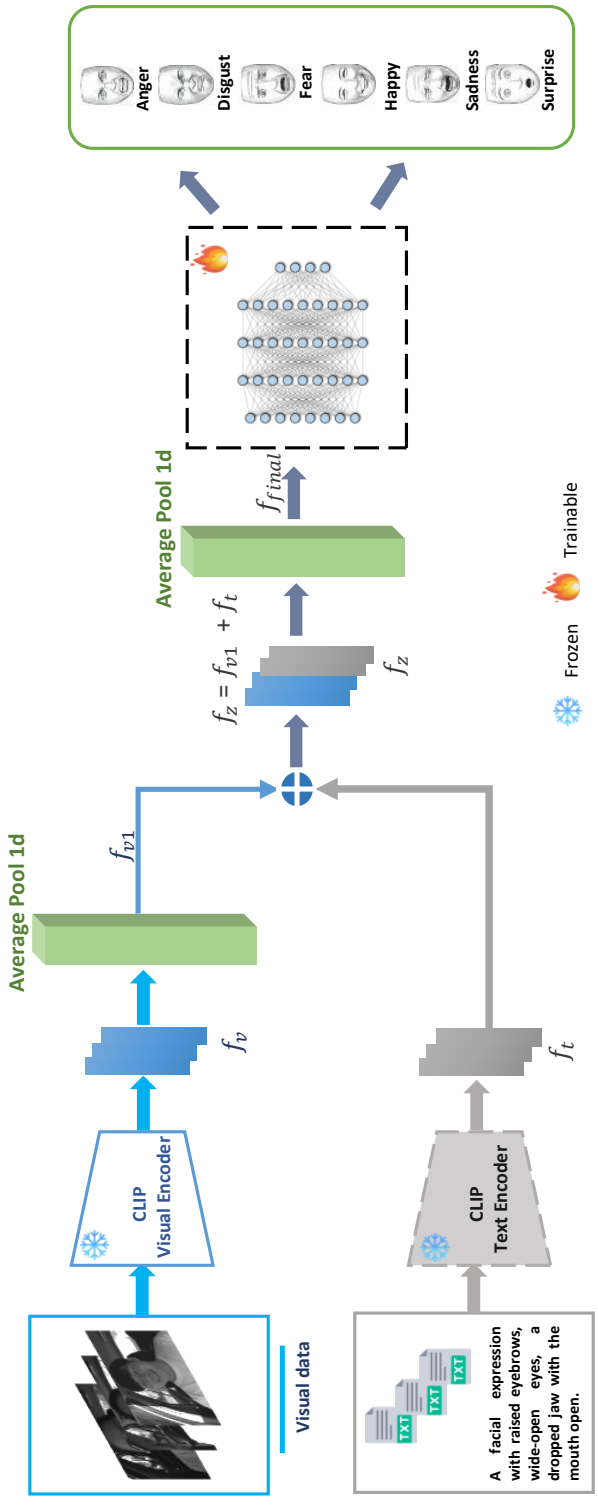


Figure 5.2: The figure illustrates our proposed CLIVP-FER architecture. The visual inputs are processed with Contrastive Language-Image Pretraining(CLIP) encoder into a vector f_v. The textual inputs are processed into a vector f_t. The visual feature vector f_v is then processed through an average pooling layer to produce f_{v1}, which is subsequently combined with f_t to form f_z. This combined vector is further processed by an additional average pooling layer, resulting in the final feature vector f_{final}, used to classify the facial expressions by a Multilayer Perceptron (MLP). The figure highlights the frozen (non-trainable) and trainable components.

proposed in [23], allows for a more nuanced and semantically rich representation of facial expressions.

The preprocessing of these captions involves tokenization and embedding using CLIP, converting the tokens into feature vectors that can be processed alongside the visual data.

Table 5.1: Facial Expressions and Their Associated Descriptions Generated by ChatGPT-4.

Facial Basic Expressions	**ChatGPT-4 Based Descriptions**
Anger	A facial expression with lowered and furrowed eyebrows, narrowed and glaring eyes, flared nostrils, a pressed mouth, and a tensed jaw.
Disgust	A facial expression with a wrinkled nose, raised upper lip, narrowed eyes, and a curled lip.
Fear	A facial expression with wide-open eyes, raised and drawn together eyebrows, a tensed open mouth, and a stretched face.
Happy	A facial expression with wide, bright eyes, raised cheeks, a broad smile showing teeth, and relaxed eyebrows.
Sadness	A facial expression with downward pointing corners of the mouth, drooping eyelids, slightly furrowed brows, and a generally downwards and subdued look.
Surprise	A facial expression with raised eyebrows, wide-open eyes, a dropped jaw with the mouth open.

5.4.3 Extracting and Refining Features Using CLIP and Average Pooling

As described in Section 5.3, CLIP has been designed with two distinct encoders: a visual encoder, E_{visual}, and a text encoder, E_{text}, which operate independently to handle their respective data modalities. For our purposes, we select the pre-trained ViT-B/32 CLIP variant as the visual encoder, which employs a ViT architecture. We use the standard Transformer from the CLIP model as the text encoder. The ViT-B/32 is particularly well-suited for this task due to its balance between expressive power and computational efficiency. Compared to other larger ViT variants, it maintains a lightweight structure while still providing rich, contextual feature representations for both text and image inputs. This makes it ideal for facial expression recognition in driving environments, where input data may be partially occluded, blurred, or captured under suboptimal lighting conditions.

The ability of ViT-B/32 to model global dependencies allows it to capture subtle emotional cues even when local facial details are unclear. Additionally, ViT-B/32 processes inputs with a stride of 32, which reduces spatial resolution but significantly improves inference speed and reduces memory requirements [131, 141].

During feature extraction, CLIP encoders process the preprocessed image and text inputs. The image is first divided into fixed-size patches and passed through a visual encoder to produce visual embeddings, while the text input is tokenized and passed through the text encoder to generate a text embedding. This process is performed without any additional training, which contributes to model simplicity and computational efficiency.

Let V denote the preprocessed image, and T the preprocessed text. The feature extraction is defined as follows:

$$f_{\mathrm{v}} = E_{\mathrm{visual}}(V) \tag{5.1}$$

$$f_{\mathrm{t}} = E_{\mathrm{text}}(T) \tag{5.2}$$

where f_{v} and f_{t} are the image and text feature vectors, respectively.

The resulting image feature vector undergoes average pooling, which condenses information, reduces dimensionality, and discards noise:

$$f_{\mathrm{v1}} = \mathrm{AvgPool1d}(f_{\mathrm{v}}) \tag{5.3}$$

We apply average pooling directly to the high-level features extracted by CLIP's Transformer-based visual encoder. Unlike pooling in CNNs, which typically operates on spatial feature maps, this approach takes advantage of the global contextual representations learned by the Vision Transformer. Subsequently, a second average pooling operation is performed on the combined image-text features set. This two-stage pooling process ensures focusing on the most relevant features for classification:

$$f_{\mathrm{z}} = \mathrm{Concat}(f_{\mathrm{v1}}, f_{\mathrm{t}}) \tag{5.4}$$

$$f_{\mathrm{final}} = \mathrm{AvgPool1d}(f_{\mathrm{z}}) \tag{5.5}$$

Here, f_{v1} represents the image features after the first average pooling operation, f_{t} denotes the text feature vector, f_{z} refers to the concatenated image-text feature representation,

and f_{final} is the final pooled representation used for training the classifier.

Finally, the condensed feature set f_{final} is used to train our classifier.

5.4.4 Emotion Classification

We utilize a specialized MLP classifier with a hidden layer that is specifically designed for categorizing emotions. This classifier is trained to accurately interpret the extracted set of final features, f_{final}, by CLIP. The structure of the classifier is as follows: The initial fully connected layer, known as `fc1`, transforms the input vector into a hidden representation that consists of 512 dimensions. This transformation is achieved using the equation:

$$h = \text{ReLU}(W_1 f_{\text{final}} + b_1) \tag{5.6}$$

Here, the weights and bias of `fc1` are denoted as W_1 and b_1, respectively. The Rectified Linear Unit activation function, ReLU, is applied in this process. To enhance generalization and prevent overfitting, a dropout layer is implemented with a rate of 0.5 on this hidden representation. The final layer, referred to as `fc2`, maps the hidden representation to the output space that corresponds to the number of emotion categories. This mapping is achieved using the equation:

$$o = W_2 h + b_2 \tag{5.7}$$

In this equation, W_2 and b_2 represent the weights and bias of `fc2`, respectively. The output denoted as o, represents the raw scores for each emotion category.

5.5 Experimental Analysis

In this section, we provide the experimental evaluation of the proposed method. We begin by briefly describing the datasets used and outlining how they were utilized in the experiments. This is followed by details of the experimental setup. We then present the results, compare our approach with state-of-the-art methods, and conduct a cross-dataset evaluation to assess the model's generalization capability. Finally, we conclude with an ablation study to examine the individual contributions of each model component.

5.5.1 Experimental Data

We evaluate the proposed architecture using the KMU-FED dataset [103], a publicly available benchmark specifically collected in a driving environment. This is the same dataset used in previous chapters, selected due to the limited availability of driver-specific facial expression datasets. For further details, see Section 2.3.1. To ensure consistent evaluation, we used a 10-fold cross-validation strategy, consistent with prior studies, as KMU-FED does not provide an official train-test split. To assess the generalization capability of our approach, we also conduct cross-database experiments, in which the model is trained on one dataset and evaluated on unseen data from a different dataset. For this purpose, we consider two widely used, in-the-wild FER datasets: FER-2013 [97] and RAF-DB [96] (see Section 2.3.1 for dataset details).

For FER-2013, we followed the standard predefined data split: 28,709 images for training, 3,589 for validation, and 3,589 for testing. Evaluation results are reported on the test set. For RAF-DB, we use the basic emotion subset, which includes 15,339 images, split into 12,271 for training and 3,068 for testing. As with FER-2013, we report results on the test set of the dataset.

5.5.2 Experimental Setup

We implemented our approach using the open-source PyTorch framework on an NVIDIA Quadro RTX 5000 GPU with 16GB RAM. During the training phase on the KMU-FED dataset, we set a batch size of 64, a learning rate of 0.003, and a weight decay of 1e-4. We employed the Adam optimizer, cross-entropy loss function, and trained for 40 epochs. Early stopping was implemented when no accuracy improvement was obtained after 10 epochs. Table 5.2 provides a summary of the implementation details.

5.5.3 Results and Discussion

Obtained Results

Table 5.3 presents the results obtained using our proposed CLIVP-FER model, alongside a comparison against some recent and state-of-the-art methods on the KMU-FED dataset. The table clearly indicates that our approach yields an impressive performance of 97.36 %,

Table 5.2: Hyperparameter Settings for CLIVP-FER on the KMU-FED dataset.

Hyperparameters	**KMU-FED dataset**
Optimizer	Adam
Loss Function	Cross-Entropy
Batch size	64
Learning rate	0.003
Weight decay	1e-4
Epochs	40
Folds	10

outperforming all other methods. Specifically, it shows a gain of 2.66 % over the hierarchical WRF method, 2.26 % over the LMRF method, and 3.09% over the pre-trained VGG16 method. This improvement is largely attributed to the fusion of CLIP's pretrained visual and textual features, which introduces additional semantic context into the representation space. By combining image features with text-derived embeddings of emotion categories, the model benefits from complementary information that enhances its ability to distinguish between facial expressions, particularly in cases where visual cues alone may be ambiguous or limited.

Table 5.3: Obtained results using our CLIVP-FER model, compared with state-of-the-art methods on the KMU-FED dataset.

Methods	**Accuracy**
Hierarchical WRF (2018) [103]	94.70%
LMRF (2020) [65]	95.10%
Pre-trained-VGG16 (2021) [217]	94.27%
Modified SqueezeNet (2022) [207]	83.40%
CLIVP-FER (Our method)	**97.36%**

The confusion matrix in Figure 5.3 confirms the strong performance of our model, with particularly high classification accuracy for expressions such as Happy and Surprise. These

results highlight the ability of the model to clearly distinguish highly expressive and visually distinct emotions. Some confusion remains between Angry and Sadness, as well as between Fear and Sadness, which can be attributed to the subtle visual overlap and shared facial cues in these expressions.

This is largely attributed to the complementary role of the text encoder, which enriches the feature space with semantic guidance derived from emotion labels. By fusing text and visual embeddings, the model benefits from both contextual meaning and visual cues. However, minor confusion persists between categories such as fear, sadness, and anger, which are not only visually similar but also semantically adjacent. This suggests that while the inclusion of text features improves robustness, distinguishing between subtle emotional states with overlapping semantics remains a challenge. Nevertheless, the overall results underscore the effectiveness of the CLIVP-FER model in interpreting a wide range of complex expressions, making it well-suited for practical applications.

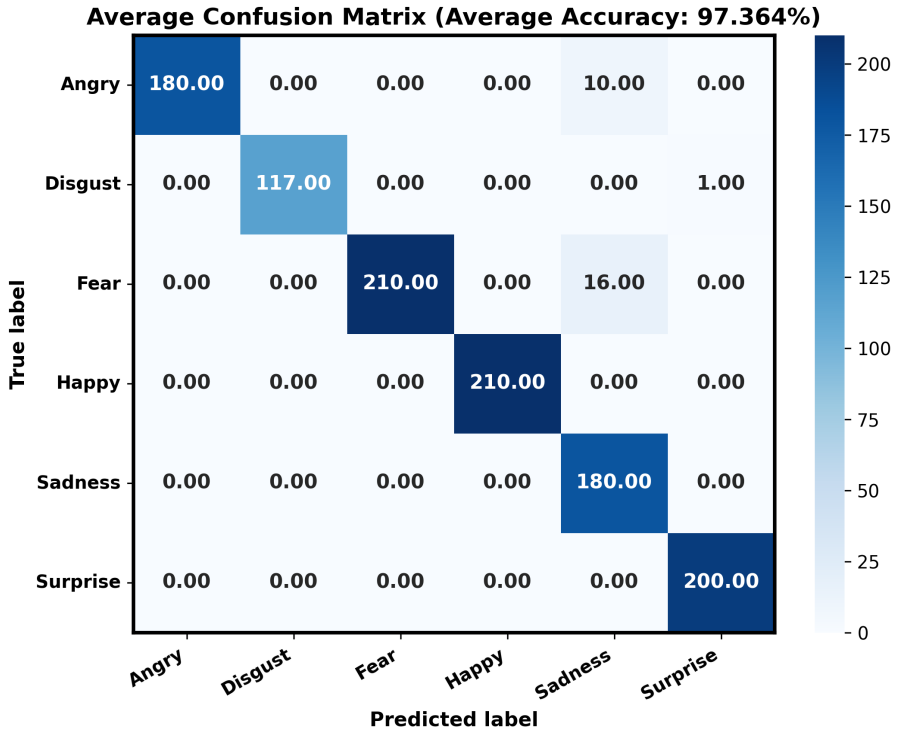


Figure 5.3: Confusion matrix of our CLIVP-FER model on the KMU-FED dataset.

Cross-Database Evaluation

In order to thoroughly investigate the generalisability and efficiency of our proposed CLIVP-FER approach, we performed a cross-database analysis by training our model on the KMU-FED dataset (captured in a driving context) and testing it on two widely used datasets collected in non-driving environments.

As reported in Table 5.4, various performance metrics, including accuracy, precision, recall, and F1 score, are used to measure the model's ability to correctly identify facial expressions across datasets. Precision and recall, although closely related to accuracy, provide more detailed information about the performance of the model. Precision indicates how many of the positive outcomes predicted by the model were actually positive, and recall indicates how many of the actual positive outcomes were correctly predicted by the model. The F1 score, a combination of precision and recall, indicates a well-balanced model that is both accurate and sensitive.

Table 5.4: Cross-database performance evaluation of our proposed CLIVP-FER model. KMU-FED dataset corresponds to a driving environment, while FER2013 and RAF-DB datasets correspond to a non-driving environment. The training is conducted on KMU-FED, and the testing is on FER2013 and RAF-DB datasets.

Datasets	KMU-FED	
	FER2013 [97]	RAF-DB [96]
Accuracy	0.89	0.92
Precision	0.92	0.85
Recall	0.87	0.87
F1-score	0.88	0.85

When trained on the KMU-FED dataset, the CLIVP-FER model achieved an accuracy of 0.89 on FER2013 and 0.92 on RAF-DB, demonstrating strong performance even when evaluated on unseen data from different environments. For a fair comparison, given the fact that our model was trained on 6 classes, we tested the model on the same number of classes, excluding the seventh class, 'Neutral', from both test sets. The model is still shown to be quite accurate, with a slight drop in performance when tested on unbalanced samples, which is due to the balanced nature of its training set. This decrease in accuracy can also be attributed to the fact that certain features present in the FER2013 and RAF-DB

datasets are somehow different from those in the driving environment, such as those found in the KMU-FED dataset.

To evaluate the speed efficiency of our model, we recorded the inference time across the two datasets, We achieved an average inference time of 2.6 ms per image on the FER2013 dataset and 2.7 ms on the RAF-DB dataset. These results not only demonstrate the robustness of our model across the two emotional datasets but also highlight its rapid inference capabilities that reflect the lightweight nature of its architecture, which are essential for real-time applications.

In summary, our CLIVP-FER model demonstrated strong cross-dataset generalization and efficient inference time, even though it is trained exclusively on a dataset captured in a driving environment. Its ability to maintain high classification performance across different data distributions and achieve rapid inference time highlights its robustness and potential for practical applications, particularly in driving contexts where accurate facial expression recognition is crucial.

Ablation Study

In order to assess the contribution of individual components and better understand the impact of different image- and text-derived features on the overall performance of our CLIVP-FER model, an ablation study is conducted. The results of this ablation study are detailed in Table 5.5.

Table 5.5: Evaluating the impact of CLIVP-FER model's components on the KMU-FED dataset. The first row shows the results of using only image features. The second row shows the results of combining image and text features. The last row demonstrates the performance of our model utilizing both visual and textual features, along with an average pooling layer.

Methods	**Accuracy**	**Precision**	**Recall**	**F1-score**
Visual	0.63	0.62	0.60	0.62
Visual+Text	0.84	0.88	0.81	0.81
Visual+Text+AvgPool	**0.97**	**0.98**	**0.97**	**0.97**

When using only the image features extracted from the CLIP vision encoder, the model yields relatively low performance, achieving an average accuracy of 0.63, a precision of 0.62, a recall of 0.60, and an F1 score of 0.62. This indicates that although visual features provide useful information, they are not sufficient to capture the nuanced variations necessary for accurate FER in complex scenarios.

Incorporating textual features in addition to the image features significantly improves the model's performance, demonstrating the effectiveness of the proposed multi-modal architecture. With the fusion of image and text representations, the model achieves a better average accuracy of 0.84. The gains in precision (0.88) and recall (0.81) further highlight the model's refined ability to accurately classify expressions and reliably identify the majority of classes. This demonstrates that semantic descriptions derived from text provide complementary information to the visual features, thereby enriching the model's contextual understanding of expressions.

Finally, the full CLIVP-FER model, leveraging the concatenation of both image and text features combined with an average pooling layer, achieves an excellent performance, with an accuracy of 0.97, a precision of 0.98, a recall of 0.97, and an F1 score of 0.97. These findings clearly highlight the strength of multimodal feature integration and demonstrate the importance of each component of the system.

5.6 Conclusion

In this chapter, we introduced CLIVP-FER, a novel approach that utilizes the CLIP vision-language model to recognize facial expressions in driving scenarios. Our approach concatenates the features extracted by a pre-trained CLIP variant from images and text descriptions, leading to a more comprehensive representation of facial expression recognition. The feature set is further refined by applying average pooling, enabling the classifier to train with more focus and pertinent information while minimizing the dimensionality of the features. We assessed the performance of our approach on a benchmark dataset of a driving environment. We conducted a cross-database evaluation, demonstrating the good generalization ability and inference efficiency of our approach. The experimental results showed significant performance enhancement, highlighting the effectiveness of our approach in recognizing facial expressions in a driving environment.

Building on these results, the next chapter extends the application of CLIP to a broader setting beyond the driving environment. While the approach differs from the previous chapter, this shift is motivated by the need to address more dynamic, general-purpose FER tasks under realistic and challenging conditions. To this end, CLIP is adapted to work with video data using efficient fine-tuning strategies, aiming to improve recognition performance while maintaining efficiency.

6 A Parameter-Efficient Fine-Tuning of VLMs for Dynamic Facial Expression Recognition

6.1 Introduction

The emergence of Vision-Language Models (VLMs), such as CLIP, offers promising solutions to a wide range of vision tasks, as previously discussed. In this chapter, we extend the use of VLMs beyond the driving environment to address general facial expression recognition challenges, with a focus on dynamic, real-world settings where expressions better reflect natural human emotions. To achieve this, we propose PE-CLIP, a parameter-efficient fine-tuning (PEFT) framework that adapts CLIP for dynamic FER. The goal is to significantly reduce the number of trainable parameters while maintaining high accuracy. PE-CLIP improves both efficiency and accuracy by incorporating adapters and prompt tuning, two lightweight mechanisms that enable effective model adaptation with minimal computational cost. As a result, the proposed model achieves good performance while requiring fewer trainable parameters.

This chapter is organized as follows: Section 6.2 outlines the motivation and related work, including CLIP-based methods and parameter-efficient fine-tuning. Section 6.3 describes the architecture and key components of the proposed PE-CLIP framework. Section 6.4 details the experimental setup, datasets used, implementation, and evaluation results. It also includes comparisons with state-of-the-art methods and an ablation study to assess the role of each component. Finally, section 6.5 concludes the chapter.

6.2 Motivation and Related Work

As discussed in Chapter 2 (Section 2.5.1), early research in FER primarily focused on static FER (SFER), which analyzes individual images to classify facial expressions [11, 234].

These methods, often applied to controlled or in-the-wild image datasets [235, 236], lack the ability to capture the temporal dynamics of facial movements, an important aspect of emotional expression. In contrast, dynamic FER (DFER) leverages video sequences to model the progression of expressions over time, offering a more natural and robust understanding of emotional cues. The availability of in-the-wild DFER datasets such as DFEW [102], FERV39K [101], and MAFW [105] has significantly advanced this field. However, these datasets introduce real-world challenges, including motion blur, occlusion, pose variation, and low resolution, which make accurate expression recognition more difficult.

To address these challenges, recent studies have begun exploring VLMs for FER due to their strong generalization capabilities and ability to learn rich multimodal representations. Models such as CLIP [141] (see Section 5.3) have shown promising results in understanding facial expressions by aligning visual and textual modalities. For example, Zhao et al. [23] introduced DFER-CLIP and extended CLIP by incorporating a transformer-based temporal module, which improved the temporal modeling capabilities. Similarly, Li et al. [184] introduced a two-stage training paradigm to adapt CLIP for DFER. Foteinopoulou et al. [185] improved the generalization capabilities of FER models by incorporating sample-level text descriptions to capture a broader range of emotions beyond the traditional basic emotions. Recently, Zhou et al. [186] proposed a novel visual language model that captures and aggregates dynamic features of expressions in semantic supervision via Inter-Frame Interaction Transformer (Inter-FIT) and Multi-Scale Temporal Aggregation (MSTA). However, despite this progress, several limitations remain. (1) Most existing methods rely on full fine-tuning of the image encoder, which results in high computational costs and resource-intensive training, making them less practical for deployment in resource-constrained environments. Furthermore, (2) temporal modeling is often handled through simplistic techniques such as average pooling or shallow transformer layers, which are insufficient for capturing the complex short- and long-term dependencies necessary for accurate dynamic FER. Additionally, (3) while textual prompts provide valuable semantic context, existing methods often fail to fully exploit enriched and task-specific textual descriptions tailored to dynamic facial sequences. This limitation reduces the alignment between visual and textual modalities, reducing overall model effectiveness.

To mitigate the high computational cost associated with fine-tuning large pre-trained models, recent works have explored parameter-efficient fine-tuning strategies [155, 157], which aim to lower computational cost by updating only a small portion of the model's parameters while

maintaining competitive performance (See Section 2.4.3). These techniques are broadly classified into two categories: adapter learning, which involves inserting lightweight modules into model layers for incremental fine-tuning [157, 237, 238], and prompt learning [239, 240, 241], which incorporates learnable tokens to adapt the input for specific tasks. For instance, to adapt static facial expression models to dynamic video-based recognition, Gowda et al. [242] introduced FE-Adapter, a single adapter leveraging dynamic dilated convolutions to improve video emotion recognition. Chen et al. [243] proposed S2D, which incorporates a learnable Temporal Modeling Adapter (TMA) based on multi-head self-attention, inserted between adjacent transformer blocks within a residual structure. However, (4) these approaches primarily focus on single-modality adaptation and employ simplistic temporal modeling techniques. They lack dedicated temporal adapters specifically designed for CLIP in Dynamic FER.

Building on the motivations and challenges discussed above, we propose PE-CLIP, a parameter-efficient framework for adapting CLIP to dynamic FER. PE-CLIP introduces lightweight components that improve temporal modeling and visual-textual alignment while reducing computational cost. Unlike previous approaches that rely on heavy architectures or limited temporal modeling, PE-CLIP integrates adapters and prompt learning to provide a more efficient and robust solution. It includes a Temporal Dynamic Adapter (TDA), a GRU-based module for capturing temporal dependencies, and a Shared Adapter (ShA) for consistent feature refinement across modalities. In addition, Multi-modal Prompt Learning (MaPLe) is employed to enhance semantic alignment through enriched textual descriptions. This design enables PE-CLIP to achieve strong performance with fewer trainable parameters, as shown in Figure 6.1. The main contributions of this work can be summarized as follows:

- We propose PE-CLIP, a new parameter-efficient framework designed for dynamic facial expression recognition. By leveraging lightweight adapter modules and multi-modal prompt learning. PE-CLIP enhances visual-textual alignment in CLIP while maintaining computational efficiency.
- We introduce a Temporal Dynamic Adapter (TDA) which is a GRU-based module with a dynamic scaling mechanism, enabling adaptive emphasis on significant temporal variations.
- We introduce a Shared Adapter (ShA), which is a vanilla adapter applied across both the textual and visual modalities to improve representations, providing consist-

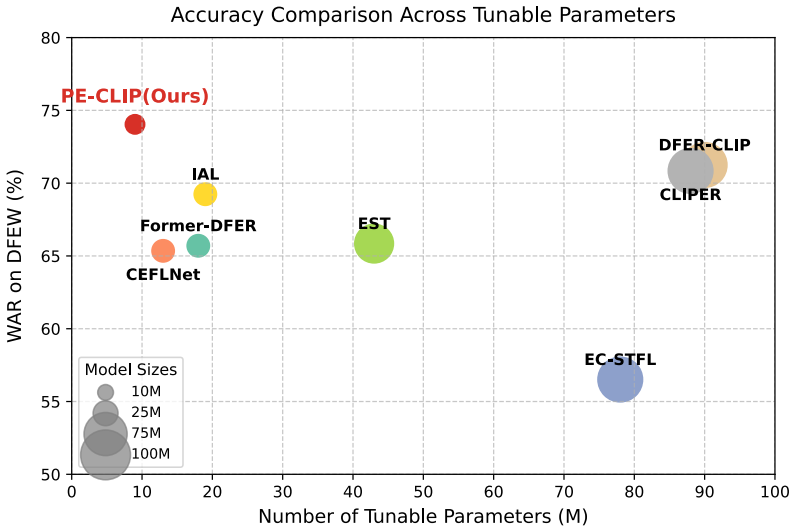


Figure 6.1: Performance comparison of dynamic facial expression recognition on the DFEW dataset. This chart highlights the trade-off between the number of tunable parameters (in millions) and model accuracy (measured as weighted average recall, WAR). The bubble size represents the model size, indicating the total number of trainable parameters. Our proposed framework, PE-CLIP (highlighted in red), achieves superior performance with significantly fewer tunable parameters (<6% of the whole model's parameters) compared to state-of-the-art methods. Models included in the comparison are IAL [244], CEFLNet [245], EC-STFL [102], Former-DFER [86], EST [199], DFER-CLIP [23], and CLIPER [184].

ent feature refinement while maintaining parameter efficiency within frozen CLIP encoders.

- We leverage Multi-modal prompt Learning (MaPLe) to adapt both textual and visual inputs. For textual input, we enrich prompts with Action Unit (AU) descriptions, providing semantically rich and contextually relevant prompts for expressions. This enhancement strengthens semantic alignment between textual and visual representations, ensuring more expressive feature adaptation for DFER.

- We conduct extensive experiments and comparison with state-of-the-art on benchmark datasets, validating the effectiveness of our proposed framework.

6.3 Proposed Framework

This section details the proposed PE-CLIP framework, which builds on the pre-trained CLIP model and adapts it for Dynamic FER using parameter-efficient fine-tuning. The core components of our methodology are described in detail. Section 6.3.1 provides an overview of the PE-CLIP framework. Section 6.3.2 focuses on visual feature extraction and temporal modeling. Section 6.3.3 explains textual prompt tuning, and section 6.3.4 presents classification via cosine similarity. The complete pipeline of the framework is illustrated in Figure 6.2.

6.3.1 Overview of PE-CLIP

As illustrated in Figure 6.2, the proposed PE-CLIP framework comprises three main components: visual feature extraction and temporal modeling, textual feature extraction, and classification. For the visual feature extraction part, a facial expression video sequence consisting of T frames is processed using the frozen CLIP visual encoder (ViT-B/16). The visual embeddings for each frame are refined by lightweight adapters, specifically the ShA, applied after the multi-head self-attention layers to efficiently refine the extracted features, and the TDA, applied after the MLP layers to model temporal dependencies across frames. In the textual feature extraction part, fine-grained prompts derived from facial action units based on the facial action coding system [246] are used to replace class names, providing richer semantic information as established in prior work [187]. These prompts are further refined by introducing learnable tokens via the MaPLe [247] method, enabling the model to better capture nuanced relationships between textual descriptions and dynamic facial expressions. The textual embeddings are then processed using the ShA adapter to refine extracted features consistently and the Textual Adapter (TA) to enhance the representation quality before alignment with the visual features. Finally, in the classification part, the refined visual and textual embeddings are projected into the shared embedding space, where cosine similarity determines the class label. Further details on these components are provided in the subsequent sections.

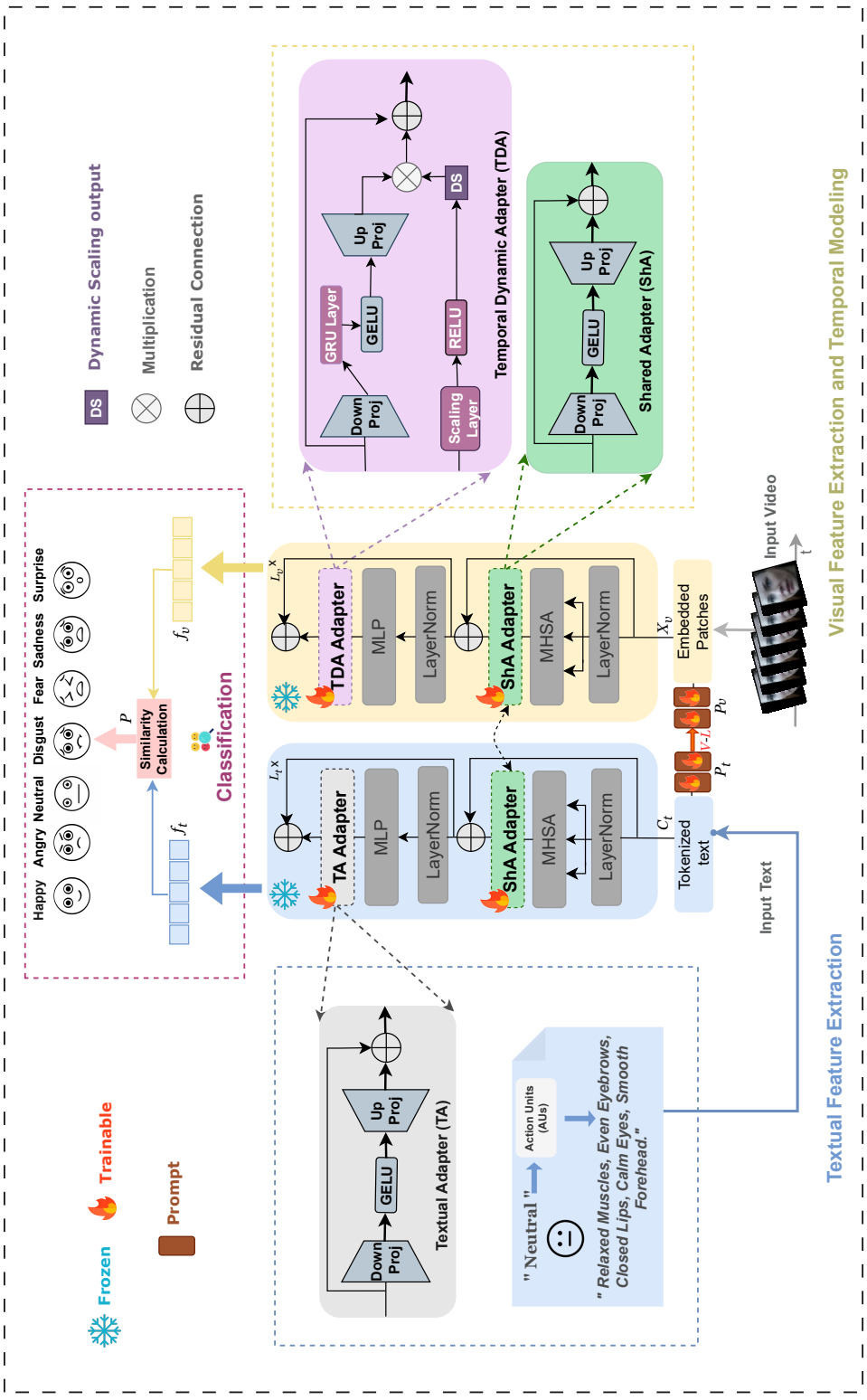


Figure 6.2: Overall architecture of the proposed PE-CLIP model. The model takes tokenized AU-based textual descriptions C_t and embedded image sequences X_t, enriched with MaPLe learnable tokens as inputs. These representations are processed through CLIP encoders, where the Shared Adapter (ShA) improves the representation learning across textual-visual modalities. Additionally, the Temporal Dynamic Adapter (TDA) with dynamic scaling, captures the key temporal dependencies in the vision branch, while the Textual Adapter (TA) enhances the textual representations. The resulting visual (f_v) and textual (f_t) embeddings are mapped into CLIP's shared space, where classification is performed via cosine similarity to associate expressions with their corresponding labels.

6.3.2 Visual Feature Extraction and Temporal Modeling

A key challenge in adapting the pre-trained CLIP model for DFER is achieving effective temporal modeling across video frames while ensuring parameter efficiency. Unlike static image recognition tasks or general video recognition, DFER requires modeling subtle short- and long-term dependencies in facial expressions. Although CLIP's pre-trained ViT encoder [248] excels in extracting spatial representations, it lacks mechanisms to directly model the temporal dynamics critical for DFER. To address this limitation, we extend the frozen CLIP visual encoder with lightweight, parameter-efficient adapters specifically designed to refine spatial features and capture temporal dependencies. The process begins with frame-level feature extraction from a facial expression video sequence $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$, where T is the temporal length, and $H \times W$ denotes the spatial resolution of each frame. Each frame $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$ is divided into non-overlapping patches of size $P \times P$, resulting in $M = (H \cdot W)/P^2$ patches per frame. These patches are flattened and projected into a d-dimensional latent space via a patch embedding layer, producing patch embeddings $\mathbf{X}v, i_{i=1}^{M} \in \mathbb{R}^d$. The learnable prompts $\mathbf{P}_v$ are prepended to these patch embeddings and then are processed within the frozen CLIP visual encoder through a series of transformer layers comprising the MHSA layer and MLPs. After processing, the class token is used to derive a frame-level feature embedding $\mathbf{v}_t \in \mathbb{R}^d$. Consequently, a sequence of frame-level embeddings $\mathbf{V} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T$ is obtained, where each $\mathbf{v}_t$ captures the spatial characteristics of the corresponding frame. These embeddings are then passed into lightweight adapters for further refinement, as described below:

Spatial Refinement with Shared Adapter (ShA)

To adapt frame-level features for DFER, we introduce the Shared Adapter, a lightweight bottleneck module applied after the MHSA layers of the CLIP visual encoder. The ShA refines spatial feature representations while maintaining computational efficiency through a reduction-expansion mechanism. Following the MHSA operation and layer normalization (LN), the ShA operates as follows:

$$\mathbf{H}_{\text{MHSA}} = \text{MHSA}(\text{LN}(\mathbf{H}_{\text{input}})). \tag{6.1}$$

The input feature $\mathbf{H}_{\text{MHSA}}$ is reduced using a down-projection matrix $\mathbf{W}_d \in \mathbb{R}^{d/r\times d}$, where r is the reduction factor:

$$\mathbf{H}_{\text{down}} = \mathbf{H}_{\text{MHSA}} \cdot \mathbf{W}_d^{\top}. \tag{6.2}$$

The reduced feature is passed through a GELU activation function to introduce non-linearity:

$$\mathbf{H}_{\text{act}} = \text{GELU}(\mathbf{H}_{\text{down}}). \tag{6.3}$$

The original dimensionality is restored using an up-projection matrix $\mathbf{W}u \in \mathbb{R}^{d \times d/r}$:

$$\mathbf{H}_{\text{up}} = \mathbf{H}_{\text{act}} \cdot \mathbf{W}_u^{\top}. \tag{6.4}$$

Finally, the adapter integrates a residual connection by adding its output back to the original input to the MHSA layer:

$$\mathbf{H}_{\text{adapt}} = \mathbf{H}_{\text{input}} + \mathbf{H}_{\text{up}}. \tag{6.5}$$

By refining spatial features through this adapter after the MHSA layers, the ShA enables task-specific adaptation without introducing significant computational overhead.

Temporal Modeling with Temporal Dynamic Adapter

Dynamic facial expressions evolve over time, necessitating robust modeling of temporal dependencies. Existing methods, specifically based on the CLIP model, frequently rely on transformer-based temporal models or simple average pooling layers. However, transformer-based approaches, though powerful, are computationally expensive, and average pooling layers often fail to capture subtle temporal dynamics, particularly in short sequences typical of DFER. To address these challenges, we propose the Temporal Dynamic Adapter, which combines a recurrent neural network, specifically, GRU [129] for temporal modeling with a learnable dynamic scaling mechanism that adaptively refines temporal feature representations based on their relevance. The TDA operates after the MLP layers of the visual encoder, At its core, the TDA employs a GRU due to its efficiency in modeling sequential dependencies while requiring fewer computational resources compared to transformer-based temporal models. This makes GRU particularly suitable for DFER, where video sequences often comprise 16 frames. The GRU processes a sequence of input features $\mathbf{H}_{\text{adapt}} \in \mathbb{R}^{T \times d}$,

where T is the number of frames and d is the feature dimension. To further optimize computational efficiency, the input features are first projected into a lower-dimensional space using a down-projection matrix $\mathbf{W}_d \in \mathbb{R}^{d/r \times d}$, where r is the reduction factor:

$$\mathbf{H}_{\text{down}} = \mathbf{H}_{\text{MLP}} \cdot \mathbf{W}_d^{\top}. \tag{6.6}$$

The GRU processes the reduced features $\mathbf{H}_{\text{down}} \in \mathbb{R}^{T \times d/r}$ and computes hidden states for each time step, enabling the model to capture temporal correlations across frames:

$$\mathbf{h}_t = \text{GRU}(\mathbf{H}_{t,\text{down}}, \mathbf{h}_{t-1}), \tag{6.7}$$

where $\mathbf{h}_{t-1}$ is the hidden state from the previous frame. This recurrent mechanism effectively models the subtle dynamics of facial expressions over time, leveraging sequential dependencies in video sequences.
To introduce non-linearity and enhance representational capacity, the GRU output $\mathbf{h}_t \in \mathbb{R}^{d/r}$ is passed through a GELU activation function:

$$\mathbf{h}_t = \text{GELU}(\mathbf{h}_t). \tag{6.8}$$

Following this, the activated features are restored to their original dimensionality using an up-projection matrix $\mathbf{W}_u \in \mathbb{R}^{d \times d/r}$:

$$\mathbf{H}_{t,\text{up}} = \mathbf{h}_t \cdot \mathbf{W}_u^{\top}. \tag{6.9}$$

By integrating this lightweight yet effective temporal adapter, the TDA strikes a balance between parameter efficiency and the ability to capture temporal dynamics.

Adaptive Emphasis with Dynamic Scaling

In adapter tuning for NLP and 2D vision tasks [157, 249], the scaling factor plays a pivotal role in determining the importance of features during adaptation. Traditionally, this factor is manually defined, which works well for static tasks but becomes less adequate

in dynamic settings like DFER. Video sequences in DFER exhibit varying levels of frame importance, some frames capture critical expression transitions, while others contain less relevant information. A fixed scaling mechanism fails to account for this variability, leading to suboptimal feature adaptation. To address this challenge, we propose a dynamic scaling mechanism inspired by [250], which enables adaptive weighting of frame's feature representation based on their contribution to temporal modeling. Instead of assigning uniform importance to all frames, this mechanism learns per-frame scaling factors through a learnable linear transformation, dynamically adjusting the contribution of each frame's feature representation. This ensures that informative temporal features are emphasized, while noisy or redundant ones are down-weighted, enhancing robustness and efficiency in temporal feature extraction.

The dynamic scaling factor is computed from MLP features before being applied after GRU processing, providing a frame-specific weighting mechanism. Specifically, a learnable linear layer generates scaling factors for each frame's feature representation:

$$\mathbf{W}_{\text{scale}} = \text{ReLU}(linear(\mathbf{H}_{\text{MLP}})) \tag{6.10}$$

where $\mathbf{H}_{t,\text{MLP}} \in \mathbb{R}^d$ represents the MLP-processed feature representation for frame t, and ReLU ensures non-negative scaling factors. These scaling factors are then applied post-GRU to adaptively focus on the most informative temporal variations as follows:

$$\mathbf{H}_{t,\text{scaled}} = \mathbf{H}_{t,\text{up}} \cdot \mathbf{W}_{\text{scale}}. \tag{6.11}$$

This adaptive scaling mechanism enhances the TDA's ability to focus on the most temporally informative frames.

6.3.3 Textual Feature Extraction and Alignment

Adapting pre-trained vision-language models such as CLIP for DFER requires constructing semantically rich textual prompts that effectively align with the fine-grained facial expressions, as demonstrated in [23, 184]. Existing methods often rely on simple class names or manually crafted prompts (e.g., "A photo of a {class}"), which lack the necessary granularity to distinguish subtle expression variations. While advanced prompt learning

approaches like CoOp [251], CoCoOp [252] introduce learnable tokens to optimize textual prompts, they primarily refine text representations while keeping the visual input unoptimized. This imbalance in adaptation limits their effectiveness for DFER, where subtle temporal and spatial cues contribute to emotion recognition. To address these challenges, we explore action unit descriptions as textual prompts, enriching semantic representations with detailed facial muscle movements. Following a similar approach to visual inputs, we incorporate MaPle-based textual prompts $\mathbf{P}_t$ into enriched textual embeddings. These are then processed through the frozen text encoder, where the ShA adapter refines feature representations, while the TA adapter further enhances the extracted features, ensuring robust and expressive text embeddings for DFER.

Table 6.1: Facial Expressions and Their Associated Action Units (AUs) Descriptions for FER [187].

Facial Basic Expressions	**AU-Based Descriptions**
Happiness	Cheek Raiser, Lip Corner Puller.
Sadness	Inner Brow Raiser, Brow Lowerer, Lip Corner Depressor.
Neutral	Relaxed Muscles, Even Eyebrows, Closed Lips, Calm Eyes, Smooth Forehead.
Anger	Brow Lowerer, Upper Lid Raiser, Lid Tightener, Lip Tightener.
Surprise	Inner Brow Raiser, Outer Brow Raiser, Upper Lid Raiser, Jaw Drop.
Disgust	Nose Wrinkler, Lip Corner Depressor, Lower Lip Depressor.
Fear	Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer, Upper Lid Raiser, Lid Tightener, Lip Stretcher, Jaw Drop.

Leveraging Action Units for Enriched Textual Prompts

Traditional textual prompts in DFER often lack semantic specificity, as they rely on abstract class names such as "Surprise" or "Fear." These simplistic prompts fail to capture detailed contextual information about the facial features and muscle movements that define each expression, thereby limiting their effectiveness in aligning with visual features. Inspired by works such as [81, 187], which emphasize the relationship between facial expressions and AUs defined in the facial action coding system, we incorporate AU-based descriptions into textual prompts to enrich their semantic content. AUs describe specific facial muscle

movements, providing a granular breakdown of facial dynamics. For example, as shown in Table 6.1, the expression "Surprise" is associated with features such as Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), and Jaw Drop (AU26). By leveraging these AU-based descriptions, we provide semantically rich prompts that better capture the nuances of facial expressions, allowing for stronger alignment with expression-relevant visual features. Formally, for a given expression class c_k, the textual prompt T_k is directly constructed using the fine-grained AUs descriptions associated with the class and represented as:

$$T_k = \{\text{AU}_1, \text{AU}_2, \ldots, \text{AU}_n\},$$

where AU_i denotes the textual description of the i-th Action Unit corresponding to the expression. These fine-grained prompts replace the abstract class names with semantically rich descriptions of facial muscle movements, enabling the model to better align visual features with textual information.

Enhancing Alignment with Multi-modal Prompt Learning (MaPLe)

To improve the alignment between visual and textual modalities, we integrate Multi-modal Prompt Learning, a framework designed to jointly adapt both image and text encoders in vision-language models [247]. Unlike single-modality prompt learning approaches, MaPLe introduces learnable tokens into both the textual and visual inputs, improving alignment. In MaPLe, the textual encoder processes AU-enriched textual prompts $\mathbf{T}_k$ (as detailed earlier) along with learnable tokens $\mathbf{P}_t \in \mathbb{R}^{N_t \times d}$, where N_t is the number of learnable tokens, and d is the embedding dimension. These learnable tokens, incorporated with the AU-Tokenized text to form $\mathbf{C}_t$, which are then passed to the textual encoder to produce the refined textual representation:

$$f_t = \text{TextEncoder}(\mathbf{C}_t), \tag{6.12}$$

where $\mathbf{C}_t$ are textual embeddings. Similarly, MaPLe adapts the visual encoder by introducing learnable tokens $\mathbf{P}_v \in \mathbb{R}^{N_v \times d}$, where N_v is the number of visual learnable tokens. These tokens are prepended to the visual embedding to form $\mathbf{X}_v$, allowing the vision encoder to refine its representations and better align with the textual embeddings. The

refined visual representation f_v is computed as:

$$f_v = \text{ImageEncoder}(\mathbf{X}_v), \tag{6.13}$$

To further enhance this alignment, MaPLe conditions the vision prompts on language prompts via a vision-language (V-L) coupling function, inducing mutual synergy between the two modalities.

6.3.4 Classification via Cosine Similarity

Leveraging the shared embedding space of CLIP, the refined visual representation f_v, which encapsulates the temporal features of the video sequence, is aligned with the refined textual embeddings f_t. The cosine similarity metric quantifies the alignment between these embeddings, ensuring effective matching of the temporal with the semantic information embedded in the textual prompts. This process can be formulated as:

$$P(k|\mathbf{f}_v) = \frac{\exp\left(\cos(\mathbf{f}_v, \mathbf{f}_{t_k})/\tau\right)}{\sum_{j=1}^{C} \exp\left(\cos(\mathbf{f}_v, \mathbf{f}_{t_j})/\tau\right)}, \tag{6.14}$$

where $\cos(\mathbf{f}_v, \mathbf{f}_{t_k})$ denotes the cosine similarity between the visual and textual embeddings, C is the total number of classes, and τ is a learnable temperature parameter.

Since the cosine similarity scores do not inherently represent probabilities, we optimize the model using a cross-entropy loss function, which internally applies a softmax operation to normalize the scores. This ensures that the model learns a probability distribution over the classes, making the classification process robust. Finally, the emotion class with the highest probability is selected as the predicted class.

6.4 Experimental Analysis

To assess the performance of PE-CLIP, we performed extensive experiments on two widely used benchmark datasets: DFEW [102] and FERV39k [101]. A brief overview of how these datasets were used in our experiments is provided in Section 6.4.1. Section 6.4.2 outlines the implementation setup. The results and comparison against state-of-the-art methods on both datasets are presented in Section 6.4.3. Furthermore, Section 6.4.4

includes ablation studies conducted on the DFEW dataset to analyze the contributions of each individual component of our framework. Finally, section 6.4.5 presents visualizations that help interpret the model's behavior.

6.4.1 Experimental Data

For the experimental evaluation, we used the two widely adopted dynamic FER datasets, DFEW and FERV39K, both collected in in-the-wild conditions (see Section 2.3.1 for more details). The DFEW dataset is already partitioned into five equal, non-overlapping subsets, thus, we followed a 5-fold cross-validation protocol to ensure consistency and fair evaluation. For FERV39K, we used the official split, where the dataset is divided into a training set of 31,088 clips (80%) and a testing set of 7,847 clips (20%), enabling fair comparisons.

6.4.2 Implementation Setup

Data Pre-processing: For both DFEW and FERV39K datasets, the preprocessed facial regions of video frames, as provided in the datasets, are directly utilized. Each video sequence is preprocessed by resizing and cropping 16 consecutive frames to a uniform size of 224×224 pixels. To mitigate overfitting and enhance generalization, various data augmentation techniques are applied, including random cropping, random horizontal flipping, and color jittering, as suggested in [243, 253, 254]. Additionally, to further increase the diversity of training data, we integrate Mixup [255] and FMix [256] techniques. During training, a random probability p determines the type of augmentation applied. If $p < 0.4$, Mixup is used, blending data and targets using a Beta distribution with a parameter $\alpha = 0.4$. If p is between 0.4 and 0.7, FMix generates binary masks based on the given α and decay power, combining the input data with shuffled samples to introduce new variations. For $p \geq 0.7$, no augmentation is applied. This dynamic augmentation strategy balances the original and augmented samples, enhancing the model's robustness to variations in training data and improving performance.

Training Setting: All experiments for the proposed PE-CLIP framework are conducted using the CLIP model with a ViT-B/16 backbone, implemented in PyTorch. Training is performed on two NVIDIA Quadro RTX 5000 GPUs, each with 16 GB of memory. To ensure parameter efficiency, only the adapter and prompt learner parameters, totaling 9M

trainable parameters, are fine-tuned, while the visual and text encoders remain frozen. The adapters employ a reduction factor r of 8 to achieve parameter efficiency without sacrificing performance. For the DFEW dataset, training is performed for 30 epochs in an end-to-end manner. The AdamW [257] optimizer is used with an initial learning rate of 2×10^{-5}, adjusted via a cosine learning rate scheduler with a warmup period of 3 epochs. Separate learning rates are set for different components: 2×10^{-4} for the adapters and 3×10^{-5} for the prompt tokens. The weight decay is set to 0.01, and the batch size is set to 8. For the FERV39K dataset, the model is trained for 20 epochs using the AdamW optimizer and cosine learning rate scheduler with a warmup period of 2 epochs. The initial learning rate is 4×10^{-5}, while the weight decay and batch size remain consistent with the DFEW settings. These configurations used in our experiments are illustrated in Table 6.2.

Table 6.2: Hyperparameter Configurations for Training PE-CLIP on DFEW and FERV39K Datasets.

Hyperparameters	**DFEW**	**FERV39K**
Batch size	8	8
Learning rate (adapters)	2×10^{-4}	2×10^{-4}
Learning rate (prompts)	3×10^{-5}	3×10^{-5}
Weight decay	0.01	0.01
Optimizer	AdamW	AdamW
Scheduler	CosineLR	CosineLR
Warmup	3	2
Reduction factor	8	8
Training epochs	30	20

6.4.3 Experimental Results

We evaluated the effectiveness of the proposed PE-CLIP method by comparing it with state-of-the-art approaches on two benchmark in-the-wild datasets, DFEW and FERV39K, using two key metrics: Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) (see Section 2.3.2).

Results on DFEW: Table 6.3 presents the average performance of PE-CLIP across 5-fold cross-validation on the DFEW dataset. Our proposed method achieves a UAR of 62.82% and a WAR of 74.04%, demonstrating its effectiveness against state-of-the-art methods. Notably, PE-CLIP outperforms vision-only approaches such as M3DFEL[258]

and Dual-STI[259], and achieves an improvement of 6.16% in UAR and 4.67% in WAR over AEN [244]. This highlights its ability to extract richer, more discriminative features for dynamic facial expression modeling. When compared to CLIP-based methods, PE-CLIP consistently surpasses their performance, with improvements of 3.21% in UAR and 2.79% in WAR over DFER-CLIP. Furthermore, PE-CLIP outperforms the recently proposed CLIP-Guided-DFER [186], achieving a UAR gain of 1.97% and a WAR gain of 1.46%, underscoring its ability to better align the vision and language representations through the adapters and prompt learning.

Table 6.3: Comparison of PE-CLIP with state-of-the-art methods on the DFEW and FERV39K datasets. The table highlights the Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) achieved by each method, alongside their respective backbones and the number of tunable parameters (in millions). "-" denotes results were not reported in the original paper.

Method	**Backbone**	**Tunable Param (M)**	**DFEW**		**FERV39k**	
			UAR	**WAR**	**UAR**	**WAR**
EC-STFL (MM'20) [102]	C3D / P3D	78	45.35	56.51	-	-
Former-DFER (MM'21) [86]	Transformer	18	53.69	65.70	37.20	46.85
CEFLNet (IS'22) [245]	ResNet-18	13	51.14	65.35	-	-
NR-DFERNet (ArXiv'22) [260]	CNN-Transformer	-	54.21	68.19	33.99	45.97
STT (ArXiv'22) [261]	ResNet-18	-	54.58	66.65	37.76	48.11
DPCNet (MM'22) [262]	ResNet-50	-	57.11	66.32	-	-
EST (PR'23) [199]	ResNet-18	43	53.94	65.85	-	-
Freq-HD (MM'23) [263]	VGG13-LSTM	-	46.85	55.68	33.07	45.26
LOGO-Former (ICASSP'23) [264]	ResNet-18	-	54.21	66.98	38.22	48.13
IAL (AAAI'23) [20]	ResNet-18	19	55.71	69.24	35.82	48.54
AEN (CVPRW'23) [244]	ResNet-18	-	56.66	69.37	38.18	47.88
M3DFEL (CVPR'23) [258]	ResNet-18-3D	-	56.10	69.25	35.94	47.67
Dual-STI (IS'24) [259]	ResNet-18	-	56.89	68.29	38.27	48.46
DFER-CLIP (BMVC'23) [23]	CLIP-ViT-B/32	90	59.61	71.25	41.27	51.65
CLIPER (ICME'24) [184]	CLIP-ViT-B/16	88	57.56	70.84	41.23	51.34
EmoCLIP (FG'24) [185]	CLIP-ViT-B/32	-	58.04	62.12	31.41	36.18
DFLM (ICSP'24) [265]	CLIP-ViT-B/32	-	59.77	71.40	41.25	51.31
CLIP-Guided-DFER (IJCB'24) [186]	CLIP-ViT-B/32	-	<u>60.85</u>	<u>72.58</u>	<u>41.43</u>	**51.83**
PE-CLIP (ours)	CLIP-ViT-B/16	**9**	**62.82**	**74.04**	**41.57**	<u>51.26</u>

Furthermore, as shown in Table 6.4, PE-CLIP demonstrates remarkable performance across various emotional categories on the DFEW dataset, further validating its robustness and effectiveness. Specifically, our method achieves 91.00% average accuracy on "happiness,"

the highest among all methods, benefiting from the distinct visual features and high representation of this emotion in the dataset. Similarly, for "sadness," "neutral," "angry," and "surprise," PE-CLIP consistently achieves high accuracy, demonstrating its ability to capture visually distinctive and high-intensity expressions. In more challenging categories, such as "fear" and "disgust," where data imbalance poses a significant challenge, PE-CLIP achieves superior results. For "fear," our model attains 45.12% average accuracy, marking a notable 8.68% improvement over the previous best method, IAL [20], reflecting its ability to capture nuanced temporal features in less prominent emotions. Similarly, for the underrepresented "disgust," PE-CLIP reaches 15.17% average accuracy, achieving an 6.20% improvement over Dual-STI [259], further highlighting its effectiveness in handling both well-represented and challenging emotional categories.

Table 6.4: Accuracy Comparison Among Various Emotion Categories on DFEW. "-" denotes results were not reported in the original paper.

Method	**Tunable Param (M)**	**Accuracy of Each Emotion**							**DFEW**	
		Hap.	**Sad.**	**Neu.**	**Ang.**	**Sur.**	**Dis.**	**Fea.**	**UAR**	**WAR**
Former-DFER (MM'21) [86]	18	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
CEFLNet (IS'22) [245]	13	84.00	68.00	67.00	70.00	52.00	0.00	17.00	51.14	65.35
NR-DFERNet (ArXiv'22) [260]	-	88.47	64.84	70.03	75.09	61.60	0.00	19.43	54.21	68.19
STT (ArXiv'22) [261]	-	87.36	67.90	64.97	71.24	53.10	3.49	34.04	54.58	66.65
EST (PR'23) [199]	43	86.87	66.58	67.18	71.84	47.53	5.52	28.49	53.43	65.85
IAL (AAAI'23) [20]	19	87.95	67.21	70.10	76.06	62.22	0.00	36.44	55.71	69.24
M3DFEL (CVPR'23) [258]	-	89.59	68.38	67.88	74.24	59.69	0.00	31.64	56.10	69.25
Dual-STI (IS'24) [259]	-	87.65	68.32	67.55	71.70	57.93	8.97	36.14	55.71	69.24
PE-CLIP (ours)	**9**	**91.00**	**79.62**	**71.94**	**79.80**	**85.10**	**15.17**	**45.12**	**62.82**	**74.04**

Results on FERV39K: The performance comparison of PE-CLIP with existing methods on the FERV39K dataset is presented in Table 6.3. Our approach achieves a UAR of 41.57% and a WAR of 51.26%, demonstrating superior results compared to vision-only based methods, and competitive results, particularly against CLIP-based methods. Compared to CLIP-Guided-DFER, PE-CLIP shows a marginal UAR improvement of +0.14%, while WAR exhibits a slight decline of -0.57%, which may be attributed to the increased variability and complexity of emotional representations in FERV39K. Unlike DFEW, which features more clearly distinguishable expressions, FERV39K introduces fine-grained nuances and imbalanced class distributions that pose challenges for all existing methods. Despite these factors, PE-CLIP maintains competitive performance, demonstrating its robustness in handling dynamic expressions across different datasets.

Model Size Efficiency: While PE-CLIP achieves competitive performance, its efficiency remains a key distinguishing factor. As shown in Table 6.3, our approach requires only 9M trainable parameters, significantly fewer than most existing methods while maintaining high performance. Compared to vision-only methods, such as Former-DFER (18M), CEFLNet (13M), and IAL (19M), PE-CLIP reduces the trainable parameter count by nearly 50% while achieving superior results. This demonstrates the advantage of using vision-language alignment for dynamic expression recognition over purely vision-based approaches. Among CLIP-based methods, such as DFER-CLIP (90M) and CLIPER (88M), which require nearly 10× more trainable parameters, PE-CLIP offers a substantial reduction in trainable parameters, achieving an impressive 90% reduction. In summary, we can conclude that PE-CLIP achieves an optimal balance between parameter efficiency and performance, requiring fewer than 6% of the model's parameters to be trained while avoiding full fine-tuning. These results highlight the effectiveness of PE-CLIP's parameter-efficient design and its practicality for DFER.

6.4.4 Ablation Analysis

To gain insights into the effectiveness of the key components in PE-CLIP, we conducted a series of ablation studies on the DFEW dataset. The analysis began with an assessment of the overall architecture. We then examined the contributions of the main components, including the proposed vision and textual adapters. Following this, we analyzed the impact of various temporal modeling techniques. Additionally, we analyzed how different reduction factor settings influence model performance and computational efficiency in terms of trainable parameters. Finally, we investigated the effect of textual prompting methods on alignment and overall performance. For consistency and clarity, all results are reported for fold 1 (fd1), aligning with prior studies [188, 243, 254].

Evaluation of the Overall Framework. Table 6.5 provides a detailed comparison of the performance of PE-CLIP overall framework and its key general components on the DFEW dataset. The baseline model, comprising frozen CLIP encoders with a single trainable vanilla adapter [155] integrated into the textual encoder after the MLP layers and class names as textual inputs, achieves a UAR of 49.82% and a WAR of 63.35% with only 0.8M trainable parameters. While effective in leveraging CLIP's pre-trained capabilities for generalizing to DFER, the baseline's performance is limited by its inability to fully capture temporal information and multimodal complexities. Adding lightweight adapters

to both the vision and text encoders significantly enhances performance, achieving a UAR of 61.74% and a WAR of 74.64%. This improvement of 11.92% in UAR and 11.29% in WAR underscores the effectiveness of adapters in refining pre-trained features for DFER while maintaining computational efficiency. Finally, incorporating Multi-modal Prompt Learning (MaPLe) with AU-based descriptions into the framework yields further gains, with an improvement of 1.28% in UAR and 0.71% in WAR. This progression highlights the critical role of semantically rich textual prompts in enhancing vision-language alignment. These results validate the overall framework of PE-CLIP, emphasizing the important contributions of the adapters and prompt learning in achieving competitive performance on DFER with only 9M trainable parameters.

Table 6.5: Performance Comparison of the Overall Framework Components on DFEW Dataset.

Model	**Params (M)**	**DFEW**	
		UAR	**WAR**
Baseline	0.8	49.82	63.35
Baseline + Adapters	6	61.74	74.64
Baseline + Adapters + Prompts (Ours)	**9**	**63.02**	**75.35**

Impact of the Proposed Adapters in PE-CLIP. Table 6.6 shows the impact of different adapter configurations within PE-CLIP, demonstrating their individual and combined contributions to DFER. All experiments in this analysis incorporate AU-MaPLe prompts to maintain a consistent textual prompting strategy. In the first configuration, with a single Textual Adapter (TA) placed after the MLP layers, the model achieves a UAR of 53.64% and a WAR of 68.18%, utilizing 4M trainable parameters. This relatively lower performance highlights the limitations of relying solely on textual adaptation, as it inadequately aligns textual features with the visual features essential for DFER. Introducing a second textual adapter after the MHSA layers leads to a UAR increase to 55.43% (+1.79%), and a slight WAR decrease to 68.09% (-0.09%). This trade-off could be attributed to a disparity in how textual features align with the dataset's class distribution, where textual features improve recall across classes but lack complementary vision adapters, reducing the model's ability to generalize effectively to frequent emotions. Shifting the focus to vision adapters, a configuration incorporating a Shared Adapter (ShA) after the MHSA layers and a Temporal Dynamic Adapter (TDA) after the MLP layers yields substantial

performance gains, achieving a UAR of 62.59% and a WAR of 74.43% with 8M trainable parameters. This represents a +7.16% UAR and +6.34% WAR improvement compared to the dual-text-adapter configuration, emphasizing the critical role of vision adapters in refining spatial features and capturing temporal dependencies. particularly through TDA's dynamic scaling mechanism, which prioritizes the most informative temporal features. However, the lack of textual embeddings refinements limits the model's ability to capture finer-grained expression nuances. Introducing ShA across both text and visual encoders while retaining the TA adapter results in a UAR of 60.76% and a WAR of 72.70%, utilizing 7M trainable parameters. This configuration benefits from ShA's ability to effectively refine the extracted features in both encoders. However, the absence of TDA restricts its capability to fully capture temporal variations, leading to slightly lower performance than the previous configuration. Finally, integrating all adapters, TA, ShA, and TDA, achieves the best performance, with a UAR of 63.02% and a WAR of 75.35%, using 9M trainable parameters. Notably, adding TDA leads to a +2.26% UAR and +2.65% WAR increase compared to the configuration without it, highlighting the important role of the temporal modeling in DFER, where effectively capturing dynamic relationships across frames is crucial. These findings confirm the complementary nature of each adapter in PE-CLIP framework, allowing it to achieve good performance while maintaining efficiency with just 9M trainable parameters.

Table 6.6: Impact of Different Adapters on PE-CLIP Performance on the DFEW Dataset ("TA": Textual Adapter, "ShA": Shared Adapter, "TDA": Temporal Dynamic Adapter).

Text		**Video**		**Params (M)**	**DFEW**	
TA	ShA	ShA	TDA		UAR	WAR
✓	✗	✗	✗	4	53.64	68.18
✓	✓	✗	✗	5	55.43	68.09
✗	✗	✓	✓	8	62.59	74.43
✓	✓	✓	✗	7	60.76	72.70
✓	✓	✓	✓	9	**63.02**	**75.35**

Impact of Dynamic Scaling in Temporal Dynamic Adapter (TDA). Table 6.7 evaluates the influence of dynamic scaling applied at different stages within the TDA adapter, demonstrating its importance in enhancing temporal modeling. Without dynamic scaling, the TDA achieves a UAR of 62.10% and a WAR of 75.22%. While the GRU effectively captures temporal dependencies, the absence of frame-specific weighting limits

the model's ability to adaptively emphasize salient temporal features, establishing this configuration as a baseline. Applying dynamic scaling at the input level before GRU processing improves UAR by 0.63% but slightly reduces WAR by 0.94%. The increase in UAR highlights an improved focus on underrepresented classes, as input-level scaling helps prioritize discriminative temporal patterns before they are modeled. However, the slight decrease in WAR suggests that early-stage scaling may disrupt the natural temporal coherence modeled by the GRU, leading to overemphasis on certain temporal variations. When applied directly to the GRU's output, dynamic scaling results in decreases of 0.48% in UAR and 0.55% in WAR, likely due to interference with the refined temporal dependencies captured by the GRU. Weighting at this stage alters the consistency of learned representations, reducing the effectiveness of the adaptation. The best performance is achieved when dynamic scaling is applied after the up-projection layer, yielding a UAR of 63.02% (+0.92%) and a WAR of 75.35% (+0.13%) compared to the baseline. This approach ensures adaptive weighting of the refined temporal features after dependencies are fully modeled, while effectively enhancing salient temporal cues and mitigating irrelevant or noisy variations. Across all configurations, the number of trainable parameters remains constant at 9M, as dynamic scaling introduces only a negligible computational overhead. These results highlight the efficiency and effectiveness of incorporating dynamic scaling into the TDA, enabling the model to achieve optimal performance without compromising its lightweight design.

Table 6.7: Impact of dynamic scale mechanism in TDA. ("w/o": without, "DS": dynamic Scale).

Setting	**Params (M)**	**DFEW**	
		UAR	**WAR**
TDA w/o DS	9	62.10	75.22
TDA w/ Input-Level Scaling	9	62.73	74.28
TDA w/ GRU Output Scaling	9	61.62	74.67
TDA w/ Post-Up-Projection Scaling (Ours)	**9**	**63.02**	**75.35**

Impact of Temporal Modeling Architectures on PE-CLIP Performance. We evaluated the performance of PE-CLIP using different RNN-based temporal adapter configurations within TDA, as detailed in Table 6.8. The Vanilla Adapter, which lacks temporal modeling, achieves a UAR of 58.80% and a WAR of 72.23%, demonstrating its inability to capture sequential dependencies across frames. Introducing an RNN-based

temporal adapter improves UAR to 62.18% (+3.38%) and WAR to 73.86% (+1.63%), demonstrating enhanced temporal modeling. However, its limited ability to balance short- and long-term dependencies restricts its effectiveness. The LSTM-based adapter, an advanced RNN variant with gated memory mechanisms, further improves WAR to 74.16% (+1.93%) but slightly reduces UAR to 61.29% (-0.89% compared to the RNN-based adapter). This suggests that LSTM's long-term dependency modeling benefits dominant classes while being slightly less sensitive to underrepresented ones. The GRU-based adapter, a computationally efficient alternative to LSTM, achieves the best performance, with a UAR of 63.02% (+4.22%) and WAR of 75.35% (+3.12%) compared to the baseline, outperforming the LSTM-based adapter by +1.73% in UAR and +1.19% in WAR. These results highlight GRU's efficiency in balancing computational cost and temporal accuracy due to its gated recurrent mechanism, making it particularly suitable for DFER.

Table 6.8: Performance Comparison of Temporal Adapter Configurations within TDA on DFEW.

Temporal Adapter	**Params (M)**	**DFEW**	
		UAR	**WAR**
Vanilla Adapter[155]	9	58.80	72.23
RNN-based Adapter	9	62.18	73.86
LSTM-based Adapter	10	61.29	74.16
GRU-based Adapter (Ours)	**9**	**63.02**	**75.35**

Effect of Reduction Factor on Efficiency and Model Performance. Table 6.9 shows the impact of varying reduction factors on the efficiency and performance of PE-CLIP. The reduction factor determines the degree of dimensionality reduction within the bottleneck layers of adapters, directly affecting both trainable parameters and recognition accuracy. With a reduction factor of 4, the model utilizes 17M trainable parameters, achieving a UAR of 61.39% and a WAR of 73.86%. While this setting effectively captures features, its higher parameter count reduces practicality for resource-constrained scenarios. Increasing the reduction factor to 8 reduces trainable parameters to 9M, while improving performance to a UAR of 63.02% (+1.63%) and a WAR of 75.35% (+1.49%) compared to the reduction factor of 4. This configuration strikes the best balance between dimensionality reduction and representation quality, demonstrating its suitability for DFER. Further increasing of the reduction factor to 16 lowers the trainable parameters to 6M but slightly decreases performance, with a UAR of 62.68% (-0.34%) and a WAR of 74.75% (-0.60%). This

suggests that too much reduction constrains the adapters' ability to retain important temporal and spatial features, leading to performance degradation. Overall, a reduction factor of 8 offers the optimal trade-off, achieving a highly parameter-efficient design while maintaining robust performance, making it well-suited for DFER.

Table 6.9: Influence of the reduction factor of adapters.

Reduction Factor	**Params (M)**	**DFEW**	
		UAR	**WAR**
4	17	61.39	73.86
8 (Ours)	**9**	**63.02**	**75.35**
16	6	62.68	74.75

Impact of Textual Prompt. We analyzed the impact of different textual inputs and prompting mechanisms on the PE-CLIP's performance, as shown in Table 6.10. Using only class names (e.g., "Happiness", "Anger") the model achieves a UAR of 61.74 % and a WAR of 74.46%. However, the lack of semantic richness limits the model's ability to capture nuanced expressions. Introducing learnable prompts via CoOp [251], which incorporates a generic learnable token into textual inputs, results in a performance decline of (-0.82%) UAR and (-1.54%) WAR compared to class names, highlighting its lack of task-specific guidance. Replacing class names with AU-based descriptions within the CoOp framework improves performance, yielding a +0.31% UAR and +0.81% WAR increase, demonstrating the value of richer semantic information from AUs. However, CoOp's generic token learning still restricts its effectiveness. A task-specific prompting approach using ChatGPT-generated descriptions [23] further enhances UAR by +1.63% and WAR by +0.21%, underscoring the importance of tailored prompts for textual-visual alignment. Finally, MaPLe with AU-based descriptions, which introduces learnable tokens for both textual and visual inputs, achieves the highest WAR (75.35%) and a UAR of (63.02%). Compared to CoOp with AU descriptions, MaPLe demonstrates a substantial improvement of +1.79% UAR and +1.62% WAR, highlighting the effectiveness of multi-modal prompting in dynamic facial expression recognition. However, compared to ChatGPT-based prompting [23], MaPLe achieves a slightly lower UAR (-0.35%) but a higher WAR (+0.68%), suggesting that ChatGPT-based prompts better support underrepresented classes, whereas MaPLe with AU descriptions optimizes alignment for dominant classes, enhancing overall performance.

Table 6.10: Performance comparison of key components on the DFEW dataset.

Setting	DFEW	
	UAR	**WAR**
Class Names	61.74	74.46
CoOp (Generic Learnable Token) [251]	60.92	72.92
CoOp + AU Descriptions	61.23	73.73
Learnable prompt by ChatGPT [253]	**63.37**	74.67
MaPLe + AU Descriptions (Ours)	<u>63.02</u>	**75.35**

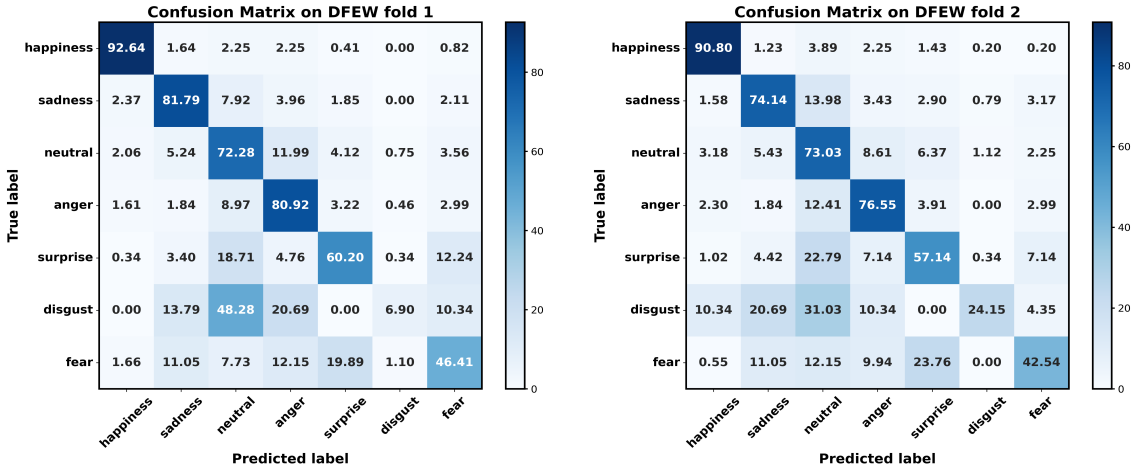
6.4.5 Visualization-Based Results

Confusion Matrices

The confusion matrices in Figure 6.3 provide further insights, aligning closely with the quantitative results in Table 6.4. Emotions like "happiness," "surprise," and "anger" exhibit the highest classification accuracies, as they feature distinctive visual cues and are well-represented in the dataset. In contrast, more complex emotions like "fear" and "disgust" show higher misclassification rates. Notably, "fear" is often misclassified as "surprise" due to shared facial cues such as raised eyebrows and widened eyes. Similarly, "disgust" is frequently confused with both "neutral" and "anger". While "anger" shares overlapping features such as furrowed brows and tightened lips, "neutral" exhibits minimal facial movement, making it challenging to distinguish from "disgust". These misclassifications underscore the inherent challenges in distinguishing visually similar emotions.

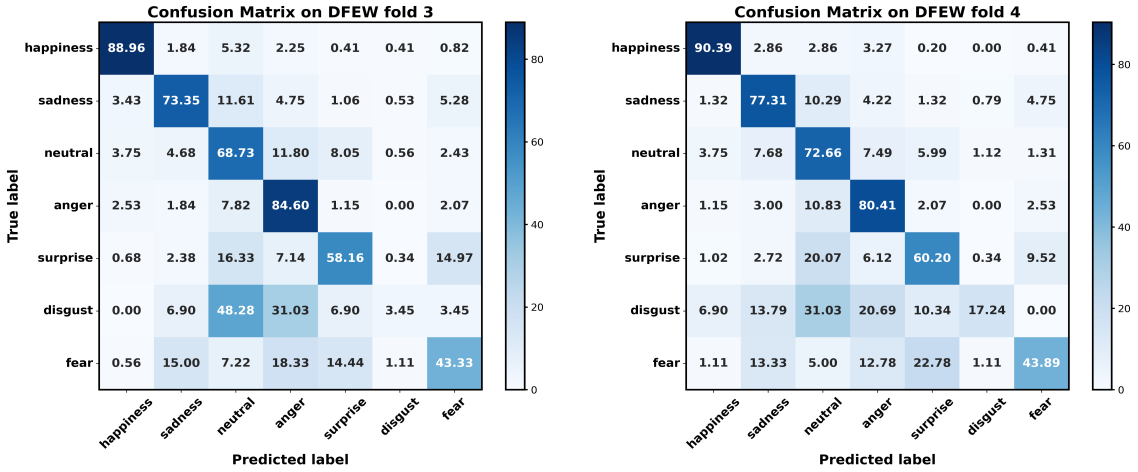
Attention Visualization

To "see" the contributions of PE-CLIP components, we visualize (in Figure 6.4) the attention map of the last transformer block using the Gradient Attention Rollout technique, including sadness and anger. From top to bottom, the visualizations correspond to (1) the input (first row), (2) the baseline model (second row), (3) the baseline + ShA adapter (third row), and (4) the full PE-CLIP framework (Baseline + ShA + TDA) (fourth row). These visualizations show that PE-CLIP progressively refines attention allocation, enabling the model to focus on emotion-relevant facial regions while effectively capturing temporal variations across frames. For instance, In "sadness", we observe distinct improvements



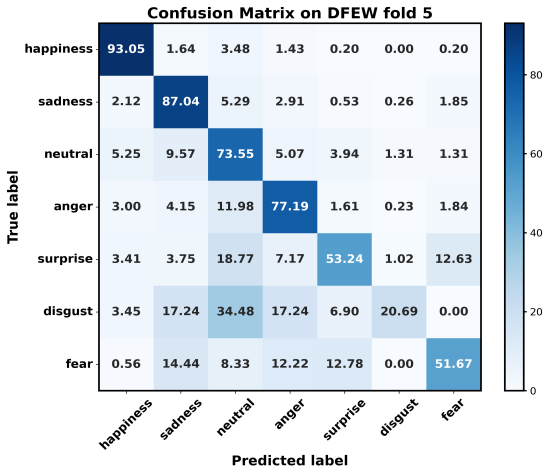
(a) Fold 1

(b) Fold 2



(c) Fold 3

(d) Fold 4



(e) Fold 5

Figure 6.3: Confusion Matrices Across 5 Folds on DFEW.

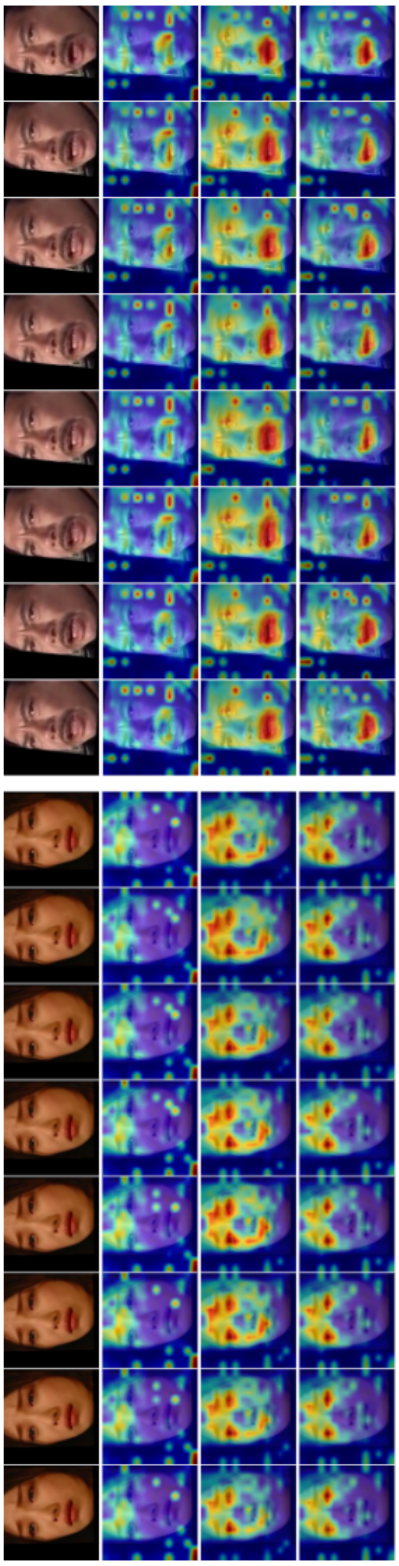


Figure 6.4: Attention visualization of our proposed PE-CLIP model. The figure presents attention maps for *sadness* and *anger* (left to right), showing how focus evolves across different model configurations. Each section includes original frames (first row), baseline model (second row), model with ShA adapters (third row), and full PE-CLIP with ShA and TDA (fourth row). Warmer colors indicate stronger attention, with PE-CLIP progressively refining focus on key facial regions, enhancing representation refinement (via ShA) and temporal modeling (via TDA).

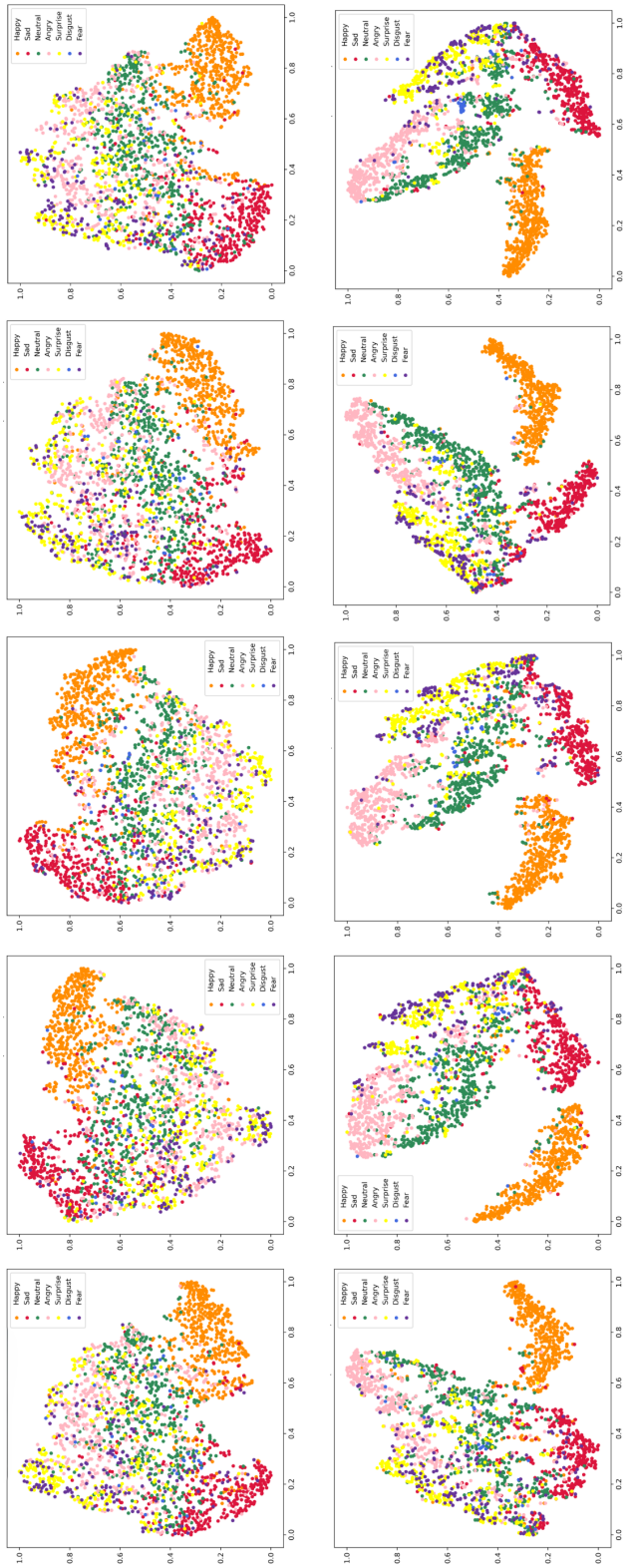


Figure 6.5: t-SNE Visualization of High-Level Feature Distributions on DFEW (fd1–5): baseline Model, PE-CLIP without adapters and prompt (Top row) vs. Proposed PE-CLIP model (Bottom row).

by adding PE-CLIP components. The baseline model displays widely dispersed attention, often focusing on irrelevant areas such as the cheeks and forehead. Adding ShA to both vision and text encoders, refines this distribution, directing the focus toward the brows, eyes, and mouth, which are strongly associated with AU1 (Inner Brow Raiser) and AU15 (Lip Corner Depressor), important indicators of sadness. With TDA integration, the attention becomes further concentrated, reinforcing the importance of these regions over time and capturing subtle temporal variations essential for distinguishing sadness, particularly the progressive furrowing of the brows across frames. A similar trend is observed for "anger", where the baseline model fails to localize important facial cues. Incorporating ShA shifts the attention toward the brows and jawline, aligning with AU4 (Brow Lowerer) and AU17 (Chin Raiser), key action units for anger. The TDA adapter further enhances this by allowing the model to track the increasing intensity of furrowed eyebrows and tightened lips over time, ensuring a more robust temporal understanding. These observations demonstrate that PE-CLIP effectively captures the dynamic transitions of facial expressions, refining attention to important regions and enhancing temporal consistency for robust DFER.

Feature Distribution Visualization

Figure 6.5 presents t-SNE (t-distributed stochastic neighbor embedding [114], presented in 2.3.2) visualizations comparing the feature distributions of the baseline model (top row) and PE-CLIP (bottom row) across five folds of the DFEW dataset. In the baseline model, features for visually similar emotions, such as "fear" and "surprise" (both involving raised eyebrows) or "disgust" and "anger" (where both are characterized by furrowed brows and tightened lips), exhibit significant overlap, leading to higher misclassification rates. Additionally, the baseline shows scattered and loosely clustered features, indicating limited temporal modeling and weak alignment between visual and textual modalities. In contrast, PE-CLIP demonstrates more well-separated and distinct clusters, reflecting its enhanced ability to capture discriminative features. Emotions with strong visual cues ("happiness", "anger", "surprise") exhibit well-defined clusters, while even challenging expressions like "fear" and "disgust" achieve improved separation. These visualizations further support the quantitative results of PE-CLIP and highlight the effectiveness of its temporal modeling (TDA) and feature refinement (ShA and TA adapters) in providing semantically rich and discriminative representations.

6.5 Conclusion

In this chapter, we proposed a final contribution, PE-CLIP, a parameter-efficient method that effectively extended CLIP for DFER by integrating lightweight adapters and multi-modal prompt learning to improve temporal modeling and feature refinement while requiring only 9M trainable parameters without full fine-tuning of CLIP's encoders. The Temporal Dynamic Adapter (TDA), incorporating a GRU-based structure with dynamic scaling, allowed the model to adaptively emphasize significant temporal variations, while the Shared Adapter (ShA) facilitated efficient feature processing across both modalities. Additionally, we leveraged multi-modal prompts (MaPLe) and Action Unit (AU)-based descriptions, providing semantically rich textual guidance to further improve vision-language consistency in dynamic facial expression recognition. Experimental results on benchmark datasets demonstrated that PE-CLIP achieved competitive performance compared to state-of-the-art methods while maintaining a lightweight design, making it a parameter-efficient for Dynamic FER applications. The ablation study sheds light on the contributions of each component of our proposed model, particularly in enhancing temporal modeling, feature refinement, and optimizing the representation learning.

7 Conclusion and Future Work

7.1 Summary of Contributions

This thesis developed deep learning-based methods aimed at improving the performance, robustness, and efficiency of facial expression recognition systems in challenging conditions, such as lighting variations, head pose changes, occlusions, and the inherent variability of facial expressions. The research initially focused on the application of FER in the driving environment, where recognizing driver emotions is crucial for enhancing road safety. To that end, models were proposed ranging from single-modality Vision Transformer-based architectures to multimodal approaches leveraging vision-language models. Subsequently, the scope was extended to general FER, targeting dynamic, in-the-wild environments where expressions are more spontaneous and uncontrolled.

The challenges addressed throughout this work have led to a series of contributions that respond directly to the research questions introduced earlier. These contributions can be summarized as follows:

- A comprehensive review of the FER field was conducted, with an emphasis on its application in driving environments. The review covered both theoretical and technical fundamentals, including key concepts, available datasets, evaluation protocols, and FER-related studies, identifying key limitations and open challenges.
- A novel transfer learning-based method, DFER-GCViT, was proposed to investigate the Global Context Vision Transformer (GC-ViT) for FER in driving environments. This model effectively captures both local and global features through self-attention mechanisms and achieves competitive accuracy under challenging conditions such as occlusions and illumination changes.
- To address the computational inefficiency of the proposed previous model, a light-weight hybrid architecture, ShuffViT-DFER, was introduced by combining ShuffleNet

V2 and EfficientViT-M2. This model offers an efficient FER solution with significantly reduced processing cost while maintaining competitive performance in resource-constrained driving setting.

- A multimodal FER framework, CLIVP-FER, was developed to integrate visual and textual information using the CLIP model. This approach enhances semantic representation and generalization capability across domains, achieving the best performance in cross-database evaluations as well with efficient inference.
- A parameter-efficient FER model, PE-CLIP, was proposed to adapt the CLIP framework for dynamic video-based FER. By incorporating lightweight adapters, prompt tuning, and AU-based textual prompts, the model achieved high accuracy with only 9M trainable parameters, providing a practical solution for FER in challenging scenarios.

These contributions highlight the essential role of deep learning in advancing the FER field, showing how carefully designed models can improve recognition accuracy while addressing practical constraints such as efficiency and environmental complexity. Collectively, they contribute to the broader objective of improving the accuracy and efficiency of FER systems across both domain-specific applications (e.g., driving environment) and general contexts (e.g., dynamic, in-the-wild settings).

7.2 Future Work

While the contributions presented in this thesis have addressed several key challenges and demonstrated promising results in facial expression recognition, particularly under constrained and dynamic conditions, there remain important directions for future research and development.

First, although the proposed models exhibited competitive performance, further improvements in **model optimization and computational efficiency** are necessary to facilitate real-time deployment on edge or embedded systems. Techniques such as pruning, quantization, and low-rank adaptation could further reduce model complexity and inference time without compromising accuracy, particularly in resource-constrained environments, where lightweight architectures and fast processing are essential.

Second, the integration of multimodal learning through vision-language models, as explored in CLIVP-FER and PE-CLIP, has demonstrated strong potential. However, investigating more advanced **multimodal fusion strategies**, such as cross-modal attention or gated fusion mechanisms, could better integrate information across modalities. Additionally, incorporating **other modalities**, including audio, contextual cues, or physiological signals, could help interpret subtle or ambiguous expressions effectively, leading to more comprehensive emotion recognition systems.

Third, ensuring robust **domain generalization** remains an open challenge. Despite the use of cross-database evaluation, such as in CLIVP-FER, domain shifts still led to performance degradation. To improve generalization, future work could explore domain adaptation techniques and self-supervised learning strategies to reduce dependency on large labeled datasets.

Fourth, the **scarcity of large-scale, annotated datasets**, particularly driver-specific datasets. While datasets such as KMU-FED have provided valuable benchmarks, their size and diversity remain insufficient for capturing the full range of expressions encountered in real driving scenarios. To address this, future research should focus on the collection and public release of more diverse, naturalistic datasets that reflect varying demographics, lighting conditions, and real-world driving contexts. This will be essential to improving both the training and evaluation of FER models.

Finally, beyond benchmark evaluation, it is important to advance toward **real-time deployment and validation in practical settings**. While strong results were achieved on benchmark datasets, further experimentation in realistic conditions and embedded systems is necessary to evaluate the feasibility and effectiveness of the proposed models under real-world conditions.

In summary, this thesis contributes to advancing FER by tackling performance, efficiency, and semantic understanding using deep learning-based methods. The proposed models offer practical and scalable solutions to challenging conditions. Future work along the directions outlined above will help drive further innovation and bring FER systems closer to robust, real-time deployment in both domain-specific and general contexts.

Résumé substantiel en français

1 Contexte et Motivation

Les expressions faciales constituent un canal fondamental de la communication non verbale humaine. Elles permettent de transmettre des informations essentielles relatives aux émotions, aux intentions et aux états cognitifs d'un individu, jouant ainsi un rôle clé dans les interactions sociales [1, 3]. La capacité à analyser automatiquement ces expressions a suscité un intérêt croissant dans le domaine de l'informatique affective, avec des applications potentielles dans des domaines variés tels que l'interaction homme–machine, la santé, la surveillance intelligente et les systèmes de sécurité [11, 53].

La reconnaissance automatique des expressions faciales (Facial Expression Recognition – FER) vise à identifier l'état émotionnel d'une personne à partir d'images ou de séquences vidéo de son visage [2, 10]. Les progrès récents de l'apprentissage profond ont permis des avancées significatives dans ce domaine, en particulier grâce aux réseaux de neurones convolutifs (CNNs) et aux architectures basées sur des mécanismes d'attention [16, 19]. Dans des conditions contrôlées, ces approches atteignent aujourd'hui des performances élevées, démontrant leur capacité à extraire des représentations discriminantes à partir des données faciales. Toutefois, la reconnaissance des expressions faciales demeure un défi majeur dans des environnements réalistes et non contrôlés. Dans de tels contextes, les performances des systèmes de FER se dégradent fortement en raison de nombreux facteurs perturbateurs, tels que les variations d'illumination, les changements de pose, les occlusions partielles du visage, la faible résolution des images ou encore la présence d'expressions de faible intensité [25, 26]. Ces difficultés sont accentuées lorsque l'analyse repose sur des séquences vidéo dynamiques, où les expressions évoluent dans le temps et où les informations pertinentes peuvent être subtiles et transitoires.

Le contexte de la conduite automobile illustre particulièrement bien ces défis. L'analyse des expressions faciales du conducteur présente un intérêt majeur pour le développement

de systèmes avancés d'assistance à la conduite et de véhicules intelligents, notamment pour l'évaluation d'états tels que la fatigue, le stress ou la distraction. Cependant, ce contexte impose des contraintes spécifiques : variations d'éclairage importantes, occlusions fréquentes dues aux accessoires ou aux gestes du conducteur, mouvements naturels de la tête (voir Figure 1), ainsi que des exigences strictes en termes de traitement en temps réel sur des systèmes embarqués aux ressources limitées.



Figure 1: Illustration des défis de la reconnaissance des expressions faciales en contexte de conduite. [200].

Ces constats mettent en évidence la complexité de la reconnaissance des expressions faciales en conditions réelles et soulignent la nécessité de développer des approches capables de faire face à ces contraintes. Ils constituent le point de départ de la problématique de recherche abordée dans cette thèse.

2 Problématique et objectifs

Malgré les progrès réalisés dans le domaine de la reconnaissance des expressions faciales, les méthodes existantes présentent encore des limitations importantes lorsqu'elles sont confrontées à des environnements réalistes et contraints [20, 23]. En particulier, les modèles actuels peinent à maintenir des performances élevées face aux variations environnementales, aux occlusions et aux changements de pose, tout en respectant les contraintes computationnelles imposées par les systèmes embarqués. De plus, de nombreuses approches montrent une capacité de généralisation limitée lorsqu'elles sont appliquées à des bases de données ou à des contextes différents de ceux utilisés lors de l'entraînement.

La problématique centrale peut être formulée de la manière suivante : *Comment concevoir des systèmes de reconnaissance des expressions faciales robustes, efficaces et efficients, adaptés à des environnements réalistes et contraints, en particulier dans le contexte de la conduite automobile ?*

Dans ce cadre, les objectifs scientifiques poursuivis sont définis comme suit :

- développer des modèles d'apprentissage profond capables d'améliorer la modélisation des informations spatiales et contextuelles du visage pour la reconnaissance des expressions faciales en conditions contraignantes, en particulier en contexte de conduite automobile.
- proposer une architecture de reconnaissance des expressions faciales à la fois légère et efficiente, adaptée à un déploiement en temps réel dans des environnements à ressources limitées, tels que les systèmes embarqués automobiles.
- explorer l'intégration conjointe d'informations visuelles et textuelles à travers des modèles vision-langage, afin d'améliorer la compréhension sémantique et les performances de reconnaissance dans des scénarios de reconnaissance des expressions faciales centrés sur le conducteur.
- étendre l'application de la reconnaissance des expressions faciales au-delà du contexte de la conduite, vers des environnements dynamiques et non contraints, en intégrant la modélisation temporelle et des stratégies d'apprentissage multimodal efficientes.

Ces objectifs structurent les travaux présentés et guident les choix méthodologiques et expérimentaux développés dans les sections suivantes.

3 Fondements théoriques et méthodologiques

La reconnaissance des expressions faciales constitue une tâche centrale de l'informatique affective, visant à analyser les signaux visuels du visage afin d'inférer l'état émotionnel d'un individu. Selon les contextes d'application, cette tâche peut être formulée de différentes manières, notamment comme un problème de classification d'émotions discrètes [39] ou comme l'estimation de représentations affectives alternatives [43, 46]. Dans le cadre de cette thèse, la reconnaissance des expressions faciales est principalement abordée comme une tâche de classification appliquée à des données visuelles du visage, considérées sous forme d'images ou de séquences vidéo.

3.1 Applications de la reconnaissance des expressions faciales en contexte de conduite

La reconnaissance des expressions faciales est utilisée dans de nombreux domaines nécessitant l'analyse des états émotionnels et cognitifs humains. Elle trouve des applications dans l'interaction homme–machine, les systèmes de surveillance intelligents, la santé, l'éducation ou encore les interfaces adaptatives, où l'interprétation des expressions faciales permet d'améliorer l'adaptation des systèmes aux utilisateurs [48, 49].

Dans cette thèse, l'accent est principalement mis sur les applications de la reconnaissance des expressions faciales en contexte de conduite automobile. Dans ce domaine, la FER constitue un composant essentiel des systèmes de surveillance et d'assistance au conducteur, permettant d'analyser des états tels que la fatigue, le stress, la distraction ou la charge cognitive [11]. Ces informations peuvent être exploitées pour déclencher des alertes, ajuster le comportement des systèmes d'assistance à la conduite ou adapter les interfaces homme–machine en fonction de l'état du conducteur. Dans les systèmes de conduite modernes [63, 64], la reconnaissance des expressions faciales est généralement intégrée dans un cadre d'analyse continue du comportement du conducteur. Elle contribue à une évaluation dynamique de son état, en complément d'autres signaux issus du véhicule ou de capteurs embarqués, afin de soutenir la prise de décision des systèmes intelligents. Contrairement aux scénarios génériques de reconnaissance des émotions, ces applications nécessitent un fonctionnement fiable sur des flux de données continus, tout en respectant des contraintes strictes de temps réel et de ressources computationnelles propres aux systèmes embarqués automobiles.

3.2 Workflow général d'un système de reconnaissance des expressions faciales

De manière générale, un système de reconnaissance des expressions faciales repose sur un pipeline structuré en plusieurs étapes successives, depuis l'acquisition des données faciales jusqu'à la prédiction de l'état émotionnel (Classification). Ce workflow constitue un cadre commun à la majorité des systèmes FER, indépendamment des méthodes de modélisation utilisées, et sert de base aux approches développées dans cette thèse. La Figure 2 illustre les principales étapes de ce processus.

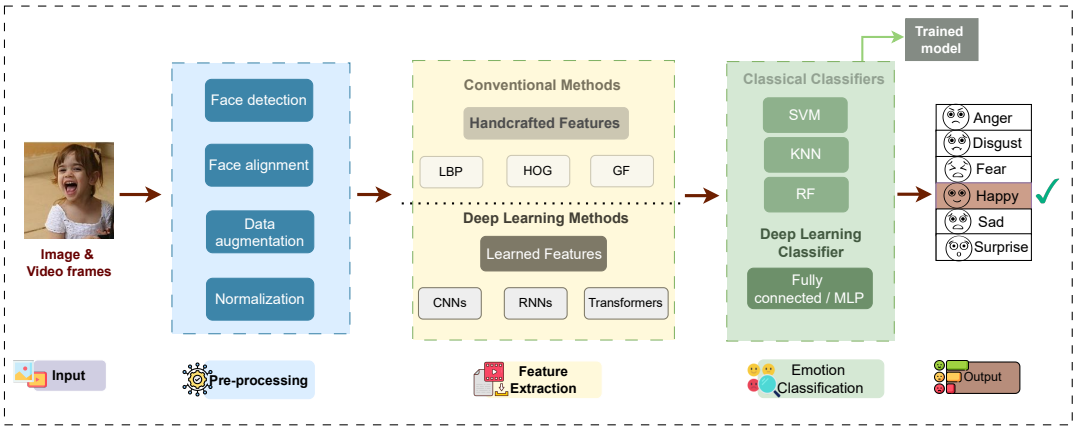


Figure 2: Workflow général d'un système de reconnaissance des expressions faciales, comprenant l'acquisition des données, le prétraitement, l'extraction de caractéristiques et la classification des émotions.

La première étape concerne l'acquisition et la représentation des données d'entrée. Selon les scénarios considérés, les entrées peuvent être de nature statique, sous forme d'images faciales isolées, ou dynamique, sous forme de séquences vidéo. Dans le cas statique, chaque image est analysée indépendamment afin d'estimer l'expression faciale correspondante [162, 163]. À l'inverse, les approches dynamiques exploitent une succession d'images permettant de capturer l'évolution temporelle des expressions, ce qui est particulièrement pertinent dans des environnements réalistes où les émotions apparaissent progressivement et peuvent être de faible intensité [9, 23].

Une fois les données acquises, une étape de prétraitement est généralement appliquée afin de préparer les entrées pour l'analyse. Cette étape comprend typiquement la détection et le recadrage du visage [67, 68], l'alignement facial, ainsi que des opérations de normalisation visant à réduire la variabilité non pertinente liée aux conditions d'acquisition [70] . Des techniques telles que la correction de l'illumination ou l'augmentation des données [71] peuvent également être employées afin d'améliorer la robustesse des systèmes face aux variations d'apparence. Le prétraitement joue un rôle important dans la stabilité et la fiabilité des systèmes de reconnaissance, en particulier lorsque les données sont acquises dans des environnements contraints.

L'étape suivante correspond à l'extraction et à la représentation des caractéristiques faciales. L'extraction de caractéristiques vise à transformer les données faciales prétraitées en une représentation numérique compacte et discriminante, adaptée à la classification

des expressions émotionnelles. La qualité de cette représentation joue un rôle déterminant dans les performances des systèmes de reconnaissance des expressions faciales, car elle conditionne la capacité du modèle à distinguer efficacement les variations émotionnelles pertinentes.

Les approches traditionnelles reposent sur des caractéristiques dites artisanales, généralement regroupées en deux grandes catégories : les caractéristiques d'apparence et les caractéristiques géométriques. Les caractéristiques d'apparence décrivent les variations de texture et d'intensité du visage, à l'aide de descripteurs tels que les motifs binaires locaux (LBP) [12], les histogrammes de gradients orientés (HOG) [13] ou les filtres de Gabor [72]. Elles peuvent être extraites globalement sur l'ensemble du visage ou localement sur des régions d'intérêt expressives. Les caractéristiques géométriques, quant à elles, modélisent les relations spatiales entre des points caractéristiques du visage, tels que les distances et angles entre repères faciaux, offrant une représentation plus directement liée aux mouvements musculaires [77]. Bien que ces approches soient relativement légères et interprétables, leurs performances dépendent fortement de la qualité du prétraitement et de la détection précise des repères faciaux, en particulier dans des environnements contraints.

Avec le développement de l'apprentissage profond, l'extraction automatique de caractéristiques est devenue prédominante en reconnaissance des expressions faciales. Ces approches permettent d'apprendre directement, à partir des données brutes, des représentations hiérarchiques adaptées à la tâche de reconnaissance émotionnelle [66]. Les modèles convolutifs sont principalement utilisés pour l'analyse des informations spatiales du visage, tandis que des architectures intégrant une modélisation temporelle sont employées pour exploiter la dynamique des expressions dans les séquences vidéo [81, 83]. Plus récemment, les architectures basées sur l'attention, et en particulier les transformateurs, ont suscité un intérêt croissant en raison de leur capacité à capturer des relations globales et à mettre en évidence les régions faciales les plus informatives [86].

Enfin, une étape de classification permet d'associer la représentation extraite à une catégorie émotionnelle, le plus souvent définie selon un ensemble d'émotions discrètes. Cette étape peut être réalisée à l'aide de méthodes classiques, de modèles d'apprentissage profond ou d'approches hybrides, selon la manière dont l'extraction des caractéristiques et la décision sont couplées [66, 87]. Dans les systèmes modernes, la classification est généralement intégrée dans un processus d'apprentissage de bout en bout, tandis que certaines approches combinent des représentations apprises avec des classificateurs conventionnels afin d'offrir

davantage de flexibilité [89, 92]. En contexte de conduite, les résultats de la classification sont exploités dans un cadre plus large de surveillance du conducteur, afin de soutenir la prise de décision des systèmes d'assistance et l'adaptation de l'interaction homme–machine en temps réel.

3.3 Jeux de données et métriques d'évaluation pour FER

La recherche en reconnaissance des expressions faciales s'appuie sur une grande variété de jeux de données, conçus pour couvrir des scénarios et des conditions d'acquisition différentes. Ces jeux de données peuvent être classés selon plusieurs dimensions clés, notamment le type de données (statiques ou dynamiques), le contexte d'acquisition (environnements contrôlés ou in-the-wild), ainsi que le domaine d'application visé.

Les jeux de données statiques sont constitués d'images faciales isolées, généralement annotées selon un ensemble d'émotions discrètes. Ils sont souvent acquis en laboratoire, dans des conditions contrôlées, avec des expressions clairement posées, et ont longtemps servi de base à l'évaluation des systèmes de FER classiques, par exemple JAFFE [99], KDEF [100], et MPI [4]. À l'inverse, les jeux de données dynamiques reposent sur des séquences vidéo , comme CK+ [94] et MMI [95], permettant d'analyser l'évolution temporelle des expressions faciales. Ces données offrent une représentation plus réaliste des comportements émotionnels, mais introduisent également des défis supplémentaires liés à la variabilité temporelle et au bruit présent dans les séquences. Afin de rapprocher les systèmes de FER des conditions d'utilisation réelles, des jeux de données acquis en environnements non contrôlés, dits in-the-wild, ont été développés. Ces bases de données se caractérisent par une forte variabilité des conditions d'éclairage, des poses, des identités et des expressions, rendant la reconnaissance émotionnelle particulièrement complexe, FER2013 [97], RAF-DB [96], DFEW [102] et FERV39K [101]. Elles constituent aujourd'hui un cadre d'évaluation essentiel pour analyser la robustesse et la capacité d'adaptation des approches proposées.

Par ailleurs, des jeux de données spécifiques à certains domaines d'application ont émergé, notamment dans le contexte de la conduite automobile, tel que KMU-FED [103], DMD [107], et DEFE [106]. Ces bases de données sont conçues pour analyser les expressions faciales du conducteur en situation réelle ou simulée, et présentent des contraintes propres aux environnements embarqués, telles que des occlusions fréquentes, des variations d'illumination importantes et des expressions souvent de faible intensité [**?** **?**]. Ces jeux de données

jouent un rôle central dans l'évaluation des méthodes développées dans cette thèse, car ils permettent de mesurer les performances des systèmes dans des conditions proches des scénarios d'utilisation visés.

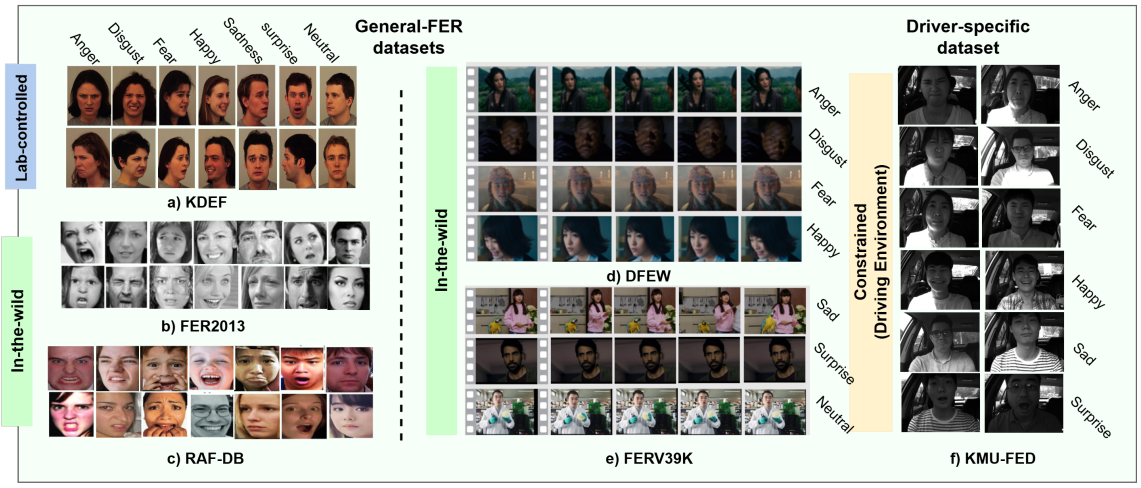


Figure 3: Exemples d'échantillons issus des jeux de données de reconnaissance des expressions faciales utilisés dans ce thèse.

La Figure 3 présente une vue d'ensemble des principaux jeux de données utilisés dans ce travail. Elle illustre la diversité des données considérées, couvrant à la fois des bases statiques et dynamiques, acquises en environnements contrôlés ou réels, ainsi que des jeux de données à usage général et spécifiques au contexte de la conduite.

3.4 Métriques d'évaluation et des techniques de visualisation

L'évaluation des méthodes de reconnaissance des expressions faciales repose à la fois sur des métriques quantitatives et sur des techniques de visualisation, en fonction de la nature des données analysées et de la formulation de la tâche. Dans cette thèse, l'accent est principalement mis sur la reconnaissance catégorielle des émotions, à partir d'images statiques et de séquences vidéo, ce qui oriente le choix des métriques d'évaluation utilisées.

Pour les données statiques, l'évaluation repose sur des métriques classiques issues de la matrice de confusion, telles que l'exactitude (accuracy), la précision, le rappel (recall) et le score F1. Ces indicateurs permettent de mesurer la capacité des modèles à prédire correctement les différentes catégories émotionnelles, tant en termes de performance globale

que de comportement par classe. Dans le cas des données dynamiques, où les expressions évoluent dans le temps et où les classes émotionnelles sont souvent déséquilibrées, des métriques spécifiquement adaptées sont privilégiées. Des mesures telles que l'Unweighted Average Recall (UAR) et le Weighted Average Recall (WAR) permettent de tenir compte de la distribution inégale des classes et d'assurer une évaluation plus équitable des performances sur l'ensemble des émotions, reflétant ainsi plus fidèlement les conditions réelles d'utilisation.

En complément de l'évaluation quantitative, des techniques de visualisation et d'interprétation sont utilisées afin d'analyser le comportement des modèles et de mieux comprendre les décisions produites. La visualisation des matrices de confusion permet d'identifier les confusions fréquentes entre certaines émotions, tandis que les méthodes d'attention et les projections de caractéristiques dans des espaces de dimension réduite offrent un éclairage sur les régions faciales et les représentations internes les plus pertinentes pour la reconnaissance émotionnelle. Ces analyses qualitatives contribuent à une meilleure compréhension du fonctionnement des modèles et renforcent la confiance dans les systèmes développés, en particulier dans des contextes applicatifs sensibles.

4 l'État de l'art

La reconnaissance des expressions faciales suscite un intérêt croissant dans la recherche en raison de sa large gamme d'applications et de son importance pour l'interaction homme–machine. Un nombre toujours plus important de travaux s'attachent à développer des systèmes de FER efficaces en utilisant à la fois des méthodes traditionnelles et des techniques avancées d'apprentissage profond. Les premières études sur la FER généraliste reposaient principalement sur des images statiques et sur des techniques d'extraction de caractéristiques traditionnelles telles que les Local Binary Patterns (LBP), les Histogrammes de Gradients Orientés (HOG), les filtres de Gabor et les descripteurs de points de repère géométriques, généralement combinés avec des classifieurs comme les machines à vecteurs de support (SVM) ou les perceptrons multicouches. Par exemple, Yang et al. [162] ont utilisé des histogrammes LBP avec un arbre de décision SVM hiérarchique pour reconnaître les expressions sur la base de données JAFFE, tandis que Dagher et al. [163] ont exploité des caractéristiques HOG avec un cadre SVM en trois étapes afin d'améliorer la discrimination entre des émotions visuellement similaires sur les bases de données JAFFE, CK+ et RaFD.

Bien que ces approches aient obtenu des performances satisfaisantes dans des conditions contrôlées, leur capacité limitée à capturer des variations faciales complexes, subtiles et non contraintes a réduit leur efficacité dans des scénarios réels.

L'avènement de l'apprentissage profond a conduit à un basculement vers des méthodes de reconnaissance des expressions faciales (FER) basées sur les réseaux de neurones convolutifs (CNN). Les CNN apprennent des caractéristiques spatiales hiérarchiques directement à partir des images et ont atteint d'excellents résultats sur des jeux de données de référence tels que FER2013, RAF-DB et AffectNet. Des architectures intégrant des mécanismes d'attention ou de régions d'intérêt ont encore amélioré les performances en se concentrant sur les zones du visage pertinentes pour l'expression, comme dans les réseaux basés sur les régions d'intérêt (ROI) ou inspirés des capsules [16, 167]. Néanmoins, les CNN standards capturent principalement des motifs locaux et peuvent être sensibles aux occlusions, aux variations de pose et d'illumination, en particulier dans des environnements non contraints. Afin de mieux modéliser la structure globale du visage, les Vision Transformers (ViT) ont récemment été adoptés pour la FER [18]. Les modèles basés sur ViT et les modèles hybrides CNN–ViT exploitent l'auto-attention pour capturer les dépendances à longue portée entre les patches d'image et ont atteint des performances à l'état de l'art sur plusieurs benchmarks. Des variantes de Transformers ont également été proposées pour la FER dynamique (DFER), où des Transformers spatiaux sont combinés avec des Transformers temporels ou des mécanismes d'attention afin de modéliser l'évolution des expressions au fil du temps. Des méthodes telles que Former-DFER [86] et les cadres de Transformers sensibles à l'intensité [20] ont obtenu des résultats compétitifs sur des jeux de données vidéo tels que DFEW, AFEW et FERV39K en modélisant explicitement l'information temporelle et contextuelle. Cependant, ces modèles s'accompagnent souvent d'un coût de calcul et de besoins en mémoire élevés, ce qui limite leur adéquation à des déploiements en temps réel ou sur des systèmes embarqués.

Plus récemment, l'émergence des modèles de vision et langage (VLM), en particulier CLIP, a ouvert de nouvelles perspectives pour la reconnaissance des expressions faciales (FER) en alignant les caractéristiques visuelles avec des descriptions textuelles sémantiques, ce qui améliore la robustesse et les performances de reconnaissance [23, 184, 185]. Parallèlement, l'apprentissage par transfert et les méthodes de réglage fin à paramètres efficaces (PEFT) ont gagné en importance en tant que stratégies pour adapter de grands réseaux préentraînés à la FER. Les approches classiques d'apprentissage par transfert réutilisent des architectures de base de type CNN ou transformeur préentraînées via l'extraction de caractéristiques ou

un réglage fin complet ou partiel, mais la mise à jour d'un grand nombre de paramètres peut s'avérer coûteuse pour les modèles de grande échelle. Les méthodes PEFT, telles que les adaptateurs et le prompt tuning, répondent à ce problème en figeant la majorité des poids préentraînés et en n'apprenant que de petits modules spécifiques à la tâche, réduisant ainsi les besoins en mémoire et le coût d'entraînement tout en préservant les performances [242, 243].

Comparé à la reconnaissance d'expressions faciales (FER) générale, les travaux consacrés à la FER spécifique aux conducteurs restent relativement limités. Les premières méthodes se sont concentrées sur des descripteurs manuels ou sur des caractéristiques géométriques légères, adaptées à un déploiement sur des systèmes embarqués dans les véhicules. Par exemple, une forêt aléatoire hiérarchique pondérée, basée sur des points de repère faciaux, a atteint une précision compétitive sur CK+, MMI et KMU-FED sans accélération GPU, démontrant que des algorithmes efficaces peuvent être adaptés aux plateformes automobiles [103]. Une forêt aléatoire multicouche ultérieure a encore amélioré la vitesse d'inférence tout en conservant de bonnes performances [65]. Des approches plus récentes ont intégré l'apprentissage profond, notamment des modèles à base de CNN, des architectures hybrides CNN–RNN, des cadres pilotés par l'optimisation et des réseaux spatio-temporels évalués sur des jeux de données tels que KMU-FED, DMD, ViVa-Face et DrivFace [111, 194, 195]. Bien que certaines de ces méthodes rapportent une grande précision sur des jeux de données de conduite contrôlés ou semi-contrôlés, leur robustesse face à la variabilité en conditions réelles et leur compatibilité avec des contraintes computationnelles strictes restent des questions ouvertes.

Dans l'ensemble, la littérature met en évidence des progrès significatifs en FER, mais des défis majeurs persistent. Les modèles existants présentent souvent des performances limitées lorsqu'ils passent de jeux de données contrôlés à des scénarios in the wild ou de conduite, en particulier en présence de fortes variations d'illumination, d'occlusions et de changements de pose. Les architectures basées sur des transformers ou sur des modèles vision–langage tels que CLIP atteignent souvent une forte précision de reconnaissance, mais elles exigent des ressources de calcul importantes, ce qui constitue un défi majeur pour les déploiements embarqués ou en temps réel. En outre, il reste difficile de modéliser la dynamique temporelle de manière à la fois efficace et efficiente, car les réseaux récurrents, les réseaux à convolutions 3D ou les modèles vidéo basés sur des transformers peinent à capturer les transitions subtiles d'expressions et les dépendances à long terme. Ces limites soulignent la nécessité de méthodes de FER qui équilibrent mieux robustesse, efficacité

computationnelle et modélisation temporelle dans des environnements réalistes.

5 Contributions

Dans le prolongement des objectifs définis précédemment et afin de répondre à la problématique centrale de cette thèse, plusieurs contributions sont proposées. Elles s'appuient sur les fondements théoriques et méthodologiques présentés et s'inscrivent dans la continuité des travaux existants, tout en apportant des avancées ciblées pour la reconnaissance des expressions faciales en environnements réalistes, et en particulier dans des contextes contraints tels que la conduite automobile. Les contributions proposées peuvent être résumées comme suit:

5.1 Modélisation du contexte global pour la reconnaissance des expressions faciales du conducteur

La première contribution s'intéresse à l'amélioration de la reconnaissance des expressions faciales du conducteur en conditions de conduite réalistes. Les approches existantes reposent majoritairement sur des représentations locales du visage, qui peuvent s'avérer insuffisantes pour capturer la structure globale des expressions, en particulier en présence de variations de pose, d'illumination ou d'occlusions partielles, fréquemment observées dans l'habitacle d'un véhicule. Pour répondre à cette limitation, cette contribution propose une approche fondée sur la modélisation du contexte global du visage, nommée DFER-GCViT (A Global Context Vision Transformer for Driver Facial Expression Recognition) [28]. L'idée centrale consiste à exploiter des mécanismes d'attention globale capables de capturer des relations à longue portée entre les différentes régions faciales, afin d'obtenir une représentation plus cohérente et plus discriminante des expressions. Cette modélisation globale permet de mieux intégrer les informations spatiales pertinentes et d'améliorer la robustesse du système face aux perturbations propres aux environnements de conduite.

Les évaluations expérimentales réalisées sur des jeux de données dédiés à la conduite automobile montrent que l'approche DFER-GCViT améliore significativement les performances de reconnaissance par rapport aux méthodes basées sur des architectures convolutionnelles classiques. Cette contribution met ainsi en évidence l'intérêt de l'intégration du contexte

global pour la reconnaissance fiable des expressions faciales du conducteur en conditions réalistes.

5.2 Architecture légère et efficace pour la reconnaissance des expressions faciales en conduite

La deuxième contribution s'inscrit dans l'objectif de concilier performances de reconnaissance et efficacité computationnelle, un enjeu central pour le déploiement de systèmes de reconnaissance des expressions faciales dans des environnements embarqués. Les architectures récentes les plus performantes reposent souvent sur des modèles profonds et coûteux, difficiles à intégrer dans des systèmes automobiles soumis à des contraintes strictes de ressources et de temps réel. Afin de répondre à ces contraintes, cette contribution propose une architecture hybride légère, nommée ShuffViT-DFER (Shuffle Vision Transformer: Lightweight, Fast and Efficient Recognition of Driver Facial Expression) [29], combinant des réseaux convolutifs efficaces avec des mécanismes d'attention inspirés des Vision Transformers. Cette approche vise à exploiter les avantages complémentaires des CNN pour l'extraction de caractéristiques locales et de vision transformer (ViT) s pour la modélisation globale, tout en réduisant significativement la complexité du modèle. Le design de l'architecture est spécifiquement orienté vers une exécution efficace sur des plateformes à ressources limitées.

Les résultats expérimentaux montrent que ShuffViT-DFER atteint des performances compétitives par rapport à des modèles plus lourds de l'état de l'art, tout en réduisant le coût computationnel et la consommation mémoire. Cette contribution démontre ainsi qu'il est possible de concevoir des modèles de reconnaissance des expressions faciales à la fois précis et adaptés aux contraintes des systèmes embarqués en contexte de conduite automobile.

5.3 Reconnaissance multimodale des expressions faciales du conducteur par modèles vision–langage

La troisième contribution explore l'intégration d'informations sémantiques afin d'améliorer la reconnaissance des expressions faciales du conducteur dans des environnements complexes. Les approches reposant uniquement sur l'information visuelle peuvent rencontrer des

difficultés face à des expressions ambiguës ou de faible intensité, fréquentes en conditions de conduite réelles, où les signaux visuels seuls ne suffisent pas toujours à une interprétation fiable. Pour dépasser ces limitations, cette contribution propose une approche multimodale basée sur les modèles vision–langage, nommée CLIVP-FER (Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment) [30]. L'approche s'appuie sur l'alignement entre des représentations visuelles du visage et des descriptions textuelles associées aux émotions, générées par ChatGPT, permettant d'enrichir la représentation des expressions faciales par une information sémantique complémentaire. Cette intégration vision–langage vise à renforcer la robustesse du système et à améliorer sa capacité de généralisation dans des conditions d'acquisition non contrôlées.

Les évaluations expérimentales réalisées sur des jeux de données de conduite automobile montrent que CLIVP-FER permet d'améliorer les performances de reconnaissance par rapport aux approches purement visuelles, en particulier dans des scénarios caractérisés par une forte variabilité environnementale. Cette contribution met en évidence l'intérêt des modèles vision–langage pour la reconnaissance des expressions faciales du conducteur en conditions réalistes.

5.4 Adaptation efficace en paramètres des modèles vision–langage pour la reconnaissance dynamique des expressions faciales

La quatrième contribution s'intéresse à la reconnaissance dynamique des expressions faciales en conditions réalistes, au-delà du seul contexte de la conduite automobile. Si les modèles vision–langage ont montré un fort potentiel pour la reconnaissance des émotions à partir d'images statiques, leur adaptation aux séquences vidéo pose des défis importants, notamment en termes de complexité computationnelle et de prise en compte des dynamiques temporelles. Pour répondre à ces enjeux, cette contribution propose un cadre d'adaptation efficace en paramètres, nommé PE-CLIP (A Parameter-Efficient Fine-Tuning of VLMs for Dynamic Facial Expression Recognition) [31], permettant d'exploiter les capacités des modèles vision–langage pour la reconnaissance dynamique des expressions faciales. L'approche repose sur l'adaptation partielle de modèles préentraînés à grande échelle, en limitant le nombre de paramètres entraînables tout en intégrant des mécanismes permettant de capturer l'évolution temporelle des expressions. Cette stratégie vise à préserver les connaissances sémantiques acquises lors du préentraînement, tout en assurant une adaptation efficace aux données vidéo.

Les expérimentations menées sur des jeux de données dynamiques acquis en conditions non contrôlées montrent que PE-CLIP atteint des performances compétitives par rapport aux méthodes de référence, tout en réduisant significativement le nombre de paramètres entraînables. Cette contribution met en évidence l'intérêt des approches d'adaptation efficace en paramètres pour concilier performance, efficacité et généralisation dans la reconnaissance dynamique des expressions faciales en environnements réalistes.

6 Conclusion et perspectives

Cette thèse s'est attachée à l'étude de la reconnaissance des expressions faciales en environnements réalistes, avec un intérêt particulier pour le contexte contraint de la conduite automobile. À travers l'analyse des limites des approches existantes et le développement de nouvelles méthodes, ce travail a mis en évidence l'importance de modéliser conjointement les informations spatiales globales, la dynamique temporelle des expressions et les contraintes d'efficacité propres aux systèmes embarqués.

Les contributions proposées ont permis d'explorer différentes stratégies visant à améliorer la robustesse et la fiabilité des systèmes de reconnaissance des expressions faciales, allant de la modélisation contextuelle globale à l'intégration d'approches multimodales vision–langage, en passant par la conception d'architectures légères et l'adaptation efficace de modèles préentraînés. Les résultats obtenus démontrent que ces approches constituent des pistes pertinentes pour faire face aux défis posés par les environnements non contrôlés et dynamiques.

Malgré ces avancées, plusieurs perspectives de recherche demeurent ouvertes. Elles concernent en premier lieu l'optimisation des modèles et l'amélioration de leur efficacité computationnelle, en vue d'un déploiement en temps réel sur des systèmes embarqués, ainsi que leur validation en conditions réelles d'utilisation. Par ailleurs, l'intégration multimodale, notamment à travers les modèles vision–langage, constitue une piste prometteuse pour enrichir l'interprétation des expressions faciales dans des situations complexes. Un autre enjeu important réside dans la généralisation des modèles face aux changements de domaine, appelant au développement de méthodes d'adaptation de domaine et d'apprentissage auto-supervisé. Enfin, la disponibilité limitée de jeux de données représentatifs, en particulier en contexte de conduite automobile, souligne la nécessité de disposer de bases plus diversifiées et proches des conditions réelles. Dans ce contexte, cette thèse contribue à poser les bases

de systèmes de reconnaissance des expressions faciales plus robustes, plus efficients et mieux adaptés aux applications réelles.

Bibliography

[1] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, 1872.

[2] Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, and Zhongxue Gan. A survey on facial expression recognition of static and dynamic emotions. *arXiv preprint arXiv:2408.15777*, 2024.

[3] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[4] Kathrin Kaulard, Douglas W Cunningham, Heinrich H Bülthoff, and Christian Wallraven. The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PloS one*, 7(3):e32321, 2012.

[5] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[6] Paul Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. 1994.

[7] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992.

[8] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[9] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022.

[10] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126:550–569, 2018.

[11] Ibtissam Saadi, Douglas W Cunningham, Taleb-Ahmed Abdelmalik, Abdenour Hadid, and Yassin El Hillali. Driver's facial expression recognition: A comprehensive survey. *Expert Systems with Applications*, page 122784, 2023.

[12] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pages 469–481. Springer, 2004.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[14] Ebenezer Owusu, Yongzhao Zhan, and Qi Rong Mao. A neural-adaboost based facial expression recognition system. *Expert Systems with Applications*, 41(7):3383–3390, 2014.

[15] Xiaorong Pu, Ke Fan, Xiong Chen, Luping Ji, and Zhihu Zhou. Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing*, 168:1173–1180, 2015.

[16] Xiao Sun, Shixin Zheng, and Hongshuai Fu. Roi-attention vectorized cnn model for static facial expression recognition. *IEEE Access*, 8:7183–7194, 2020.

[17] Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, and Masoumeh Zareapoor. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115:101–106, 2018.

[18] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual trans-

formers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, 2021.

[19] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. Facial expression recognition based on squeeze vision transformer. *Sensors*, 22(10):3729, 2022.

[20] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 67–75, 2023.

[21] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[22] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.

[23] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023.

[24] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. *arXiv preprint arXiv:2303.00193*, 2023.

[25] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[26] Mrinalini Patil and S Veni. Driver emotion recognition for enhancement of human machine interface in vehicles. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0420–0424. IEEE, 2019.

[27] Ibtissam Saadi, Douglas W Cunningham, Abdelmalik Taleb-Ahmed, Abdenour Hadid, and Yassin El Hillali. Driver's facial expression recognition: A comprehensive survey. *Expert Systems with Applications*, 242:122784, 2024.

[28] Ibtissam Saadi, Douglas W. Cunningham, Taleb-Ahmed Abdelmalik, Abdenour Hadid, and Yassin El Hillali. Driver's facial expression recognition using global context vision transformer. In *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, pages 1–8, 2023.

[29] Ibtissam Saadi, Douglas W Cunningham, Taleb-ahmed Abdelmalik, Abdenour Hadid, and Yassin El Hillali. Shuffle vision transformer: Lightweight, fast and efficient recognition of driver facial expression. *arXiv preprint arXiv:2409.03438*, 2024.

[30] Ibtissam Saadi, Abdenour Hadid, Douglas W. Cunningham, Abdelmalik Taleb-Ahmed, and Yassin El Hillali. Leveraging vision language models for facial expression recognition in driving environment. In Orhan Konak, Bert Arnrich, Gerald Bieber, Arjan Kuijper, and Sebastian Fudickar, editors, *Sensor-Based Activity Recognition and Artificial Intelligence*, pages 81–93, Cham, 2025. Springer Nature Switzerland.

[31] Ibtissam Saadi, Abdenour Hadid, Douglas W Cunningham, Abdelmalik Taleb-Ahmed, and Yassin El Hillali. Pe-clip: A parameter-efficient fine-tuning of vision language models for dynamic facial expression recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[32] Albert Mehrabian and Morton Wiener. Decoding of inconsistent communications. *Journal of personality and social psychology*, 6(1):109, 1967.

[33] *What's in a Face?: How Your Looks Betray Your Personality*. New scientist. Reed Busuness Information, 2009.

[34] André Chastel et al. *Leonardo on Art and the Artist*. Courier Corporation, 2002.

[35] Stephen Greenblatt. Toward a universal language of motion: Reflections on a seventeenth century muscle man. *LiNQ*, 21(2):56–62, 1994.

[36] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[37] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[38] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[39] Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[40] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8):5619–5627, 2022.

[41] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[42] Gautam Srivastava and Surajit Bag. Modern-day marketing concepts based on face recognition and neuro-marketing: a review and future research directions. *Benchmarking: An International Journal*, 31(2):410–438, 2024.

[43] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[44] Robert Plutchik. *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

[45] M Sunitha, T Adilakshmi, and Renuka. Emotion-based music recommendation system to generate a dynamic playlist. In *Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022*, pages 359–371. Springer, 2023.

[46] Iris Bakker, Theo Van Der Voordt, Peter Vink, and Jan De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current psychology*, 33:405–421, 2014.

[47] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[48] Gajendra Kumar, Tanaya Das, and Kuldeep Singh. Early detection of depression through facial expression recognition and electroencephalogram-based artificial intelligence-assisted graphical user interface. *Neural Computing and Applications*, 36(12):6937–6954, 2024.

[49] José Almeida and Fátima Rodrigues. Facial expression recognition system for stress detection with deep learning. In *ICEIS (1)*, pages 256–263, 2021.

[50] Bo Jin, Yue Qu, Liang Zhang, and Zhan Gao. Diagnosing parkinson disease through facial expression recognition: video analysis. *Journal of medical Internet research*, 22(7):e18697, 2020.

[51] Denys Fontaine, Valentin Vielzeuf, Philippe Genestier, Pascal Limeux, Serena Santucci-Sivilotto, Emmanuel Mory, Nelly Darmon, Michel Lanteri-Minet, May Mokhtar, Melanie Laine, et al. Artificial intelligence to evaluate postoperative pain based on facial expression recognition. *European Journal of Pain*, 26(6):1282–1291, 2022.

[52] Haochen Xu and Manhua Liu. A deep attention transformer network for pain estimation with facial expression video. In *Biometric Recognition: 15th Chinese Conference, CCBR 2021, Shanghai, China, September 10–12, 2021, Proceedings 15*, pages 112–119. Springer, 2021.

[53] Ashraf Abbas M Al-Modwahi, Onkemetse Sebetela, Lefoko Nehemiah Batleng, Behrang Parhizkar, and Arash Habibi Lashkari. Facial expression recognition intelligent security system for real time surveillance. In *Proc. of World Congress in Computer Science, Computer Engineering, and Applied Computing*, 2012.

[54] Shih-Chung Hsu, Hsin-Hui Huang, and Chung-Lin Huang. Facial expression recognition for human-robot interaction. In *2017 First IEEE International Conference on Robotic Computing (IRC)*, pages 1–7. IEEE, 2017.

[55] Fadi Dornaika and Bogdan Raducanu. Efficient facial expression recognition for human robot interaction. In *Computational and Ambient Intelligence: 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings 9*, pages 700–708. Springer, 2007.

[56] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 2021.

[57] Weiqing Wang, Kunliang Xu, Hongli Niu, and Xiangrong Miao. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. *Complexity*, 2020:1–9, 2020.

[58] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. Towards personalised gaming via facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 10, pages 30–36, 2014.

[59] Wanghua Deng and Ruoxue Wu. Real-time driver-drowsiness detection system using facial features. *Ieee Access*, 7:118727–118738, 2019.

[60] Sajid Ali Khan, Shariq Hussain, Sun Xiaoming, and Shunkun Yang. An effective framework for driver fatigue recognition based on intelligent facial expressions analysis. *IEEE Access*, 6:67459–67468, 2018.

[61] Shruti Mohanty, Shruti V Hegde, Supriya Prasad, and J Manikandan. Design of real-time drowsiness detection system using dlib. In *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 1–4. IEEE, 2019.

[62] CV Hari and Praveen Sankaran. Driver distraction analysis using face pose cues. *Expert Systems with Applications*, 179:115036, 2021.

[63] Hua Gao, Anil Yüce, and Jean-Philippe Thiran. Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5961–5965. IEEE, 2014.

[64] Wenbo Li, Guanzhong Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, Dongpu Cao, and Fei-Yue Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Transactions on Computational Social Systems*, 9(3):667–678, 2021.

[65] Mira Jeong, Jaeyeal Nam, and Byoung Chul Ko. Lightweight multilayer random forests for monitoring driver emotional status. *Ieee Access*, 8:60344–60354, 2020.

[66] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[67] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[68] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[69] Jin Zhang, Xue Mei, Huan Liu, Shenqiang Yuan, and Tiancheng Qian. Detecting negative emotional stress based on facial expression in real time. In *2019 IEEE 4th international conference on signal and image processing (ICSIP)*, pages 430–434. IEEE, 2019.

[70] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.

[71] Chirag Dalvi, Manish Rathod, Shruti Patil, Shilpa Gite, and Ketan Kotecha. A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions. *Ieee Access*, 9:165806–165840, 2021.

[72] Wenfei Gu, Cheng Xiang, YV Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern recognition*, 45(1):80–91, 2012.

[73] SL Happy, Anjith George, and Aurobinda Routray. A real time facial expression classification system using local binary patterns. In *2012 4th International conference on intelligent human computer interaction (IHCI)*, pages 1–5. IEEE, 2012.

[74] Deepak Ghimire, Sunghwan Jeong, Joonwhoan Lee, and San Hyun Park. Facial expression recognition based on local region specific features and support vector

machines. *Multimedia Tools and Applications*, 76:7803–7821, 2017.

[75] Yiming Wang, Hui Yu, Brett Stevens, and Honghai Liu. Dynamic facial expression recognition using local patch and lbp-top. In *2015 8th International conference on human system interaction (HSI)*, pages 362–367. IEEE, 2015.

[76] Pranav Kumar, SL Happy, and Aurobinda Routray. A real-time robust facial expression recognition system using hog features. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pages 289–293. IEEE, 2016.

[77] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013.

[78] Timothy F. Cootes, Gareth J. Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[79] Farkhod Makhmudkhujaev, Mohammad Abdullah-Al-Wadud, Md Tauhid Bin Iqbal, Byungyong Ryu, and Oksam Chae. Facial expression recognition with local prominent directional pattern. *Signal Processing: Image Communication*, 74:1–12, 2019.

[80] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

[81] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

[82] Ankit S Vyas, Harshadkumar B Prajapati, and Vipul K Dabhi. Survey on face expression recognition using cnn. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 102–106. IEEE, 2019.

[83] Ruicong Zhi and Ming Wan. Dynamic facial expression feature learning based on sparse rnn. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 1373–1377. IEEE, 2019.

[84] Rajesh Singh, Sumeet Saurav, Tarun Kumar, Ravi Saini, Anil Vohra, and Sanjay Singh. Facial expression recognition in videos using hybrid cnn & convlstm. *International Journal of Information Technology*, 15(4):1819–1830, 2023.

[85] Tzuu-Hseng S Li, Ping-Huan Kuo, Ting-Nan Tsai, and Po-Chien Luan. Cnn and lstm based facial expression analysis model for a humanoid robot. *IEEE Access*, 7:93998–94011, 2019.

[86] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021.

[87] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[88] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017.

[89] Kuang Liu, Mingmin Zhang, and Zhigeng Pan. Facial expression recognition with cnn ensemble. In *2016 international conference on cyberworlds (CW)*, pages 163–166. IEEE, 2016.

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

[91] Deepanshu Khanna, Neeru Jindal, Prashant Singh Rana, and Harpreet Singh. Enhanced spatio-temporal 3d cnn for facial expression classification in videos. *Multimedia Tools and Applications*, 83(4):9911–9928, 2024.

[92] Vinay Kukreja, Abhinav Kumar, et al. Facial expression recognition with cnn-svm for emotional state classification. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–5. IEEE, 2023.

[93] Ayoub Sassi, Wael Jaafar, Safa Cherif, Jihene Ben Abderrazak, and Halim

Yanikomeroglu. Video traffic analysis for real-time emotion recognition and visualization in online learning. *IEEE Access*, 11:99376–99386, 2023.

[94] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[95] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

[96] Shan Li, Weihong Deng, and JunPing 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*, pages 2852–2861, 2017.

[97] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[98] Hajer Guerdelli, Claudio Ferrari, Walid Barhoumi, Haythem Ghazouani, and Stefano Berretti. Macro-and micro-expressions facial datasets: A survey. *Sensors*, 22(4):1524, 2022.

[99] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[100] Daniel Lundqvist, Anders Flykt, and Arne Öhman. Karolinska directed emotional faces. *PsycTESTS Dataset*, 91:630, 1998.

[101] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for

facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.

[102] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

[103] Mira Jeong and Byoung Chul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12):4270, 2018.

[104] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012.

[105] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022.

[106] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios. *IEEE Transactions on Affective Computing*, 2021.

[107] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 387–405. Springer, 2020.

[108] Sourabh Vora, Akshay Rangesh, and Mohan M Trivedi. On generalizing driver gaze zone estimation using convolutional neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854. IEEE, 2017.

[109] Katerine Diaz-Chito, Aura Hernández-Sabaté, and Antonio M López. A reduced

feature set for driver head pose estimation. *Applied Soft Computing*, 45:98–107, 2016.

[110] Harshit Varma, Nagarajan Ganapathy, and Thomas M Deserno. Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning. In *Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications*, volume 12037, pages 57–63. SPIE, 2022.

[111] Bindu Verma and Ayesha Choudhary. A framework for driver emotion recognition using deep learning and grassmann manifolds. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1421–1426. IEEE, 2018.

[112] Ashish Tiwari. Supervised learning: From theory to applications. In *Artificial intelligence and machine learning for EDGE computing*, pages 23–32. Elsevier, 2022.

[113] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[114] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[115] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[116] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[117] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[118] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[119] Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):1–20, 2021.

[120] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in

medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

[121] Ammar Almomani, Khalid Nahar, Mohammad Alauthman, Mohammed Azmi Al-Betar, Qussai Yaseen, and Brij B Gupta. Image cyberbullying detection and recognition using transfer deep machine learning. *International Journal of Cognitive Computing in Engineering*, 5:14–26, 2024.

[122] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105, 2012.

[123] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[124] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[125] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[126] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[127] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[128] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[129] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint*

arXiv:1406.1078, 2014.

[130] Ricardo Dintén and Marta Zorrilla. Design, building and deployment of smart applications for anomaly detection and failure prediction in industrial use cases. *Information*, 15(9):557, 2024.

[131] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[132] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[133] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[134] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[135] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[136] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021.

[137] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,

pages 17980–17989, 2022.

[138] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[139] Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Weiming Hu, Xiaohu Qie, and Jianping Wu. Vilem: Visual-language error modeling for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11018–11027, 2023.

[140] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

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

[142] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *European Conference on Computer Vision (ECCV)*, 2020.

[143] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[144] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Align: Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 4, 2021.

[145] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss

for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[146] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[147] Liunian Harold Li, Mark Yatskar, D Yin, CJ Hsieh, and KW Chang. Visualbert: A simple and performant baseline for vision and language. arxiv 2019. *arXiv preprint arXiv:1908.03557*, 3, 1908.

[148] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[149] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[150] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[151] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[152] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[153] Fuzhen Zhuang, Zhicong Qi, Keng Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He, and Zhongzhi Shi. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[154] Chao Chen, Nor Ashidi Mat Isa, and Xin Liu. A review of convolutional neural network based methods for medical image classification. *Computers in Biology and Medicine*, 185:109507, 2025.

[155] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[156] Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 1–16, 2022.

[157] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[158] Haoxuan Zhang, Mingyu Wang, Xiang Xie, and Mingkui Tan. Prompt tuning: A review of methods and applications in vision and language models. In *ACM Computing Surveys*, 2023.

[159] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[160] Hamit Soyel and Hasan Demirel. Localized discriminative scale invariant feature transform based facial expression recognition. *Computers & Electrical Engineering*, 38(5):1299–1309, 2012.

[161] Deepak Ghimire and Joonwhoan Lee. Extreme learning machine ensemble using bagging for facial expression recognition. *Journal of Information Processing Systems*, 10(3):443–458, 2014.

[162] Mei-xia Yang, Shu-mao Zheng, and Yang Li. Facial expression recognition based on lbp and svm decision tree. In *2015 International Industrial Informatics and Computer Engineering Conference*, pages 651–655. Atlantis Press, 2015.

[163] Issam Dagher, Elio Dahdah, and Morshed Al Shakik. Facial expression recognition using three-stage support vector machines. *Visual Computing for Industry, Biomedicine, and Art*, 2:1–9, 2019.

[164] Hajar Chouhayebi, Jamal Riffi, Mohamed Adnane Mahraz, Ali Yahyaouy, and Hamid Tairi. Facial expression recognition using machine learning. In *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–6. IEEE, 2021.

[165] Huilin Ge, Zhiyu Zhu, Yuewei Dai, Biao Wang, and Xuedong Wu. Facial expression recognition based on deep learning. *Computer Methods and Programs in Biomedicine*, 215:106621, 2022.

[166] Xiaoming Zhao, Xugan Shi, and Shiqing Zhang. Facial expression recognition via deep learning. *IETE technical review*, 32(5):347–355, 2015.

[167] Tayyip Ozcan and Alper Basturk. Static facial expression recognition using convolutional neural networks based on transfer learning and hyperparameter optimization. *Multimedia Tools and Applications*, 79(35):26587–26604, 2020.

[168] Xing Jin, Xulin Song, Xiyin Wu, and Wenzhu Yan. Transformer embedded spectral-based graph network for facial expression recognition. *International Journal of Machine Learning and Cybernetics*, 15(6):2063–2077, 2024.

[169] Chunlei Li, Xiao Li, Xueping Wang, Di Huang, Zhoufeng Liu, and Liang Liao. Fg-agr: Fine-grained associative graph representation for facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):882–896, 2023.

[170] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 445–450, New York, NY, USA, 2016. Association for Computing Machinery.

[171] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recogni-

tion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17958–17968, 2023.

[172] Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. Multi-modal recurrent attention networks for facial expression recognition. *IEEE Transactions on Image Processing*, 29:6977–6991, 2020.

[173] Haposan Vincentius Manalu and Achmad Pratama Rifai. Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intelligent Systems with Applications*, 21:200339, 2024.

[174] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[175] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[176] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014.

[177] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[178] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[179] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[180] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[181] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yu-Ming Tang, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*, 2024.

[182] Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039, 2024.

[183] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6764–6772, 2024.

[184] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[185] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024.

[186] Junliang Zhang, Xu Liu, Yu Liang, Xiaole Xian, Weicheng Xie, Linlin Shen, and Siyang Song. Clip-guided bidirectional prompt and semantic supervision for dynamic facial expression recognition. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024.

[187] Weihong Ren, Yu Gao, Zhi Han, Zhiyong Wang, Jiaole Wang, Honghai Liu, et al. Facial expression monitoring via fine-grained vision-language alignment. *IEEE Transactions on Automation Science and Engineering*, 2024.

[188] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2301–2310, 2024.

[189] Afizan Azman, Kirbana Jai Raman, Imran Artwel Junior Mhlanga, Siti Zainab Ibrahim, Sumendra Yogarayan, Mohd Fikri Azli Abdullah, Siti Fatimah Abdul Razak, Anang Hudaya Muhamad Amin, and Kalaiarasi Sonai Muthu. Real time driver anger detection. In *Information Science and Applications 2018: ICISA 2018*, pages 157–167. Springer, 2019.

[190] He-xuan Hu, Lu-feng Zhang, Tian-jin Yang, and Qiang Hu. A lightweight two-stream model for driver emotion recognition. In *Journal of Physics: Conference Series*, volume 2400, page 012002. IOP Publishing, 2022.

[191] Khalid Zaman, Zhaoyun Sun, Sayyed Mudassar Shah, Muhammad Shoaib, Lili Pei, and Altaf Hussain. Driver emotions recognition based on improved faster r-cnn and neural architectural search network. *Symmetry*, 14(4):687, 2022.

[192] Goutam Kumar Sahoo, Jayakrishna Ponduru, Santos Kumar Das, and Poonam Singh. Deep leaning-based facial expression recognition in fer2013 database: An in-vehicle application. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2022.

[193] Mariya Tauqeer, Saddaf Rubab, Muhammad Attique Khan, Rizwan Ali Naqvi, Kashif Javed, Abdullah Alqahtani, Shtwai Alsubai, and Adel Binbusayyis. Driver's emotion and behavior classification system based on internet of things and deep learning for advanced driver assistance system (adas). *Computer Communications*, 194:258–267, 2022.

[194] SS Sudha and SS Suganya. On-road driver facial expression emotion recognition with parallel multi-verse optimizer (pmvo) and optical flow reconstruction for partial occlusion in internet of things (iot). *Measurement: Sensors*, 26:100711, 2023.

[195] Yucheng Shang, Mutian Yang, Jianwei Cui, Linwei Cui, Zizheng Huang, and Xiang Li. Driver emotion and fatigue state detection based on time series fusion. *Electronics*,

12(1):26, 2023.

[196] Deepak Kumar Jain, Ashit Kumar Dutta, Elena Verdú, Shtwai Alsubai, and Abdul Rahaman Wahab Sait. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image and Vision Computing*, 133:104659, 2023.

[197] Suparshya Babu Sukhavasi, Susrutha Babu Sukhavasi, Khaled Elleithy, Ahmed El-Sayed, and Abdelrahman Elleithy. A hybrid model for driver emotion detection using feature fusion approach. *International journal of environmental research and public health*, 19(5):3085, 2022.

[198] Susrutha Babu Sukhavasi, Suparshya Babu Sukhavasi, Khaled Elleithy, Ahmed El-Sayed, and Abdelrahman Elleithy. Deep neural network approach for pose, illumination, and occlusion invariant driver emotion detection. *International Journal of Environmental Research and Public Health*, 19(4):2352, 2022.

[199] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023.

[200] Huafei Xiao, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. On-road driver emotion recognition using facial expression. *Applied Sciences*, 12(2):807, 2022.

[201] Chang Liu, Kaoru Hirota, and Yaping Dai. Patch attention convolutional vision transformer for facial expression recognition with occlusion. *Information Sciences*, 619:781–794, 2023.

[202] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021.

[203] Zhijie Huang, Yuezhao Yu, and Chao Gou. Driver facial expression recognition based on vit and stargan. In *2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 254–257. IEEE, 2021.

[204] Yunsheng Ma and Ziran Wang. Vit-dd: Multi-task vision transformer for semi-supervised driver distraction detection. In *IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, 2023.

[205] Ali Hatamizadeh, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. *arXiv preprint arXiv:2206.09959*, 2022.

[206] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[207] Goutam Kumar Sahoo, Santos Kumar Das, and Poonam Singh. Deep learning-based facial emotion recognition for driver healthcare. In *2022 National Conference on Communications (NCC)*, pages 154–159. IEEE, 2022.

[208] Wentao Hua, Fei Dai, Liya Huang, Jian Xiong, and Guan Gui. Hero: Human emotions recognition for realizing intelligent internet of things. *IEEE Access*, 7:24321–24332, 2019.

[209] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021.

[210] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[211] Lie Yang, Haohan Yang, Bin-Bin Hu, Yan Wang, and Chen Lv. A robust driver emotion recognition method based on high-purity feature separation. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[212] Anwer Mustafa Hilal, Dalia H Elkamchouchi, Saud S Alotaibi, Mohammed Maray, Mahmoud Othman, Amgad Atta Abdelmageed, Abu Sarwar Zamani, and Mohamed I Eldesouki. Manta ray foraging optimization with transfer learning driven facial emotion recognition. *Sustainability*, 14(21):14308, 2022.

[213] Yueh-Lin Wu, Hsin-Yun Tsai, Yi-Chi Huang, and Bo-Hao Chen. Accurate emotion

recognition for driving risk prevention in driver monitoring system. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 796–797. IEEE, 2018.

[214] Xiaobo Chen, Jian Du, Fuwen Deng, and Feng Zhao. Transferable driver facial expression recognition based on joint discriminative correlation alignment network with enhanced feature attention. *IET Intelligent Transport Systems*, 2023.

[215] Goutam Kumar Sahoo, Santos Kumar Das, and Poonam Singh. Performance comparison of facial emotion recognition: A transfer learning-based driver assistance framework for in-vehicle applications. *Circuits, Systems, and Signal Processing*, pages 1–28, 2023.

[216] Pawan Wawage and Yogesh Deshpande. Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. *International Journal of Performability Engineering*, 18(11):791, 2022.

[217] Alessandro Leone, Andrea Caroppo, Andrea Manni, and Pietro Siciliano. Vision-based road rage detection framework in automotive safety applications. *Sensors*, 21(9):2942, 2021.

[218] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023.

[219] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[220] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[221] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[222] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[223] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11998–12008, 2022.

[224] Yahui Nan, Jianguo Ju, Qingyi Hua, Haoming Zhang, and Bo Wang. A-mobilenet: An approach of facial expression recognition. *Alexandria Engineering Journal*, 61(6):4435–4444, 2022.

[225] Yanju Liu, Yange Li, Xinhai Yi, Zuojin Hu, Huiyu Zhang, and Yanzhong Liu. Lightweight vit model for micro-expression recognition enhanced by transfer learning. *Frontiers in Neurorobotics*, 16:922761, 2022.

[226] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2):281–305, 2012.

[227] Yuanyuan Liu, Jiabei Zeng, Shiguang Shan, and Zhuo Zheng. Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 458–465. IEEE, 2018.

[228] Vijayalakshmi GV Mahesh, Chengji Chen, Vijayarajan Rajangam, Alex Noel Joseph Raj, and Palani Thanaraj Krishnan. Shape and texture aware facial expression recognition using spatial pyramid zernike moments and law's textures feature set. *IEEE Access*, 9:52509–52522, 2021.

[229] Tarim Dar, Ali Javed, Sami Bourouis, Hany S Hussein, and Hammam Alshazly. Efficient-swishnet based system for facial emotion recognition. *IEEE Access*, 10:71311–71328, 2022.

[230] Tehmina Kalsum, Zahid Mehmood, et al. A novel lightweight deep convolutional neural network model for human emotions recognition in diverse environments.

Journal of Sensors, 2023, 2023.

[231] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[232] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition. *arXiv preprint arXiv:2310.16640*, 2023.

[233] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024.

[234] Joyati Chattopadhyay, Souvik Kundu, Arpita Chakraborty, and Jyoti Sekhar Banerjee. Facial expression recognition for human computer interaction. *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India*, pages 1181–1192, 2020.

[235] Rohan Wadhawan and Tapan K Gandhi. Landmark-aware and part-based ensemble transfer learning network for static facial expression recognition from images. *IEEE transactions on artificial intelligence*, 4(2):349–361, 2022.

[236] Sumeet Saurav, Prashant Gidde, Ravi Saini, and Sanjay Singh. Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *The Visual Computer*, 38(3):1083–1096, 2022.

[237] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

[238] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16085–16093, 2024.

[239] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[240] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, 2023.

[241] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6540–6548, 2024.

[242] Shreyank N Gowda, Boyan Gao, and David A Clifton. Fe-adapter: Adapting image-based emotion classifiers to videos. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2024.

[243] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, 2024.

[244] Bokyeung Lee, Hyunuk Shin, Bonhwa Ku, and Hanseok Ko. Frame level emotion guided dynamic facial expression recognition with emotion grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5681–5691, 2023.

[245] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195, 2022.

[246] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[247] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages

19113–19122, 2023.

[248] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[249] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

[250] Xin Zhou, Dingkang Liang, Wei Xu, Xingkui Zhu, Yihan Xu, Zhikang Zou, and Xiang Bai. Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14707–14717, 2024.

[251] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[252] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[253] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[254] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.

[255] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[256] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.

[257] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[258] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17958–17968, 2023.

[259] Min Li, Xiaoqin Zhang, Chenxiang Fan, Tangfei Liao, and Guobao Xiao. Dual-sti: Dual-path spatial-temporal interaction learning for dynamic facial expression recognition. *Information Sciences*, page 120953, 2024.

[260] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022.

[261] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022.

[262] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022.

[263] Zeng Tao, Yan Wang, Zhaoyu Chen, Boyang Wang, Shaoqi Yan, Kaixun Jiang, Shuyong Gao, and Wenqiang Zhang. Freq-hd: An interpretable frequency-based high-dynamics affective clip selection method for in-the-wild facial expression recognition in videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 843–852, 2023.

[264] Fuyan Ma, Bin Sun, and Shutao Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[265] Yitao Han and Qiming Li. Dflm: A dynamic facial-language model based on clip. In *2024 9th International Conference on Intelligent Computing and Signal Processing*

(ICSP), pages 1132–1137, 2024.