CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Learning Domain-Invariant Features for Facial Expression Recognition Using Unsupervised Domain Adaptation

by

Atteq u Rehman

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Software Engineering

2025

CERTIFICATE OF APPROVAL

Learning Domain-Invariant Features for Facial Expression Recognition Using Unsupervised Domain Adaptation

by

Atteq u Rehman

(MAI241004)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ayyaz Hussain	QAU, Islamabad
(b)	Internal Examiner	Dr. Mohmmad Masroor Ahmed	CUST, Islamabad

---

Dr. M. Abdul Qadir
Thesis Supervisor
December, 2025

Dr. Nadeem Anjum
Head
Dept. of Software Engineering
December, 2025

Dr. M. Abdul Qadir
Dean
Faculty of Computing
December, 2025

Author's Declaration

I, **Atteq u Rehman** hereby state that my MS thesis titled "**Learning Domain-Invariant Features for Facial Expression Recognition Using Unsupervised Domain Adaptation** " is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Atteq u Rehman)

Registration No:MAI241004

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Learning Domain-Invariant Features for Facial Expression Recognition Using Unsupervised Domain Adaptation**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Atteq u Rehman)

Registration No: MAI241004

Acknowledgement

First and foremost, I express my deepest gratitude to Allah (S.W.T) for blessing me with the knowledge, strength, courage, and patience to complete this research. His guidance has been my constant companion throughout this journey. To my loving parents, whose unwavering support and encouragement have been the bedrock of my journey; to "Fay" and my cherished friends, whose laughter and companionship have illuminated even the most challenging days; and to my esteemed supervisor Dr. M. Abdul Qadir , whose invaluable guidance and mentorship have shaped this work into a milestone. I also take a moment to appreciate myself. Reflecting on this journey, I am thankful for my resilience, staying positive, maintaining self-belief in tough times, and balancing research with personal responsibilities. These efforts have been crucial in overcoming obstacles and achieving this milestone. Finally, I extend my sincere gratitude to everyone who has supported me. This thesis reflects the collective encouragement and wisdom I've received.

(Atteq u Rehman)

Abstract

Facial Expressions are one of the primary sources of understanding human feelings and emotions. Machine learning model's ability to detect emotional states can be beneficial in many real-life applications. Performance of these models decreases in the case of Cross Domain validation. The issue of cross-domain facial expression recognition, where a model trained on one particular dataset failed to generalize to other domains or real-life settings because of the discrepancies in lighting, pose, and image quality, needed to be resolved. This study aims to develop a resilient model that focus on learning domain in-variant features to improves Cross-Domain Facial Expressions Recognition (CD-FER) accuracy when the training and testing datasets are sourced from different origins.

To achieve this, we have introduced a novel model called ResEmoteNet-DANN. This model combines Domain-Adversarial Neural Networks (DANN), Squeeze-and-Excitation (SE) blocks, and residual connections to extract features that stay consistent and gives better classification accuracy across different domains. Benchmark datasets such as FER2013, AffectNet, JAFFE, and ExpW were used to test this model. The findings demonstrate that although traditional models attain high in-domain accuracy (e.g.,79% on FER2013 and 72% on AffectNet), they suffer greatly in cross-domain situations, where accuracy falls to 32% and 38%, respectively. By contrast, our ResEmoteNet-DANN achieves a significant improvement over the baseline, achieving 58.04% (FER2013→AffectNet) and 56.8% (AffectNet→FER2013). Additionally, the proposed model outperforms the standard DANN 53.17% and provides competitive results against the SOTA AGRA model 65.03% in the difficult AffectNet→ExpW adaptation, achieving 62.9% while maintaining a significantly lower model complexity. The model outperforms SOTA binary cross-domain emotion classification for EM-UDA (74.55%) in the A→F emotion recognition with an accuracy of 80.72%. These findings show how well domain-adversarial learning works to create generalizable FER systems, offering a scalable solution for practical uses. These results demonstrate that Unsupervised Domain Adaptation(UDA) with adversarial learning is an effective way to develop

generalizable FER systems. In the future, further research can be done on multimodal and temporal FER and recognition of emotion intensities and complex expressions to take affective computing further.

Contents

List of Figures

List of Tables

Abbreviations

AAS	Active Assessment Strategy
A→E	AffectNet to ExpW Domain Transfer
A→F	AffectNet to FER2013 Domain Transfer
AGRA	Adversarial Graph Representation Alignment
AI	Artificial Intelligence
AU	Action Unit
AUROC	Area Under the Receiver Operating Characteristic Curve
CD-DCT	Cross-Domain Dynamic Class Threshold
CD-FER	Cross-Domain Facial Expression Recognition
CK+	Extended Cohn–Kanade Dataset
CNN	Convolutional Neural Network
DANN	Domain-Adversarial Neural Network
DCNN	Deep Convolutional Neural Network
ExpW	Expression in-the-Wild Dataset
F→A	FER2013 to AffectNet Domain Transfer
F→J	FER2013 to JAFFE Domain Transfer
FER	Facial Expression Recognition
FER2013	Facial Expression Recognition 2013 Dataset
GCN	Graph Convolutional Network
GRL	Gradient Reversal Layer
HMM	Hidden Markov Model
JAFFE	Japanese Female Facial Expression Dataset
KNN	K-Nearest Neighbors
KSDA	Kernel Subclass Discriminant Analysis

LA-CMFER	Local–Adaptive Cross-Modal Facial Expression Recognition
LBP	Local Binary Patterns
LR	Learning Rate
ML	Machine Learning
MSE	Mean Squared Error
OSH	Optimal Separating Hyperplane
PCA	Principal Component Analysis
RAF-DB	Real-world Affective Faces Database
RBF-SVM	Radial Basis Function Support Vector Machine
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue Color Channels
SE	Squeeze-and-Excitation
SGD	Stochastic Gradient Descent
SOTA	State of the Art
STN	Spatial Transformer Network
SVM	Support Vector Machine
UDA	Unsupervised Domain Adaptation
WCDA	Weighted Cross-Domain Alignment

Chapter 1

Introduction

1.1 Background and Motivation

Human communication is not just spoken words; it relies on both verbal and non-verbal cues to convey meaning effectively. Beyond just words, decades of research suggest that emotions are communicated more accurately through facial expressions, tone of voice, body language, and other non-verbal channels. Emotions have a fundamental effect on any human behavior and their decisions [1]. Historical studies of non-verbal communication have demonstrated that when verbal and non-verbal signals disagree, listeners tend to trust the non-verbal component, with only about 7% of the perceived message coming from words, 38% from vocal tone, and 55% from facial expression [2].

Some of the other studies also reinforces the value of facial cues, Sauter's review found that emotions such as happiness and amusement are recognized more reliably from smiles and laughter and that this holds across many cultures [3]. Ekman, Sorenson, and Friesen likewise reported high cross-cultural agreement when people identified six basic emotions (illustrated in Figure 1.1) happiness, fear, disgust, anger, surprise, and sadness from still images, even in non-literate societies [4].

In facial expression research, emotions are often plotted in a circumplex of valence

arousal, where it (pleasant ↔ unpleasant) is represented on the horizontal axis and arousal (calm ↔ excited) is represented on the vertical axis (Figure 1.2) [5].

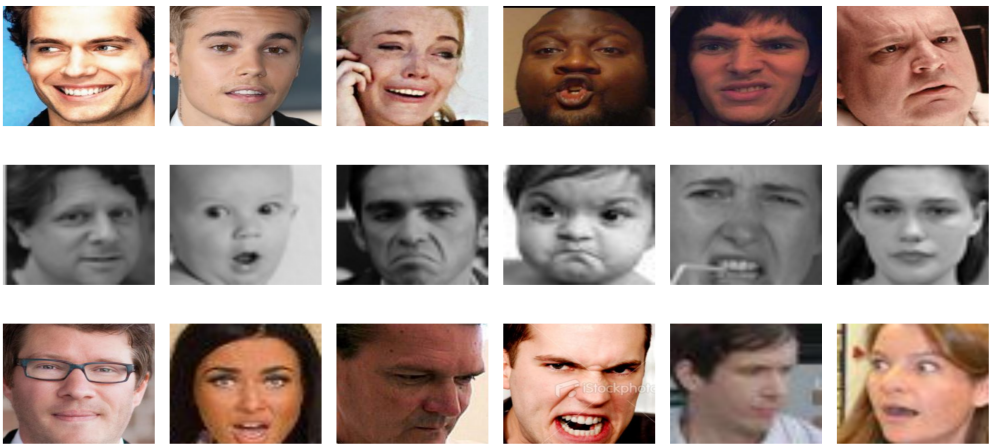

FIGURE 1.1: Six Basic Expressions Postulated by Ekman and Friesen
From Left to Right: Happy, Surprise, Sad, Anger, Disgust, Fear

Neutral faces cluster near the center, while the six basic emotions occupy predictable quadrants, splitting them into further positive versus negative categories. In addition to communication, Facial Expression Recognition (FER) is effective in various areas, as it directly affects most daily activities and becomes integral to the habits carried out by humans [6].

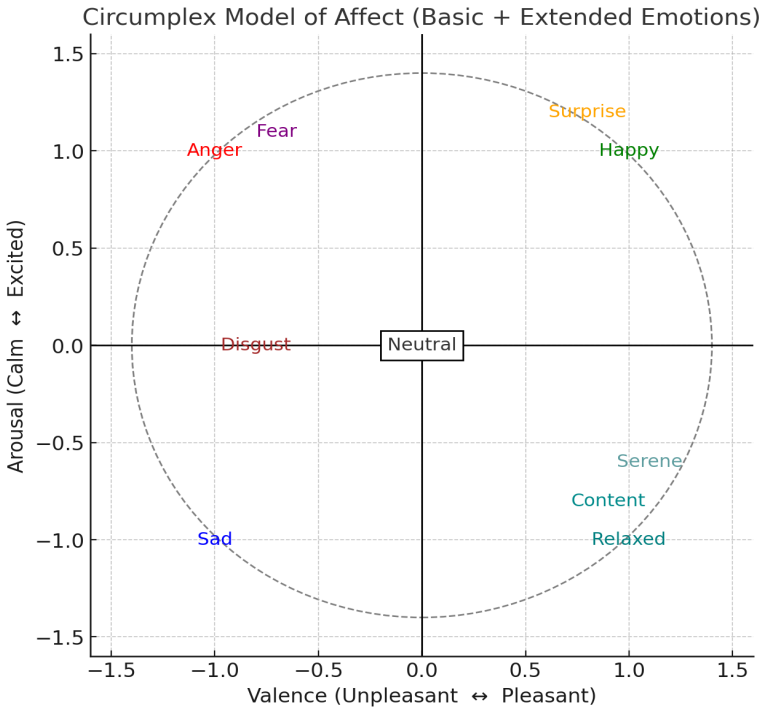

FIGURE 1.2: Emotions

1.1.1 Facial Expression Recognition and its Applications

"Facial Expression Recognition (FER) is a computer-vision process aimed at detecting and classifying human emotional expressions." Cîrneanu, Popescu & Iordache, 2023 [7].

Facial expressions are established to be crucial in communicating human emotions, FER has become increasingly integral to various real-world consumer applications, encompassing driver safety, healthcare, education and marketing, particularly in advanced driver-assistance systems. Facial cues, such as yawning, blinking, or prolonged eye closure, can be used to detect driver fatigue and drowsiness, enabling preventive alerts that improve road safety [8].

Many healthcare providers across the globe use FER as an indicator of emotional well-being, which is crucial for mental health assessments and timely interventions. Emotions such as sadness, stress, or pain can indicate underlying conditions or responses to treatment. For example, the help of FER enables healthcare providers to detect signs of discomfort or distress in patients, allowing for timely care [9]. Additionally, FER is also very helpful in diagnosing conditions like autism by analyzing facial-expression patterns that differ from those of Normal humans, which can reveal both elemental and compound emotions [10]. This application is particularly valuable in pediatric care, where non-verbal cues are crucial.

In educational environments, employing FER to recognize emotions such as interest, confusion, or boredom enables the assessment of emotional engagement and attention levels, thereby promoting customized learning experiences. For example, the real-time evaluation of student concentration in online learning settings or intelligent classrooms can improve teaching methodologies [11][12]. The FER-enabled Education Aids IoT System exemplifies how FER can be incorporated into online learning platforms to observe and respond to students' emotional states [13].

Being able to detect emotions is very useful in the Advertising and Marketing domains, as it can enhance customer interactions by analyzing consumer reactions to products or services. By analyzing customers' specific emotions, such as

happiness, surprise, or dissatisfaction, businesses can gather valuable insights into consumer preferences that can help to tailor their marketing strategies and enhance customer experiences. For example, FER-based sentiment analysis can aid in the development of products and advertising [14]. Having real-time feedback from facial expressions during product testing can provide direction in business planning and lead to improved customer satisfaction.

For security and law enforcement, FER detects micro-expressions to spot suspicious behavior; it can also detect emotions of fear or deception. For instance, the use of the Vibra image system, introduced to the airports, for the detection of people who are about to commit a crime by analyzing their micro-expressions to detect potential terror suspects promises to give security a boost [15]. This app also has a low-cost real-world application in life-or-death situations, where the quick identification of emotional cues could help avert danger.

Along with these, there are many other real-life applications of FER, from Virtual reality to Gaming. So, as we know, if we can teach machines how to read facial expressions and, based on that, classify the emotions of a person better than human accuracy, it will be very beneficial; for that particular purpose, there is an excellent history of research on automated FER, recognizing emotions through these automation systems can be categorized into two main parts of feature generation: classic machine-learning feature extraction and automatic extraction via deep neural networks [16].

1.1.2 Evolution of Automated FER Approaches

The field of Facial Expression Recognition (FER) has undergone substantial methodological advances during the last few decades. The first FER models used traditional machine-learning approaches which required human developers to create specific features by hand. The first approaches to facial expression analysis used Principal Component Analysis (PCA) and Local Binary Patterns (LBP) , Gabor filters , Hidden Markov Models (HMM) , K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) together with Viola–Jones detector for facial detection.

The approaches achieved excellent results on laboratory-controlled datasets because they operated under stable conditions of pose and lighting and occlusion. The need for manual feature extraction in these approaches made them less capable of handling real-world conditions.

With the emergence of high-quality GPUs with sufficient memory to process larger datasets, real progress began with deep-learning algorithms. Classic machine-learning algorithms require extensive preprocessing and manual feature extraction. However, deep-learning models can extract the most distinguishable features independently. They can make more accurate predictions, and since they are trained on a large amount of data, they are better at generalizing to real-world scenarios compared to classic machine-learning algorithms discussed earlier. The practical and successful application of deep Convolutional Neural Networks (CNNs) to general image-recognition tasks (e.g., AlexNet, VGG, ResNet on ImageNet) suggests that similar approaches could also rapidly learn expressive facial features automatically. Facial expression recognition (FER) research began actively utilizing CNN architecture approaches around 2013 and 2015, taking advantage of publicly available face-expression datasets with increased extension (FER2013, RAF-DB, AffectNet, etc.) [17].

CNNs (Convolutional Neural Networks) are rapidly established as the foundation of FER tasks. A CNN is capable of accepting raw images of pixels as input and automatically learning a hierarchy of feature filters (edges, textures, parts of faces, etc.) from those images that are most suitable for the expression-classification task. Local connectivity and weight sharing are the primary characteristics that enable CNNs to handle large image datasets efficiently. Such end-to-end learning has significantly lowered the reliance on manual pre-processing or feature extraction [18]. Therefore, if an ample dataset is available, CNN can learn the optimal features required for FER tasks. Initially proposed CNN architectures for FER were relatively simple, often consisting of just two convolutional layers (which apply learnable filters to generate feature maps highlighting patterns like edges), followed by pooling layers to reduce dimensionality and computational burden [19]. Finally, a fully connected layer performed the classification based on these

extracted features [19].

As the research progressed in CNN, new architectures emerged, and models were proposed explicitly for FER tasks. Additionally, transfer learning approaches were utilized. For example, an influential model was , which was initially a deep CNN for face recognition. That VGG-Face (a 37-layer VGG-based network) learned rich facial features and representations specific for face recognition, which turned out to be useful for expression recognition as well. Other CNN architectures, such as ResNet [20], GoogLeNet [21], transformers [22], and Vision Transformer (ViT) [23], also showed significant improvements for image-related tasks that were eventually useful for FER.

Interpreting human feelings by machine FER hereafter has, therefore, become a central problem in computer vision and machine learning [24].

1.2 Evolution of FER Datasets

Machine learning and deep-learning algorithms, as well as training structures, rely heavily on the dataset. To train a model for improved accuracy and precision, we need a proper dataset. There is also a brief history of the evolution of datasets. At first, datasets were created in a perfectly controlled lab environment, for example, Multi-PIE [25], which included the images from different facial postures, angles, and different lighting conditions; although the Multi-PIE database [25] extends the earlier PIE collection by recording multiple poses, view angles, and lighting conditions under laboratory control, it still suffers from two problems that limit its usefulness for modern deep learning. First, its scale is modest: recent neural scaling studies show that accuracy continues to rise with millions, rather than thousands, of training images [26][27][28]. Second, the data are almost noise-free and highly repetitive, so models that excel on Multi-PIE often generalize poorly to in-the-wild photographs containing occlusion, motion blur, or sensor artifacts [29][30][31].

A recent reevaluation reaches the same conclusion and recommends supplementing Multi-PIE with larger, more diverse datasets to achieve reliable real-world

performance [32]. Real-world scenarios can present the test element from different distributions, qualities, formats, and representation differences.

The Cohn–Kanade database, developed by Kanade et al. in 2000, offered a collection of image sequences of facial expressions, significantly facilitating temporal analysis of facial dynamics, which was extended as Extended Cohn-Kanade (CK+) in 2010 [33].

FER-2013, presented by Goodfellow et. al. in 2013 [17], is one of the pioneers in improving generalization to real-world scenarios due to its size, which features more than 35000 unrestricted images gathered from the internet [17].
Another significant improvement was Emotionet, created by Benitez-Quiroz et al. , which further supplemented this collection with approximately one million web sourced AU-annotated images [34]. A significant addition was AffectNet by Mollahoseini et al. (2019), which includes around one million images categorized by emotions in unrestricted web contexts [35]. Similarly, Li et al. (2017) introduced the RAF-DB dataset, concentrating on unconstrained real-world situations [36]. The ExpW dataset, proposed by Zhang et al. (2018), comprises diverse images from movies and web sources, thereby enhancing the cross-cultural applicability of FER models [37]. Despite these advancements, the transition from controlled laboratory datasets to large-scale in-the-wild collections continues to highlight the widening complexity of real-world FER tasks. Consequently, each new dataset has pushed the community toward models that are more robust, context-aware.

1.3 Domain Shift Problem in FER

FER models, such as EfficientFER [38], ResEmoteNet [39], EmoNeXt [40], Emo-FAN [41], and Deep-Residual Bi-LSTM Fusion [42], perform very well when both training and testing are conducted within the same domain. Still, these methods usually fail to generalize to data in other domains due to differences in lighting, pose, and quality, leading to a significant performance drop [29][30][31]. This type of performance degradation occurs when there is a distribution gap between the

source and target datasets; this distribution gap between domains is often referred to as domain shift. There are some primary reasons for the domain shift specific to the FER.

Variations in Lighting Conditions: Differences between indoor and outdoor scenes, or variations in light intensity, can make it difficult for models to interpret expressions related to features [43].

1.3.1 Differences in Camera Angles

The camera angle is crucial because it captures the perspective from which faces such as frontal faces, side profiles, or portraits—are recorded, and it can affect the consistency of expressions relevant to feature extraction [44].

1.3.2 Image Resolution Differences

Differences in image resolution can cause domain shift. High-resolution images contain more precise facial-expression details, aiding feature extraction, whereas low-resolution images may obscure details and lead to performance gaps [45].

An instant solution to mitigate this problem can be to train a new model for the target dataset. However, that approach demands a substantial amount of labeled data, and the annotation process is widely recognized as time-consuming, costly, and labor-intensive [46][47]. For every new domain, real-life scenario, fresh annotations are usually required before a model can generalize reliably in that domain [46].

1.3.3 Cultural Differences in Expressions

Although expressions are often considered similar worldwide, emotional displays can differ across cultures due to societal norms or habits. A model trained on a dataset from one cultural group may not perform as well on datasets from other groups [48].

1.4 Need for Unsupervised Domain Adaptation

As mentioned previously, the domain shift issue causes FER models to frequently experience a sharp decline in performance when implemented in new domains. Gathering and annotating a lot of labeled data for each new target domain would be an easy way to address this problem, but deep learning techniques require a lot of data, and this kind of annotation is generally acknowledged to be expensive, time-consuming, and unfeasible in many practical applications. [2, 46, 49]. This problem is particularly pronounced in FER, where emotion labels can be arbitrary and contingent on contextual and cultural elements, making it more challenging to get trustworthy supervision. [50–52].

Unsupervised Domain Adaptation (UDA) has therefore emerged as a promising approach for improving cross-domain robustness without requiring labeled data in the target domain. In UDA, labeled data from a source domain and unlabeled samples from a target domain are used to train a model with the goal of minimizing the distribution disparity between the two while maintaining discriminative performance. [46, 49, 53]. Typical UDA strategies include adversarial learning, where a domain discriminator is trained jointly with the feature extractor (e.g., Domain-Adversarial Neural Networks, DANN [54]), self-training with pseudo-labels on target data, and self-supervised contrastive learning frameworks such as SimCLR, which encourage compact, domain-invariant representations [55–57]. Recent work on cross-domain FER confirms that explicitly addressing domain shift leads to consistent gains over non-adaptive baselines. Improved generalization across datasets is reported using techniques based on learning and selection of global-local representation [51, 52], self-training and similarity transfer [56], AU-guided adaptation [58], and bi-directional fusion of active and stable information [59, 60].

When considered collectively, the latest research shows that UDA is not only helpful but also essential for developing FER systems that can function consistently in a variety of unlabeled real-world domains. [46, 49, 53, 59]. Nevertheless, UDA techniques frequently rely on sizable and complex deep neural architectures, many of

which have tens of millions of trainable parameters, leading to significant memory and computational overhead. For instance, recent research has shown that in order for standard UDA frameworks to be deployable in real-world scenarios, parameter counts must be reduced. [61]. For real-world FER deployment, where efficiency and low latency are crucial, such models are unfeasible due to their high GPU resource requirements and lengthy training times. The need for lightweight, computationally efficient UDA frameworks that can achieve strong cross-domain generalization without excessive parameter complexity is highlighted by the fact that these heavy models become impractical to train or deploy in resource-constrained environments like embedded driver-safety systems, mobile healthcare applications, or edge-based classroom monitoring [62]. Additionally, the challenges associated with computational overload often restrict the adaptability of traditional UDA pipelines when applied to continuous real-time FER tasks. This further emphasizes the importance of designing models that maintain strong representational capacity while operating under strict hardware limitations.

1.5 Problem Statement

Current FER models perform well when trained and validated on the same datasets, but suffer significant performance drops under cross-domain shifts in lighting, pose, culture, or resolution. Many of the UDA methods help reduce this disparity, but they still fall short in terms of accuracy, and some of the techniques consist of millions of parameters that strain memory and extend training duration.

1.6 Research Objectives

i. Learn Domain-Invariant Features: Develop a lightweight model that can distinguish emotion-related content from domain-specific differences (e.g., lighting, pose, image quality) [63][64]. This enables strong multi-source performance.

ii. Apply Adversarial Domain Adaptation: Use adversarial training with a domain discriminator so the feature extractor learns domain-invariant representations, reducing source–target discrepancy.

iii. Propose a Novel Model for Cross-Domain FER: Design a ResEmoteNet-inspired [39] architecture where SE blocks [65] attend over latent common features, and compare against standard SE placement and baselines.

iv. Evaluate the Model Across Different Domains: Assess performance on FER-2013, JAFFE, AffectNet-7, and ExpW using accuracy, confusion matrices, and component-wise analysis for understanding domain robustness.

1.7 Research Questions

i. How does a single-source FER model generalize to unseen data from a target domain, and to what extent can UDA techniques mitigate performance degradation due to domain shift?

ii. How does the proposed hybrid model perform on diverse facial expression datasets with varying lighting conditions, camera angles, and image capture settings?

iii. Can modifying the ResEmoteNet architecture by replacing a common or disentangled latent space with a simple CNN feature extractor within Squeeze-and-Excitation block achieve competitive cross-domain FER performance for binary (positive/negative) emotion classification? Specifically, can this modification reach performance comparable to within-domain training?

Chapter 2

Literature Review

This Section provides an extensive Literature Review related to the evolution of Facial Expression Recognition, the evolution of datasets used to train machine-learning and deep-learning models for FER, and the use of unsupervised domain adaptation techniques to mitigate the domain gap issue, particularly when labeled data is insufficient.

The core task of FER is to analyze facial expressions. Therefore, FER techniques are highly dependent on the detection of faces in images or videos, which relies on the recognition of facial features in these images or videos. The pictures or videos contain additional details, such as different backgrounds and surroundings, in addition to human faces.

2.1 Classic Machine-Learning Methods for FER

The earliest history of face recognition can be traced back decades, led by pioneers such as Woodrow Bledsoe and Helen Chan Wolf. They focused on developing rudimentary algorithms for identifying individuals. The algorithms are based on manual measurements of basic facial features (eyes, nose, mouth, etc.) from photographs. That technique was limited to controlled environments and also required significant human effort.

The specific research aimed at recognizing emotions from facial expressions was based on these foundational face-recognition studies. Initially, classic machine-learning methods were used based on handcrafted features.
Bassili investigated facial motion analysis. The primary purpose of his research was to determine how facial movements can be used to recognize basic human emotions. He categorized emotions into six fundamental emotions: anger, disgust, fear, joy, sadness, and surprise. This study is considered a foundational contribution to the field of emotion recognition because their work provided an early framework for understanding how these emotions are expressed and identified through dynamic facial cues. That research was very influential in understanding the relationship between facial expressions and emotional states. This foundational contribution to emotion recognition is reviewed in recent surveys [66] which trace the evolution of studies from Bassili's early work to modern FER systems."

Ushida and Yamaguchi developed one of the earliest systems that employed fuzzy logic for facial expression recognition. Their method involved the use of fuzzy-rule systems that utilized distances and angles derived from manually located facial feature points (such as mouth corners and eye corners) and textural features like patterns derived from image intensity (pixel values) in specific regions (like wrinkles near eyes/mouth, cheek puffiness) to produce linguistic classifications. These classifications were later analyzed using Mamdani IF-THEN rules, a fundamental technique in fuzzy inference, to identify seven expressions: neutral, happiness, sadness, surprise, fear, anger, and disgust. This groundbreaking system showed how fuzzy logic could be used in areas like human-computer interaction. Recent comprehensive surveys of FER systems affirm that Ushida and Yamaguchi's work laid foundational ground for applying fuzzy logic to human-computer interaction in expression analysis [67].

Matthew Turk and Alex Pentland developed the Eigenfaces method, which utilized Principal Component Analysis (PCA) to address face recognition. They treated faces as holistic patterns and, through dimensionality reduction using PCA, their method efficiently captured the most significant variations in facial appearance. A mathematical foundation for automated face representation and identification

emerged from this method. Although primarily designed for face recognition, the core PCA framework later became fundamentally crucial for facial expression analysis (FER) because its method facilitated efficient feature extraction from facial imagery. According to recent FER survey studies, Eigenfaces and other classical PCA-based models had a major impact on early automatic expression recognition research and laid the groundwork for the development of contemporary deep-learning FER systems. [68].

Lyons et al. were the first to implement Gabor wavelets to extract features from facial images for expression recognition. Their work demonstrated that multiscale, multi-orientation Gabor filters sensitive to frequency and orientation effectively capture significant facial changes critical for FER, such as wrinkles and furrows. Concurrently, Zhang et al. (co-authored by Lyons) also validated this approach by showing that Gabor features outperformed geometric methods in emotion classification. By applying these biologically inspired wavelets to facial regions, they improved the accuracy of classifying the six basic emotions over holistic approaches, such as PCA. This particular application of Gabor filters made them a quality standard for handcrafted feature extraction in early FER pipelines as they got 92% correctness on JAFFE. Despite computational limitations, the latest studies still confirm their continuing role in extracting facial features helpful in classifying human emotions [68].

Vapnik and co-workers proposed Support Vector Machines (SVMs). An SVM classifies unseen data by finding, during training, a hyperplane that keeps the largest possible margin between classes, thus maximising the distance of the closest points from either class to the hyperplane. Vapnik calls this the Optimal Separating Hyperplane (OSH), designed to minimise misclassification on both training and unseen data. Recent FER survey studies confirm that SVM remained a dominant classifier in early automated expression-analysis systems and is still widely employed as a benchmark and hybrid component in modern FER research due to its robustness on small datasets and high discriminative capability [68].

Wang and his team developed their emotion-reading program using Haar features, these are tiny "if-this-then-that" rules that compare brightness between pairs of rectangular patches on a face (e.g., "Is the left side lighter than the right side?").

Individually weak, these rules were combined into a robust classifier through AdaBoost. Over 1,400 rounds, AdaBoost iteratively selected the Haar feature whose look-up-table classifier best reduced class-weighted error, assigning each a vote weight (α_t) based on its accuracy. The algorithm then increased the weight of misclassified training samples for the next round. The final strong classifier was a sign-weighted sum of these selected weak rules. Crucially, this structure leveraged integral-image arithmetic, enabling high-speed computation—about 0.11 ms per face on 2004 hardware. Trained on an augmented JAFFE dataset (5,112 images synthesized from the original 213 photos of ten Japanese females), the system achieved high accuracy: 98.9% on JAFFE and 92.4% on an independent 206-image test set. This significantly outperformed a slower RBF-SVM baseline (91.6% accuracy, roughly 300 times slower). The speed and accuracy of this boosted Haar-LUT ensemble made it a popular handcrafted baseline for real-time facial-expression recognition (FER), capable of identifying anger, disgust, fear, happiness, neutral, sadness, and surprise, until deep CNNs became practical. Due to its efficiency and robustness, the boosted Haar-feature framework became a widely adopted handcrafted baseline for real-time FER tasks before deep CNN-based approaches became dominant [7, 68].

Along with the classical machine-learning techniques mentioned earlier in detail, other ML techniques such as Decision Trees, K-Nearest Neighbor (KNN), Bayesian Networks, and Hidden Markov Models (HMM) have also demonstrated satisfactory performance on controlled, smaller datasets. Still, the general trend has shifted toward deep-learning models because of their enhanced ability to extract optimal features from extensive data autonomously.

2.2 Deep Learning Methods for FER

Deep learning is a machine learning technique in which artificial neural networks are loosely inspired by the neurons of the human brain. Deep neural architectures enable more discriminative features for emotion classification by learning hierarchical representations directly from raw pixel inputs through back-propagation, in

contrast to conventional handcrafted feature-based techniques. Large-scale FER datasets and GPU-accelerated training made this paradigm shift possible, enabling deep models to outperform pipelines based on PCA, LBP, and Gabor (like those investigated by Padgett & Cottrell in early neural FER studies). Nevertheless, deep FER systems are still limited in their application in real-world scenarios due to issues like overfitting on sparse training data, sensitivity to cross-domain variation, and high computational cost [7, 18, 38, 68]. These restrictions encouraged the creation of modern domain-adaptation methods that enhance generalization across datasets without necessitating sizable labeled target samples.

The next significant step in data volume related to FER occurred in 2016, when Benitez-Quiroz, Srinivasan, and Martinez introduced EmotioNet, a massive collection of approximately one million face photos gathered from the web [34]. Instead of hand-labeling every image, they built an automatic AU detector that combines 66 landmark-based shape measurements with landmark-centred Gabor texture features and classifies them using kernel subclass discriminant analysis (KSDA).

These feature vectors $\mathbf{x} \in \mathbb{R}^d$ are mapped into a high-dimensional space through a kernel $\boldsymbol{\phi}(\mathbf{x})$. And then the detector applies Kernel Sub-class Discriminant Analysis (KSDA), which finds projection weights $\mathbf{w}$ by solving:-

$$(\mathbf{S}_w + \lambda \mathbf{I})^{-1} \mathbf{S}_b \, \mathbf{w} \;=\; \gamma \, \mathbf{w}, \tag{2.1}$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within- and between-subclass scatter matrices in feature space and λ is a small regulariser. Each Action Unit k is then detected with a sigmoid scorer:-

$$P(\mathrm{AU}_k \mid \mathbf{x}) \;=\; \sigma\big(\mathbf{w}_k^{\top} \boldsymbol{\phi}(\mathbf{x}) + b_k\big). \tag{2.2}$$

Facial Action Units (AUs) are the basic muscle movements defined by the Facial Action Coding System for example, AU 12 (lip-corner puller, a smile) or AU 1 (inner-brow raiser) and any simple or complex facial expression can be described as a combination of these building blocks [69]. Although the automatically generated AU labels are noisy, the KSDA pipeline recognised AUs accurately in real time (over 30 images per second on 2016 hardware) and the system semantically

assigned emotion categories based on detected AU patterns. The authors evaluated annotation accuracy using expert verification of a subset of images, enabling future work to benchmark progress precisely.

In 2017, Hasani and Mahoor in their research used cascading deep residual Inception modules with a linear-chain Conditional Random Field (CRF), which markedly improved video-based FER accuracy. In the first phase, they extracted per-frame spatial features by using three Inception-ResNet blocks (parallel convolutional paths with identity skip-connections). Then they fed those features into a CRF to model the temporal dependencies between consecutive frames, falling back to a simple softmax layer when evaluating still images. On the extended CK+ set, their CRF-augmented model reached 93.04% accuracy (vs. 85.77% without CRF) in subject-independent evaluation. On FERA, it achieved 66.66% accuracy in subject-independent testing. Their method outperformed earlier CNN baselines [42].

Facial Identity attributes act as confounding noise in facial expression recognition (FER) by obscuring transient expression signals within static facial morphology [40, 70, 71]. To separate identity-related features from expression-related features, Yang, Ciftci, and Yin introduced De-Expression Residue Learning (DeRL) in 2018 to disentangle expression from identity for FER [72]. A conditional GAN first generates a subject's neutral face from an expressive input. The "residue (pure expression information left behind in the process)" defined as the expressive component preserved in the generator's intermediate layers, is then extracted and fed to a multi-branch CNN for classification. Evaluated on CK+ [33] with a 10-fold subject-independent protocol, DeRL achieved 97.7% accuracy by isolating expression signals from identity variations. It also set new state-of-the-art results on Oulu-CASIA [73] (88.0%) and competitive benchmarks on MMI [74] (73.23%), BU-3DFE [75] (84.17%), and BP4D+ [76] (81.39%), demonstrating cross-dataset robustness [72]. The DeRL model demonstrated the ability to separate facial identity from emotional cues effectively. However, the emotion recognition (FER) research community is currently focused on developing lightweight and computationally efficient convolutional neural network (CNN) architectures, as training

models with a higher number of parameters incurs a significant computational cost.

Amal et al. [77] introduced a Convolutional Neural Network (CNN) architecture-based real-time facial emotion recognition system that was trained using the FER2013 dataset. LBP-based face detection is combined with additional hand-crafted features, such as Histogram of Oriented Gradients (HOG) descriptors and facial landmarks extracted using dlib, to improve the model's processing of raw pixel inputs. Their tests showed that adding engineered features to CNN input significantly enhanced performance, reaching 75.1% accuracy as opposed to 59.1% when using only raw pixel data. The suggested approach performed better than the baseline model, but its reliance on manually created features and sensitivity to changes in background, lighting, and facial pose suggest that it is not very robust in unrestricted real-world settings. For cross-domain FER tasks in particular, these limitations emphasize the need for more generalized and domain-invariant feature learning methods like UDA.

Using the FER2013 and RAF-DB datasets, Qutub and Atay [78] evaluated deep CNN architectures for facial emotion recognition by contrasting a baseline CNN with EfficientNet-B0, ResNet18, VGGNet16, and VGGNet19. The superiority of deeper networks with transfer learning have been demonstrated by their experiments, which showed that VGGNet19 performed best on FER2013 with 71.02% accuracy and ResNet18 obtained 86.02% accuracy on RAF-DB. However, They achieved high accuracy but no cross-domain testing is done, and performance under real-world variations like illumination, occlusion, or pose changes is not assessed. Moreover, the strongest models are computationally expensive and unsuitable for real-time edge-device deployment due to their large parameter counts (e.g., 45M in VGGNet19).

In [79] El Boudouri and Bohi (2025) introduced EmoNeXt, a novel framework intended to attain SOTA accuracy on the FER2013 dataset[17]. The ConvNeXt [80] architecture, a modernized CNN that incorporates design concepts from Vision Transformers, is used at its core. In order to increase its representational power for facial expressions, EmoNeXt adds three essential components to this foundation.

Spatial Transformer Network (STN): An STN [81], a differentiable module that learns to apply spatial transformations (such as scaling and rotation) to the input image, is where the model starts. As seen in Figure 2.1, this makes the network more invariant to changes in pose and alignment by enabling it to automatically focus on and spatially normalize the most feature-rich areas of the face.

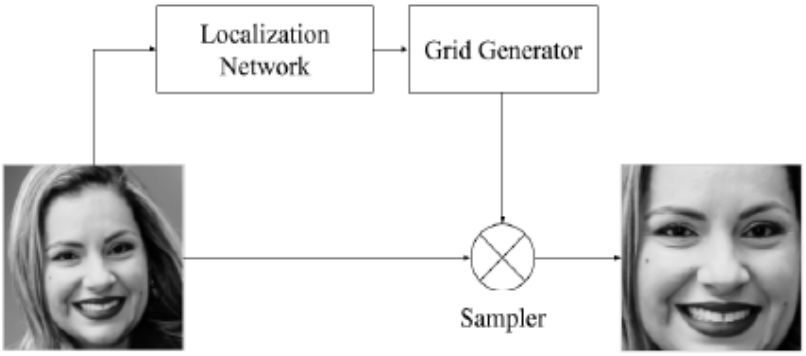

FIGURE 2.1: Architecture of Spatial Transformer Module [79]

Squeeze-and-Excitation (SE) Blocks: After each of the four main ConvNeXt stages, the authors integrate SE blocks [65]. These blocks perform channel-wise attention by first squeezing global spatial information into a channel descriptor via global average pooling, then exciting (recalibrating) the channels through a gating mechanism based on their importance. As seen in Figure 2.2, this enables EmoNeXt to highlight informative features pertinent to emotion classification.

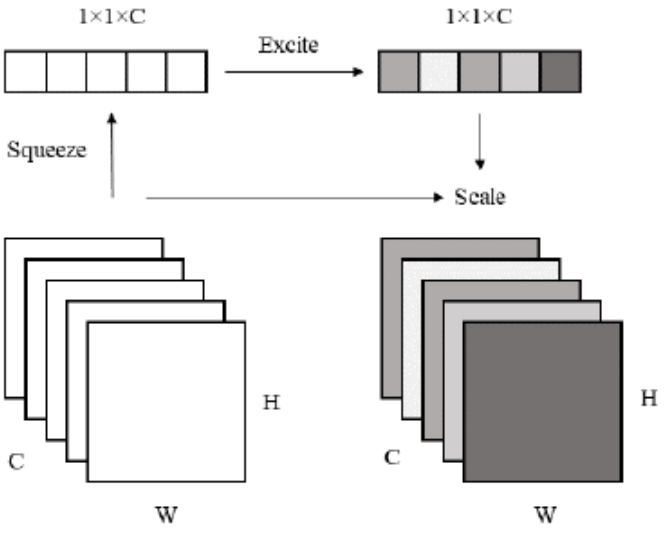

FIGURE 2.2: Squeeze and Excitation Block [79]

The EmoNeXt architecture incorporates Squeeze-and-Excitation (SE) blocks after each of its four main ConvNeXt stages to implement a form of channel-wise attention. The fundamental purpose of these blocks is to adaptively recalibrate the feature maps, allowing the network to emphasize informative channels and suppress less useful ones. The SE block operates through a sequential process of squeezing and excitation [79]. The squeeze operation first compresses each two-dimensional feature map into a single scalar value by using global average pooling, which captures the global distribution of responses for each channel and embeds this holistic information into a compact descriptor. This squeezed vector is then passed to the excitation operation, which consists of a small gating network [81]. This network, typically formed by two fully connected layers, learns to model the non-linear, channel-wise dependencies and produces a set of scaling weights. These weights, normalized between zero and one by a sigmoid activation, represent the importance of each channel in the context of the current input. Finally, the original feature maps are rescaled by multiplying them with these learned weights, effectively performing a feature recalibration that enhances discriminative power for the emotion recognition task [79].

Self-Attention Regularization (SA): A novel regularization term is added to the standard cross-entropy loss. This term is calculated from the self-attention weights of the final feature vector. By minimizing the variance of these attention weights, the SA regularization encourages the model to produce a more compact and balanced feature representation, preventing it from over-relying on a small subset of features and potentially improving generalization [79] as depicted in Figure 2.3.

EmoNeXt introduces a novel Self-Attention Regularization (SA) [79] term to its loss function to encourage the learning of a more compact and balanced feature representation. This mechanism addresses the issue where a deep network might over-rely on a narrow subset of features, which can hinder generalization. The process begins by applying a self-attention mechanism to the final feature vector extracted by the backbone network. This involves creating Query and Key projections from the feature vector and then computing self-attention weights through scaled dot-product and softmax normalization. These weights indicate the relative

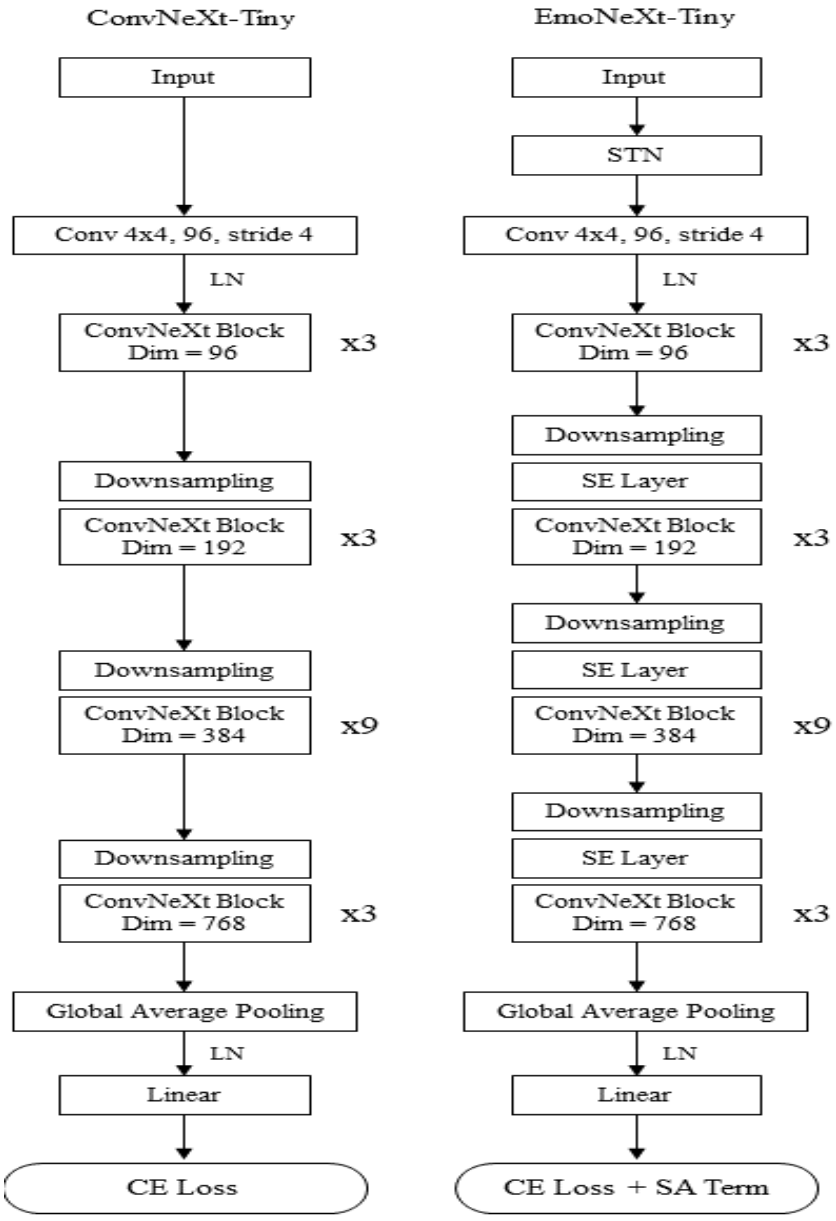

FIGURE 2.3: Architecture designes for ConvNeXt and EmoNeXt [79]

importance each feature element assigns to every other element. The core of the SA regularization is to minimize the variance of these self-attention weights. By penalizing the deviation of individual weights from their mean, the loss term encourages the network to distribute its attention more uniformly across the entire feature vector [39]. EmoNeXt [79], employed a rigorous training regimen, including advanced techniques like AdamW optimization with cosine decay, extensive data augmentation (RandomCropping, RandomRotation), and regularization methods (Stochastic Depth, Label Smoothing). Furthermore, they leveraged pre-trained weights from ImageNet-22k and used Mixed Precision Training for efficiency. This

comprehensive approach allowed their largest model, EmoNeXt-XLarge, to achieve a then-state-of-the-art accuracy of (76.12%) on the FER2013 dataset, outperforming other contemporary models like Segmentation VGG-19 (75.97%) and the original ConvNeXt variants.

The shortfall of EmoNeXt [79], like many high-accuracy in-domain models, is its lack of an explicit mechanism to handle domain shift, as it is designed and optimized solely for a single dataset FER2013 [17]; consequently, despite its impressive in-domain accuracy of 76.12%, it experiences severe performance degradation in cross-domain scenarios. This limitation is compounded by its high model complexity, where the EmoNeXt-XLarge architecture—with a feature channel configuration of (256, 512, 1024, 2048), millions of parameters, an STN, multiple SE blocks, and a large 224x224 input resolution—demands substantial computational resources, rendering it unsuitable for resource-constrained or real-time applications. Therefore, while its reported accuracy is notable within the controlled in-domain benchmark of FER2013 [17], this focus underscores a lack of proven robustness for real-world, unconstrained environments, which stands in direct contrast to the core contribution of this thesis: to preserve accuracy across domains with an efficient model like ResEmoteNet-DANN, rather than to maximize in-domain performance alone.

ResEmoteNet [39], proposed by Arnab Kumar Roy et al., is also one of the state-of-the-art models for FER. ResEmoteNet combines three well-established CNN components into a streamlined process. First, a simple sequence of three layers of convolution, BatchNorm, and MaxPool captures basic edges and textures. This block can be referred to as a feature extractor. The output from this feature extractor then passes through a Squeeze-and-Excitation (SE) block. It carries out two main functions: Squeeze, which applies global average pooling to summarize spatial information from each channel into a single global descriptor, and Excitation, which utilizes a sigmoid-activated gating mechanism to understand channel dependencies. The method employed by the SE block enables the network to learn various attention weights, emphasizing the significance of each input component for the network's output. Subsequently, three residual blocks are

employed to gather more profound semantic insights; the identity skip-connections help stabilize gradients and allow the network to learn complex mappings without performance degradation. The resultant feature map is compressed via adaptive-average pooling into a fixed-length vector, which is subsequently processed by a single linear layer to produce a seven-way softmax distribution that represents fundamental emotions. ResEmoteNet outperforms models such as Segmentation-VGG-19, LHC-Net, and the CNN-based EmoNeXt on all three datasets, achieving remarkable performance scores of 79.79% on FER-2013, 94.76% on RAF-DB, and 72.93% on AffectNet [39]. The authors credit these advancements to the SE module's ability to recalibrate channel-wise features and maintain residual stability. The following hyperparameters optimized the training structure, enumitem

(i) SGD optimization (initial LR=1e-3)

(ii) Plateau-based LR scheduler (reduce by 0.1 upon stagnation)

(iii) Minimal augmentation (only random horizontal flipping)

ResEmoteNet demonstrates that a hybrid of CNN, SENets, and ResNet learning, without relying on heavy transformer blocks, can achieve or surpass SOTA accuracy, even with limited GPU resources.

Nonetheless, while deep CNNs have demonstrated effectiveness in facial expression recognition (FER), the majority of these models discussed previously, such as EmoNeXt and ResEmoteNet, perform well when training and testing datasets come from the same distribution. However, when the data distribution changes due to domain shift, the performance of these models, which are not specifically trained using UDA architects, drastically decreases, limiting their ability to generalize across different domains. Therefore, let us examine recent studies focused on domain adaptation for FER.

DANN is pioneering deep learning techniques to mitigate domain gaps [54].
Recent surveys still refer to it as "the pioneering work DANN" when discussing about domain adaptation [31]. It learns domain-invariant features through adversarial training and it comprises on three basic components:-

Component	**Role in the network**	**Mathematical form**
Feature extractor G_f	Maps an input image x to a latent feature vector f	$f = G_f(x;\, \theta_f)$
Label predictor G_y	Predicts the class label $\hat{y}$ (source domain only)	$\hat{y} = G_y(f;\, \theta_y)$
Domain classifier G_d	Predicts the domain label $\hat{d}$ via a *gradient-reversal layer (GRL)* inserted before it	$\hat{d} = G_d\big(\mathrm{GRL}(f); \theta_d\big)$

TABLE 2.1: Core Components of DANN [54].

Losses:

Let $\mathcal{D}_s$ be the labelled source set, $\mathcal{D}_t$ the unlabelled target set, and $d_i \in \{0, 1\}$ the domain label (0 = source, 1 = target). With cross-entropy ℓ_{CE}, the classification and domain losses are :-

$$\mathcal{L}_y = \frac{1}{|\mathcal{D}_s|} \sum_{(x_i, y_i) \in \mathcal{D}_s} \ell_{\mathrm{CE}}\Big(G_y\big(G_f(x_i)\big),\, y_i\Big), \tag{2.3}$$

$$\mathcal{L}_d = \frac{1}{|\mathcal{D}_s| + |\mathcal{D}_t|} \sum_{x_i \in \mathcal{D}_s \cup \mathcal{D}_t} \ell_{\mathrm{CE}}\Big(G_d\big(G_f(x_i)\big),\, d_i\Big). \tag{2.4}$$

Saddle Point Objective:

DANN solves the min–max problem

$$\min_{\theta_f, \theta_y} \max_{\theta_d} \; \mathcal{L}_y \; - \; \lambda\, \mathcal{L}_d, \tag{2.5}$$

where λ balances task accuracy against domain invariance.

During back-propagation, the GRL multiplies the gradient passed to G_f by $-\lambda$, so that

$$\nabla_{\theta_f} \; \longleftarrow \; \frac{\partial \mathcal{L}_y}{\partial \theta_f} \; - \; \lambda\, \frac{\partial \mathcal{L}_d}{\partial \theta_f},$$

forcing G_f to confuse the domain classifier while still supporting the label predictor [54]. Wang et al. addressed the issue of domain shift in Facial Expression Recognition (FER) through their proposed architecture, AdaFER. AdaFER uses fixed pre-trained AU detector to derive AU codes, followed by two adaptation methods. AU-Guided Annotating (AGA) generates pseudo-labels for target-domain images by aligning AU codes with those from source images, either using

the most commonly occurring expression (hard label) or the distribution of expressions (soft label, handled through KL divergence) derived from matched source neighbours [58].

Formally, the classification-style loss employed in AGA is

$$\mathcal{L}_c \;=\; \mathrm{CE}\big(P_s, Y_s\big) \;+\; \beta\Big(\mathrm{CE}\big(P_t, Y_{S\text{-hard}}\big) + \mathrm{KL}\big(P_t, Y_{T\text{-soft}}\big)\Big), \tag{2.6}$$

where P_s and P_t are the predicted class distributions for source and target images, Y_s are ground-truth source labels, $Y_{S\text{-hard}}$ the hard pseudo-labels, $Y_{T\text{-soft}}$ the soft pseudo-labels, and β balances the two target terms.

AU-Guided Triplet Training (AGT) creates cross-domain triplet sets, where anchor and positive pairs share the same AU codes while the negative differs, employing a margin-based triplet loss to group AU-equivalent faces across domains:

$$\mathcal{L}_{\text{tri}} \;=\; \max\Big\{0,\; \gamma + \|F_a - F_p\|_2^2 - \|F_a - F_n\|_2^2\Big\}, \tag{2.7}$$

with F_a, F_p, F_n the ℓ_2-normalised features of anchor, positive and negative images and γ the margin.

These elements are optimised collectively in an end-to-end manner using a ResNet-18 backbone (pre-trained on MS-Celeb-1M). The overall objective is therefore

$$\mathcal{L}_{\text{all}} \;=\; \mathcal{L}_c \;+\; \varepsilon\,\mathcal{L}_{\text{tri}}, \tag{2.8}$$

where the hyper-parameter ε trades off classification and metric-learning terms.

AdaFER achieves state-of-the-art unsupervised cross-domain accuracy without additional inference overhead, recording scores of 81.40% (CK+), 61.37% (JAFFE), 57.29% (FER2013) and 70.86% (ExpW), surpassing previous methods by two to six percentage points by addressing annotation bias through AU consistency [58].

In 2023, Guo et al. (2023) [56] introduced the USTST framework to address

UD-FER by integrating self-training with similarity transfer. In USTST, the Cross-Swin-Transformer (CST) replaces the standard self-attention with a cross-attention mechanism that emphasizes anatomically similar regions(e.g., eyes/-mouth), then enters an alternating self-training–resampling loop. During Self-Training Resampling (STR), it fuses facial kernel patches from high-confidence target images with source domain samples to reduce data discrepancy. A clustering-based reliability check and entropy thresholds were used to refine pseudo-labels for compound expressions inside the Knowledge Transfer (KT) module. To enhance Domain alignment, adversarial training and Maximum Mean Discrepancy (MMD) losses were used. They evaluated the USTST framework on RAF-DB (source) and chose five target datasets (CK+, JAFFE, SFEW, FER2013, ExpW). USTST achieves a mean accuracy of 70.22%, outperforming state-of-the-art methods on CK+ (91.02%), JAFFE (67.85%), and FER2013 (69.08%) while showing competitive results on SFEW and ExpW. This demonstrates that using similarity-guided self-training and robust pseudo-label curation effectively together mitigates domain shift in FER.

Gao et al. (2024) proposed a framework named AGLRLS to tackle cross-domain facial-expression recognition (FER), targeting two key shortcomings of prior work: (i) excessive reliance on global feature alignment, which ignores transferable local cues, and (ii) inadequate discriminative supervision for target-domain features [82]. The framework processes each face through seven parallel pathways—one global, five local (eyes, nose, mouth corners), and one fused stream. Every pathway undergoes its own adversarial alignment via dedicated domain discriminators, so both global and local features become domain-invariant [82].

To boost discriminability, a semantic-aware pseudo-labeling engine assigns labels to each stream: classifiers predict target labels independently, while dynamic class-adaptive thresholds (based on per-class confidence trends) discard unreliable pseudo-labels; the remaining labels guide classifier learning, sharpening decision boundaries.

A Global Local Prediction Consistency (GLPC) module then fuses predictions at inference: it trusts the high-confidence global-local classifier first and falls back to the global classifier when uncertainty remains, finally masking and aggregating

predictions from all seven paths [82].

With this design, AGLRLS attains state-of-the-art results on the RAF-DB→CK+, JAFFE, SFEW 2.0, FER2013, ExpW benchmark, raising mean accuracy by +2.94 pp using ResNet-50 and +5.21 pp with the lightweight MobileNet-v2 backbone, surpassing the previous AGRA method [83] and demonstrating the benefit of local-aware alignment even on compact models. However, a main limitation of the AGLRLS framework is its computational complexity. The strength of the model comes from the use of strong backbone networks, such as ResNet50 or MobileNet-v2, as feature extractors, along with seven parallel classifiers and domain discriminators. This structure includes millions of parameters, which require a lot of GPU memory and computational power for training. The two-stage training process also adds to the overall training time. This high demand for resources could limit its use in settings with fewer resources or in applications that need quick deployment.

To enhance facial expression recognition (FER) for cross-domain datasets, Yang et al. (2024) introduced LA-CMFER [60]. This method addresses the differences between datasets (inter-domain shifts) as well as the inconsistencies found within datasets (intra-domain shifts). Their process involves two main components: first, a dual-level alignment that focuses on difficult-to-classify face images and organizes similar expressions based on labels, and then a twin network branches that examine both the entire face (global) and specific regions such as the eyes and mouth (local). To reduce error, these branches verify predictions against each other to eliminate inaccurate AI-generated labels.

Furthermore, they evaluated their technique on six widely-used FER benchmarks, including RAF-DB, AffectNet, and ExpW, and LA-CMFER surpassed existing techniques in adapting to new contexts. They also demonstrated its efficacy for real-life practical applications such as emotion-aware AI systems.

Zhu et al. (2025) introduce FER-DAS [59]to address potential issues in CD-FER, including methods incorporating adversarial mechanisms [55, 83]. They found, achieving stability and convergence in adversarial learning presents significant challenges. FER-DAS is an innovative framework for recognizing facial expressions across different domains that enhances alignment by deliberately merging

active and stable samples. Their Active Assessment Strategy (AAS) identifies "Active" target images (exhibiting high uncertainty and variance) and selectively incorporates them into training to focus on challenging examples. Meanwhile, the Cross-Domain Dynamic Class Threshold (CD-DCT) module generates data-driven thresholds for every class, making detected high-confidence "stable" samples in both domains more robust while alleviating class imbalance and noise. Finally, Weighted Cross-Domain Alignment (WCDA) leverages these adaptive thresholds to identify sample importance relative to class centers, enabling fine-grained, class-specific alignment. They evaluated their model on FER datasets affected by lighting, pose, and demographic variations. They claimed that FER-DAS surpasses existing adversarial and static-threshold approaches by up to 4 percentage points. Their Study highlights the effectiveness of integrating active learning with dynamic thresholding and weighted alignment. As previously discussed, some FER applications simplify the task by grouping emotions into positive versus negative. Cross-domain FER (CD-FER), leveraging unsupervised domain adaptation (UDA), has also been studied to mitigate domain shift.

EM-UDA [57] proposed by Priti R. Jain et al., is one such technique. In their study, they introduce a model based on adversarial domain adaptation, called EM-UDA, which employs deep convolutional neural networks (DCNNs) to categorize facial emotions into positive and negative classes. Their methodology uses AffectNet as a labeled source dataset and trains it along with unlabeled data from either the CK+ or FER 2013 datasets.

They used the same original DANN architecture that contains a feature extractor, a domain classifier equipped with a gradient reversal layer (GRL), and a label predictor. Along with this, they used VGG-19 as a backbone architecture.

The results of their experiments showed EM-UDA records an accuracy of 83.9% and an F1-score of 82.8% when evaluated on CK+, surpassing the baseline transfer learning model by 5.6% in terms of accuracy. When tested on the FER 2013, EM-UDA achieves an accuracy of 74.55% and an F1-score of 74.87%, exceeding the baseline by roughly 5%. These results demonstrate that unsupervised domain adaptation effectively addresses the challenges posed by limited labeled data and domain shift between datasets.

TABLE 2.2: Classical and CNN Based FER (No Explicit UDA)

Technique	**Domain Gap Handling**	**Accuracy Limitation**	**Model Complexity**
EmotioNet (2016) [34]	Focus on AU detection; no UDA; AU stats are domain-dependent.	Emotion accuracy tied to AU quality; noisy AU labels.	KSDA + AU pipeline; real-time, low–moderate cost.
Hasani & Mahoor (2017) [42]	Temporal CRF model; no explicit cross-domain handling.	CK+ 93.04%, FERA 66.66%; drops cross-dataset.	Inception-ResNet + CRF; moderate–high compute.
DeRL (2018) [72]	Disentangles identity vs. expression; not UDA.	Strong on controlled sets; weaker in-the-wild cross-domain.	cGAN + multi-branch CNN; training heavy.
Amal et al. (2022) [77]	Real-time FER using CNN trained on FER2013, in-domain only; no UDA; performance degrades in unconstrained settings.	75.1% with features / 59.1% raw pixels; not evaluated cross-dataset.	Lightweight CNN + LBP + HOG + dlib landmarks; limited generalization.
Qutub & Atay (2023) [78]	In-domain FER; standard CNN without UDA; accuracy affected by pose, lighting, and occlusion variations.	71.02% FER2013; performance drops substantially in real-world scenarios.	Lightweight CNN; suitable for controlled settings; weak cross-domain generalization.
EmoNeXt (2025) [84]	STN helps spatial invariance; no UDA.	76.12% FER2013; degrades with demographic/low-res shift.	ConvNeXt + STN + SE; moderately heavy.
ResEmoteNet (2025) [39]	Robust in-domain; no UDA.	79.79% FER2013, 94.76% RAF-DB, 72.93% AffectNet; cross-domain drop.	Lightweight CNN + SE + residuals; efficient.

TABLE 2.3: UDA-based FER (Cross-Domain Focus)

Technique	**Domain Gap Handling**	**Accuracy Limitation**	**Model Complexity**
DANN (2016) [54]	Adversarial feature alignment (GRL) reduces domain gap.	Improves cross-domain transfer but not SOTA; struggles on subtle AUs.	Simple GRL add-on to CNN; low–moderate complexity.
AdaFER (2022) [58]	AU-guided pseudo-labels + triplet learning; assumes AU reliability across domains.	CK+ 81.40%, JAFFE 61.37%, FER2013 57.29%; pseudo-label noise limits performance.	ResNet-18 + fixed AU detector; moderate; additional preprocessing required.
USTST (2023) [56]	Self-training + similarity transfer; CST focuses on local discriminative regions.	FER2013 69.08%; sensitive to thresholds and clustering; brittle with noisy targets.	Transformer-based; alternating STR/KT loops; high memory and training time.
AGLRLS (2024) [82]	Global + local alignment through seven parallel discriminators.	+2.9–5.2 pp improvement over prior work; diminishing gains on rare classes.	Heavy model (multi-head discriminators); requires large GPU resources and two-stage training.
LA-CMFER (2024) [60]	Dual-level alignment; **multi-source** training increases robustness.	Strong across six benchmarks; requires multiple labeled source datasets.	Twin-branch network + consistency mechanisms; multi-source setup adds complexity.
FER-DAS (2025) [59]	Dynamic active sample selection + adaptive class thresholds for alignment.	+4 pp over adversarial baselines; needs retuning for each domain pair.	Medium–high complexity; additional thresholding and weighting modules.
EM-UDA (2024) [57]	DANN-style UDA using VGG backbone.	SOTA for binary emotion classification; lacks full basic emotion coverage.	Parameter-heavy (VGG-19); slow training and high memory usage.

Chapter 3

Methodology

In this chapter, we will discuss our proposed framework for cross-domain facial emotion recognition (FER). Our approach is a blend of adversarial Unsupervised Domain Adaptation (UDA), residual convolutional blocks, and squeeze-and-excitation (SE) modules to create a robust and compact model. Our Model is based on DANN (Domain-Adversarial Neural Networks) [54] and ResEmoteNet architectures [39]. This framework focuses on learning features that are irrelevant to a specific domain. For example, whether it is taken in a high-resolution studio setting like CK+ or from an in-the-wild dataset like FER2013 with motion blur or occlusion, a smile conveying the emotion "happiness" should be identifiable.

The overall strategy is to extract common or disentangled, latent space that captures relevant features independent of the domain. This latent representation obtained through the Adversarial training is then fed into a modified version of the ResEmoteNet architecture [39], where the Squeeze and Excitation block [65] uses these learned features.

3.1 Framework Components

The framework consists of the following main components:

i. Datasets

ii. Data Preprocessing

iii. Proposed Architecture: ResEmoteNet-DANN

iv. DANN Encoder (Domain Irrelevant Feature Extractor)

v. Gradient Reversal Layer (GRL) and Domain Discriminator

vi. Squeeze-and-Excitation Blocks

vii. Residual Blocks

viii. Classification Head

ix. Loss Functions and Training Procedure

x. Evaluation Metrics

3.2 Datasets

During the phase of model development and implementation, we used four benchmark datasets: AffectNet [35], ExpW [37], FER2013 [17], and JAFFE [85].

3.2.1 AffectNet

AffectNet is one of the largest collections of human face images used for studying emotions. It contains millions of pictures of people's faces that were collected from the internet, each showing different facial expressions like happiness, sadness, anger, surprise, fear, disgust, or a neutral look as shown in Figure 3.1. These

FIGURE 3.1: AffectNet Dataset Sample

images are labeled with the emotion they represent, so computers can be trained

to recognize and understand feelings from faces. In simple words, AffectNet is like a huge photo album of faces with emotions written under them, created to help machines learn how to "read" human emotions.

3.2.2 ExpW

The ExpW (Expression in the Wild) dataset [37] is also a very valuable benchmark dataset for FER Tasks. This dataset consists of over 90,000 images that are gathered from diverse settings and annotated with seven emotion labels: "angry" (0), "disgust" (1), "fear" (2), "happy" (3), "sad" (4), "surprise" (5), and "neutral" (6). Figure 3.2 shows the ExpW Dataset Samples.

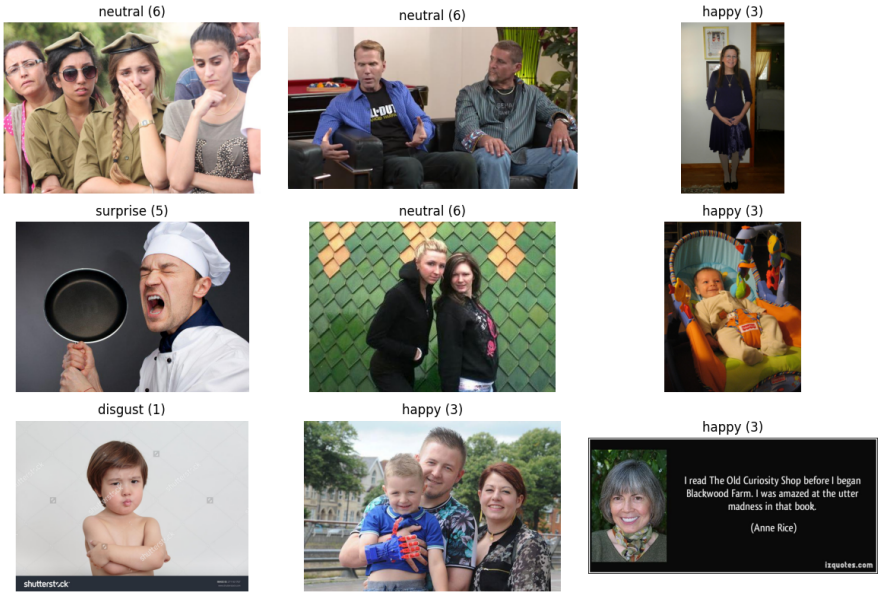

FIGURE 3.2: ExpW Dataset Samples

3.2.3 FER2013

FER2013 is a collection of black-and-white images representing facial expressions that help computers learn how to recognize emotions. It has about 35,000 small images, each showing one of seven basic emotions: happy, sad, angry, surprised, fearful, disgusted, or neutral. The pictures are simple and useful for training purposes. In simple words, FER2013 is like a teaching book full of tiny face

photos with their emotions written on them, so computers can practice and get better at 'reading' human feelings. The sample of this data set is shown in Figure 3.3.

FIGURE 3.3: FER2013 Dataset Sample

3.2.4 JAFFE

JAFFE (Japanese Female Facial Expression) is a small collection of face images. It contains images of Japanese women, each showing different facial expressions like happiness, sadness, anger, surprise, fear, disgust, and neutral. The dataset is often used in studies to train and test how well computers can recognize emotions from faces. In simple words, JAFFE is like a small photo album of women's faces, each showing a clear emotion, created to help machines learn how to understand human emotions. Sample of JAFFE dataset is depicted in Figure 3.4.

FIGURE 3.4: JAFFE Dataset Sample

3.3 Data Preprocessing

There are specific pre-processing steps involved, particularly for each benchmark dataset that we used in this research. Figure 3.1 shows the data samples.

3.3.1 Preprocessing of AffectNet

The AffectNet dataset [35], a well-known resource for recognizing facial expressions, contains over a million images annotated with emotion labels. Organized in

training and validation sets, where images are located in train_set/images and val_set/images, with corresponding annotations structured in train_set /annotations and val_set/annotations, respectively.

3.3.1.1 Filtering and Label Mapping

The original AffectNet dataset contains ten emotion categories: 0—"neutral" 1—"happy," 2—"sad," 3—"surprise," 4—"fear," 5—"disgust," 6—"anger," 7—"contempt," 8—"unknown," and 9—"no_face." However, to fulfill our requirements, we needed a specific subset of these emotions, which required a preprocessing step to filter, remap, and reorganize the dataset according to the research aims and model needs.

To focus on the primary emotional expressions relevant to this study, images labeled as "contempt," "unknown," and "no_face" were left out. As a result, only images labeled as "happy," "surprise," "sad," "anger," "disgust," "fear," and "neutral" were kept. This filtering of the labels was essential to ensure that the research objectives were fulfilled and to align with the structure of the other dataset used in this study, hence improving its utility and performance in the specific context [35]. Initially, the selected emotions were mapped to numerical indices based on their alphabetical order:

i. "anger" → 0

ii. "disgust" → 1

iii. "fear" → 2

iv. "happy" → 3

v. "sad" → 4

vi. "surprise" → 5

Later, to ensure compatibility with a pretrained model designed for the FER-2013 dataset [17], the labels were rearranged to align with the conventions of FER-2013. The final label mapping was as follows:-

i. "happy" → 0

ii. "surprise" → 1

iii. "sad" → 2

iv. "anger" → 3

v. "disgust" → 4

vi. "fear" → 5

vii. "neutral" → 6

This reordering helped in training the Model for CD-FER using UDA techniques resulting in the enhancement of both training efficiency and effectiveness.

3.3.1.2 Validation Set

The Original AffectNet-provided validation set was kept as it is (in the context of the number of images and the images used in every class). That maintained the integrity of benchmark validation.

3.3.1.3 Verification

We also conducted a verification process to confirm the integrity of the dataset by making sure that unwanted labels are removed and verifying consistency between images and their corresponding labels in the CSV.

3.3.2 Preprocessing of ExpW

The dataset is structured in a label file (label.lst) that contains annotations and a folder of images (all_images). We utilized a preprocessing pipeline to prepare this dataset for the training of our CD-FER model, which involved data loading and

filtering of relevant expressions, splitting the dataset, organizing directories, and generating CSV files, along with a verification phase through visualization.

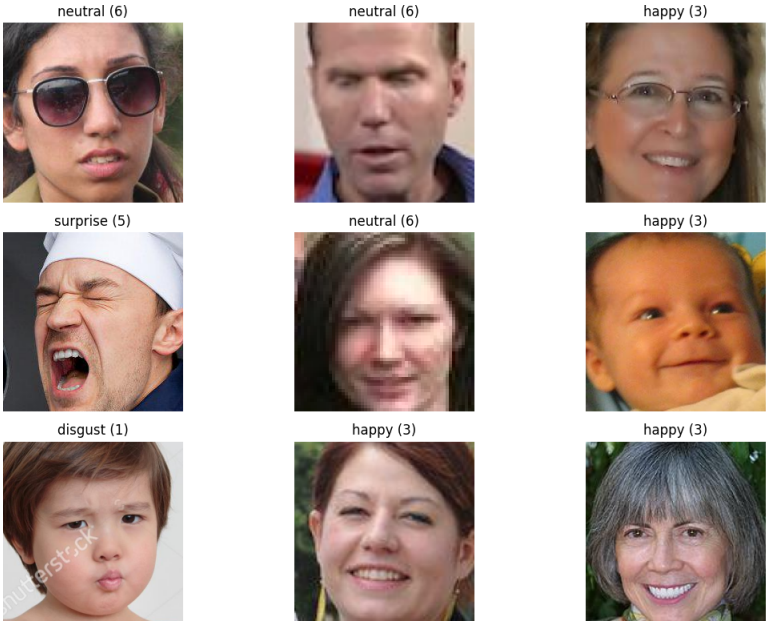

FIGURE 3.5: ExpW Dataset Samples After pre-processing

3.3.2.1 Data Loading of ExpW

The preprocessing process started with loading the label file (label.lst), which is a space-separated text file without headers, into a pandas DataFrame. This file contains columns that represent the image filename, face ID, bounding box coordinates (top, left, right, bottom), confidence score, and emotion label. To train our model, we need the primary facial expression for each image (even if it contains multiple faces in an image) and minimize the noise created by secondary faces.

Only the entries where face_id == 0 were kept, representing the most significant face detected in each image. After this filtering, we used the pandas library to filter the features, only the image filename and emotion label, which are the essential elements needed for the classification task.

This cleaned subset ensured that each training instance corresponded to a single, unambiguous facial expression, reducing the chances of conflicting labels during

learning. Moreover, removing auxiliary faces improved the consistency of emotion annotations across the dataset. This preprocessing step ultimately produced a reliable and noise-reduced dataset for downstream model training.

3.3.2.2 Dataset Splitting

To facilitate model training, validation, and testing, the filtered dataset was segmented into three parts: training, validation, and test sets. A stratified sampling method was employed to maintain the distribution of emotion labels across all splits, addressing class imbalance issues present in the dataset. The proportions set for the splits were 70% for training, 15% for validation, and 15% for testing, yielding 37,426 images for training, 8,020 for validation, and 8,020 for testing. This stratification was performed using the train_test_split function from the scikit-learn library, with a designated random state (random_state=42) to ensure reproducibility. The sizes of each subset were verified through printed outputs, affirming the reliability of the splitting process.

3.3.2.3 Directory Organization and CSV File Generation

For each dataset segment (training, validation, and test), a specific directory was established under the ExpWdata/ root directory (e.g., ExpWdata/train, ExpWdata/ val, ExpWdata/test). We utilized the shutil module to transfer images belonging to each segment from the original all_images directory to their respective segment folders. This approach allows us to prevent duplicate copying if an image is already present in the target location. This process ensured a neat and self-contained dataset structure. Furthermore, CSV label files (train_labels.csv, val_labels.csv, test_labels.csv) were created for each segment, containing the image filenames along with their related emotion labels.

This structured organization not only streamlines the data pipeline but also ensures seamless integration with PyTorch dataloaders during model training. By separating images and labels for each split, potential indexing or path related

errors are minimized. Such a modular layout also supports future extensions, including data augmentation and cross-dataset evaluations.

3.3.3 Preprocessing of FER 2013

The FER 2013 (Facial Expression Recognition 2013) dataset [17] is a well established benchmark for FER. This dataset contains 35,887 grayscale facial images, individually every image is classified into one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Initially, the Original FER2013 dataset is in CSV format. The dataset consists of entries containing an emotion label, a space-separated sequence of pixel values that represent a 48×48 image, and also has a reference that specifies the dataset division (training, validation, or test). To align this dataset with our model's needs, a preprocessing pipeline was created to convert the CSV data into an organized directory of image files and labels into corresponding CSV files for each dataset split.

Revised processing of each dataset split was carried out systematically as outlined below:

i. Directory Setup: A directory specific to the split (e.g., data/train) was generated using ensure_dir, which utilized mkdir with parents=True and exist_ok=True to address any existing directories appropriately.

ii. CSV Management: From the input CSV file an output label CSV was created with a header row ["file","label"], following our required dataset practices.

iii. Data Retrieval and Transformation: For every dataset row, an integer class label is retrieved, and the corresponding image is obtained from the "pixel string" via pixels_to_img.

iv. Image Naming and Saving: Images were labeled in the format {split_name}_{index:06d}.png, using a zero-padded six-digit numbering systemfor the row index to ensure uniqueness and traceability. The image was subsequently stored in the designated split directory.

v. Label Logging: The filename, together with its corresponding label, was added to the label CSV.

3.3.4 Data Loading and Augmentation

The process of preparing data for deep learning model training involves several vital steps, including transformations and augmentations, to improve model performance. In this research, these processes were used for CD-FER-specific purposes; for that, we created a personalized PyTorch dataset class and separate transformation pipelines for both training and evaluation. This section describes the data loading process and explains the significance of data augmentation.

Establishing a reliable data pipeline is essential, especially for cross-domain FER tasks where appearance variations are substantial. By carefully designing the dataset class and transformation strategy, the model receives consistent, well-structured inputs throughout training. These steps collectively enhance robustness and reduce the risk of overfitting to any particular domain.

3.3.4.1 Data Loading Mechanism

Data loading is being done through a custom dataset class, MyFour4All, which extends PyTorch's Dataset class. Through which, image filenames and their corresponding labels from a CSV file utilizing Python's csv module is retrieved, providing a more lightweight option than pandas for basic two-column information.

The constructor takes in arguments such as the path to the CSV file, the directory path for images, along with a training flag that indicates if it's for training or evaluation. Transformations are executed. Establishing a reliable data pipeline is essential, especially for cross-domain FER tasks where appearance variations are substantial. By carefully designing the dataset class and transformation strategy, the model receives consistent, well-structured inputs throughout training. This structured preprocessing ensures that domain-specific discrepancies are minimized before the data reaches the learning model.

3.3.5 Problem Setup and Notation

To address the cross-domain FER challenge introduced in Chapter 1, we formally define the problem setting. We consider two data domains: a labeled *source* domain and an unlabeled *target* domain. Let

$$\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}, \qquad x_i^s \in \mathbb{R}^N, \quad y_i^s \in \{0, 1, 2, 3, 4, 5, 6\}, \tag{3.1}$$

denote the labeled source data and corresponding expression class labels, and let

$$\mathcal{T} = \{x_j^t\}_{j=1}^{N_t}, \qquad x_j^t \in \mathbb{R}^N, \tag{3.2}$$

denote the unlabeled target data. Both domains share the same label space $\mathcal{Y} = \{0, \ldots, 6\}$. We assume the marginal data distributions may differ,

$$p_s(x) \neq p_t(x), \tag{3.3}$$

3.3.5.1 Notation Summary

i. $x_i^s, x_j^t \in \mathbb{R}^N$ represent source and target input images, where N is the flattened pixel dimension.

ii. $y_i^s \in \mathcal{Y}$ denotes the facial emotions label associated with source sample x_i^s.

iii. $\mathcal{Y} = \{0, 1, 2, 3, 4, 5, 6\}$ represents the shared expression label space corresponding to the seven basic emotions.

iv. $p_s(x)$ and $p_t(x)$ represent the marginal data distributions of the source and target domains, respectively.

v. N_s and N_t represent the total number of samples in the source and target datasets.

vi. $f_\theta(\cdot)$ denotes the feature extraction function parameterized by θ, which maps input images to a latent representation used for both emotion classification and domain discrimination.

3.4 Proposed Model: ResEmoteNet-DANN

To address the significant challenge in FER that arises from variability across datasets caused by domain shifts, variations in lighting, posture, resolution, or cultural expression, which make it hard for the models to generalize, our novel ResEmoteNet-DANN model fuses cutting-edge Convolutional Neural Network (CNN) methods [39] with domain adaptation techniques [54], resulting in strong performance across various datasets.

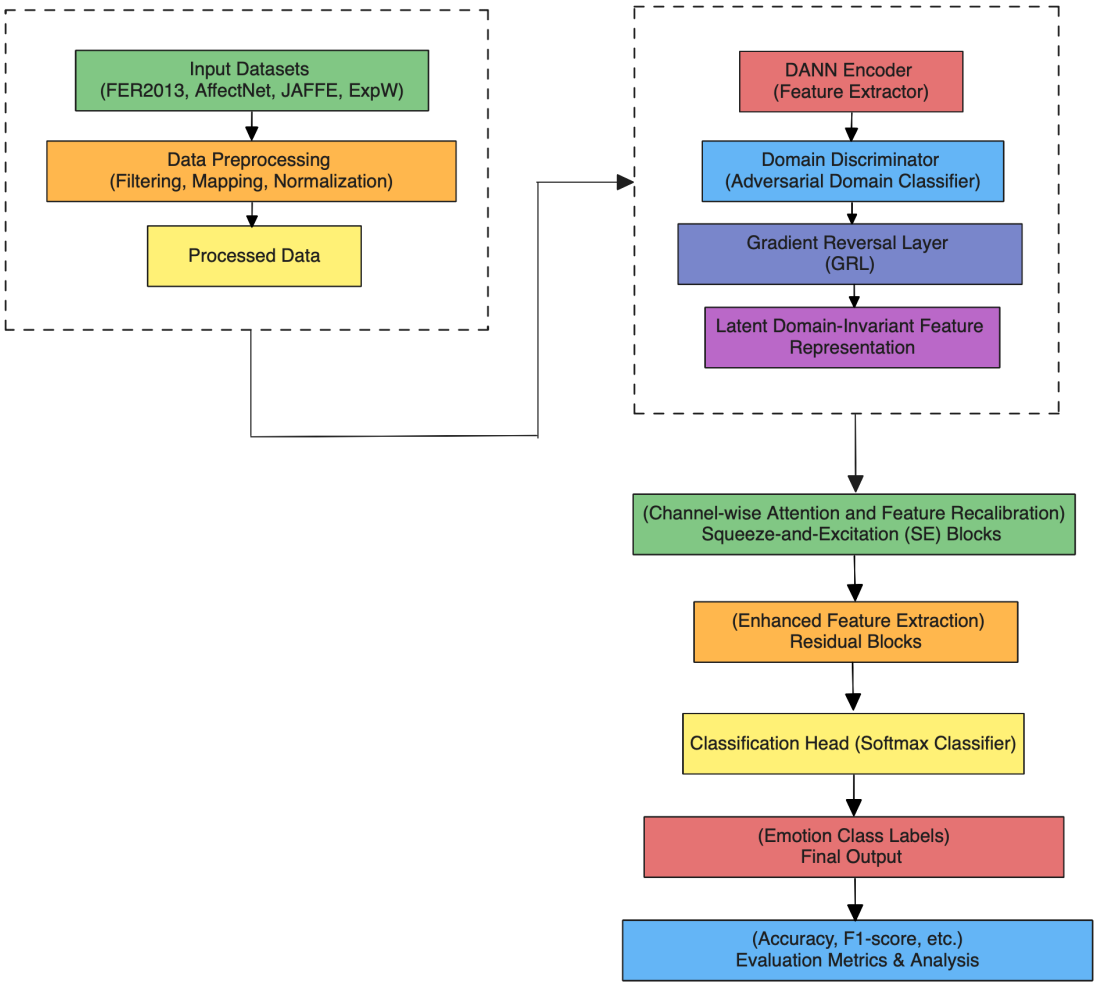

FIGURE 3.6: Pipeline of the Proposed ResEmoteNet-DANN Model

ResEmoteNet-DANN features a Domain Adversarial Neural Network (DANN) encoder [54], Squeeze-and-Excitation (SE) blocks [65], residual connections [20], and a domain discriminator, all integrated into a unified framework. It helps model to extract discriminative and domain-invariant representations.

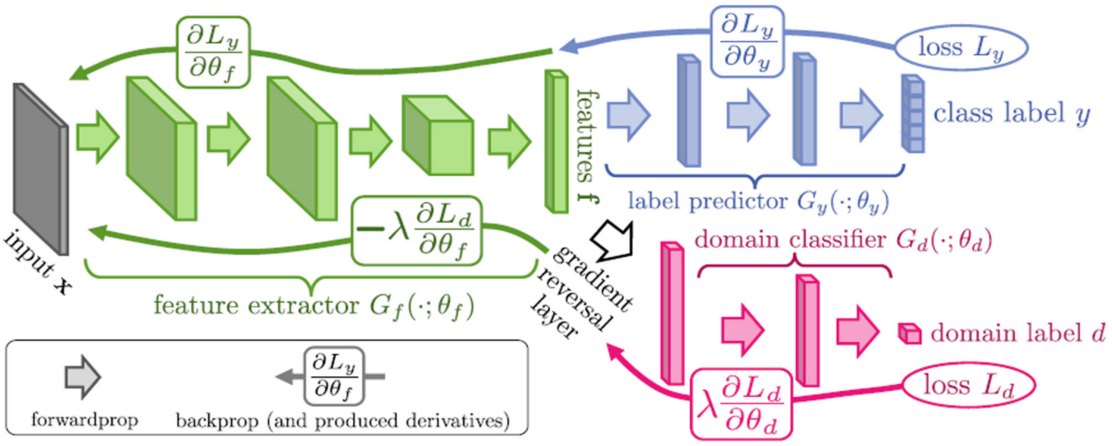

FIGURE 3.7: DANN Framework [54]

3.4.1 DANN Encoder

The DANN encoder serves as the core component of ResEmoteNet-DANN, which is used to extract the domain-invariant features that are appropriate for both emotion classification and domain adaptation [54].

3.4.1.1 Components of DANN Encoder

i. Input: Accepts RGB images with three channels at dimensions (64 × 64). For grayscale datasets, it adjusts to accommodate a single-channel input.

ii. First Convolutional Layer: This layer is based on a (5 × 5) kernel and consists of 64 filters. By using a stride of 1 and padding of 2, it is followed by batch normalization.
In this layer, the ReLU activation function is used, and (2 × 2) max-pooling (with a stride of 2), which reduces spatial dimensions to (32 × 32).

iii. Second Convolutional Layer: The second layer applies a (5 × 5) kernel and consists of 128 filters. Using a stride of 1 and padding of 2, it is combined with batch normalization and uses ReLU as an activation function.
The (2 × 2) max-pooling layer is applied, resulting in feature maps of size (16 × 16) with 128 channels.

3.4.1.2 Functionality

The encoder performs a series of convolutional operations to extract hierarchical features. First, it extracts low-level details (such as edges) and advances to more complex representations (such as facial landmarks). Max-pooling minimizes spatial resolution and emphasizes the most significant features.

3.4.2 Domain Discriminator

The domain discriminator plays a very important role in facilitating domain adaptation. It guarantees that the features obtained from the encoder are independent of domain-specific traits, or we can specify them as domain-irrelevant features [54].

3.4.2.1 Components of Domain Discriminator

i. Domain Discriminator gets the flattened output from the encoder (32,768 dimensions: $(128 \times 16 \times 16)$).

ii. Domain Discriminator consists of two fully connected layers.

iii. The first layer transforms the input to 100 dimensions using ReLU activation, and the second layer gives a single value output that indicates domain prediction (either source or target).

iv. A Gradient Reversal Layer (GRL) is positioned before the discriminator. This inverts gradients during backpropagation with a scaling factor (λ). During training, domain labels (0 = source, 1 = target) are used with binary cross-entropy loss [54].

In GRL, the adversarial loss is multiplied by a scaling factor λ that determines the strength of the domain-confusion signal back propagated to the feature extractor. The feature extractor may prematurely prioritize domain alignment before learning discriminative emotion features from the source domain if λ is set too high at the beginning of training.

Performance on both domains declines as a result of unstable optimization and negative transfer. Instead of applying the adversarial strength consistently from the start, it is progressively increased during training to avoid this problem.

The scaling factor λ typically follows an annealing schedule:

$$\lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \tag{3.4}$$

where p is training progress ($0 \rightarrow 1$), and $\gamma = 10$ as in Ganin et al. [54].

3.4.2.2 Functionality

Domain Discriminator is trained in an adversarial manner, akin to a minimax game [54]. The discriminator aims to identify the domain of the input features, while the encoder tries to generate features that mislead the discriminator. Both try to beat each other and improve as the training process continues. The GRL assists in this by reversing the gradient direction, which helps in aligning feature distributions across different domains.

3.4.3 Squeeze-and-Excitation Blocks

Squeeze-and-Excitation (SE) blocks [65] enhance a network's capability to recognize important features in image data. They utilize channel-wise attention as they evaluate each feature channel individually to determine its importance. This allows the network to extract crucial features while ignoring the less significant ones [65]. Workflow of SE Block is shown in Figure 3.4. By adaptively recalibrating channel responses, SE blocks improve feature representation without a significant increase in computational cost.

This mechanism strengthens discriminative features, which is particularly beneficial for cross-domain facial emotion recognition tasks. Moreover, incorporating SE blocks helps stabilize learning by ensuring that the model consistently prioritizes expressive, domain-invariant cues across varying datasets.

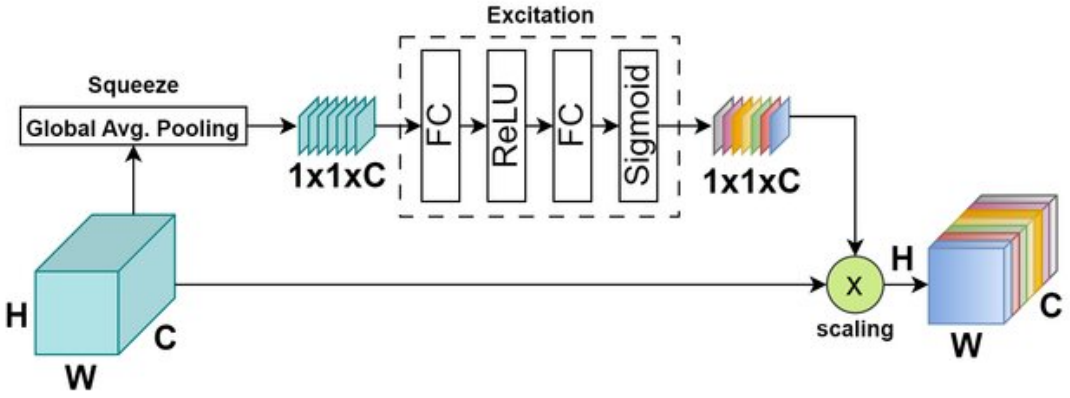

FIGURE 3.8: Squeeze-and-Excitation (SE) Blocks [65]

3.4.3.1 Components of Squeeze-and-Excitation Blocks

i. Squeeze: First, SE blocks "squeeze" the spatial information of each channel into a single representative value through global average pooling. Equation 3.5 shows the mathematical representation of the squeeze function [65].

$$z_c = F_{\text{sq}}(u_c) = \frac{1}{H\,W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{3.5}$$

Notation Explanation:

a. $u_c(i,j)$: value of feature map u at spatial location (i,j) of channel c

b. H, W: height and width of the feature map

c. z_c: output scalar that represents the global descriptor of channel c

d. $F_{\text{sq}}(\cdot)$: global average pooling function

How it works: The squeeze operation averages all $H \times W$ pixel activations of each feature map, reducing $u_c \in \mathbb{R}^{H \times W}$ to a scalar $z_c \in \mathbb{R}$. This results in a compact vector that summarizes the semantic channel strength throughout the entire image.

ii. Excitation: The excitation step adaptively learns the importance of each feature channel by applying two fully-connected layers that model inter-channel dependencies. Equation 3.6 represents the excitation function [65].

$$s = F_{\text{ex}}(z, W) = \sigma\big(g(z, W)\big) = \sigma\big(W_2\,\text{ReLU}(W_1 z)\big) \tag{3.6}$$

Notation Explanation:

a. z: vector of squeezed descriptors $(z_1, \ldots, z_C)$

b. W_1, W_2: learnable weight matrices of the fully-connected layers

c. ReLU: Rectified Linear Unit activation

d. $\sigma(\cdot)$: sigmoid activation that outputs scale factors in $(0, 1)$

e. $s \in \mathbb{R}^C$: learned channel-wise modulation vector

How it works:

a. First (1×1) convolution reduces the number of channels by a factor of 16, followed by a ReLU activation function.

b. Second (1×1) convolution restores the channel count to its original number, and a sigmoid activation produces scaling factors ranging from 0 to 1.

iii. Scaling: Finally, SE blocks use channel-wise multiplication to reweight the initial feature maps:

$$\tilde{u}_c = s_c \cdot u_c,$$

where the significance of channel c is determined by s_c. Discriminative power is increased by amplifying important channels and suppressing irrelevant ones [65].

3.4.3.2 Intuitive Interpretation of SE Block

By dynamically highlighting expression-relevant features (like eyebrow movement) while lowering background noise, SE blocks enable the network to learn "what to pay attention to." This improves FER performance, particularly in difficult cross-domain environments.

3.4.4 Residual Blocks

Residual blocks include "skip connections" or "shortcut connections" [20]. These connections facilitate the direct flow of gradients through the network during backpropagation, helping to mitigate the vanishing gradient issue. As a result, training very deep networks becomes more stable [20]. For example, deeper architectures typically lose crucial facial details (like subtle muscle movements or micro-expressions) when training a deep CNN for facial expression recognition on challenging datasets like AffectNet or ExpW because of weakened gradient signals. The model can retain expression relevant information that might otherwise be lost by using a skip connection, which preserves the original feature map and combines it with the transformed features. This improves both convergence stability and recognition accuracy.

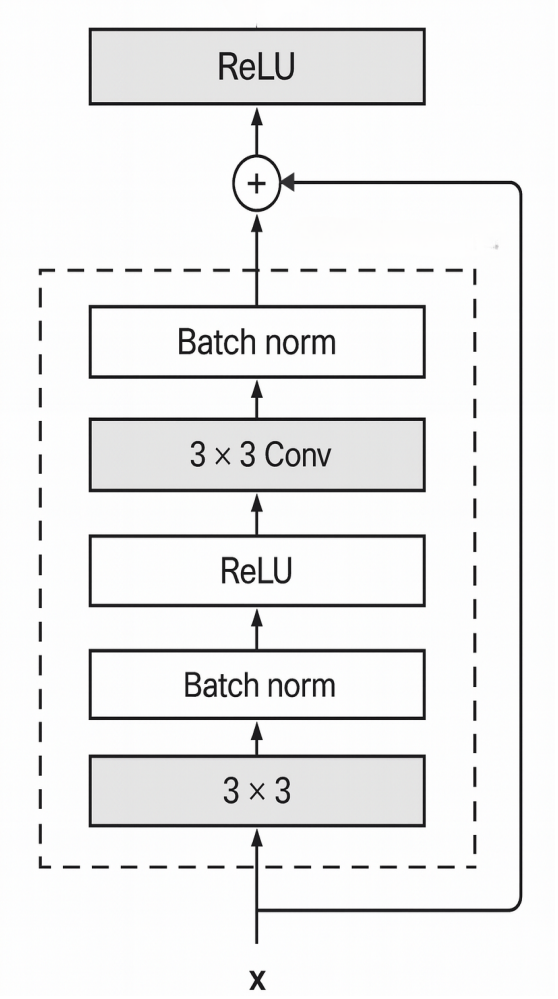

FIGURE 3.9: Structure of Residual Blocks [20]

3.4.4.1 Components of Residual Blocks

1. Structure: Each block comprises two (3×3) convolutional layers equipped with padding, batch normalization, and ReLU activation, plus a shortcut connection that adds the input to the output, along with a (1×1) convolution if there is a difference in channel dimensions [20].

2. Configuration: The three blocks increase channel depth from 128 to 128, 128 to 256, and 256 to 512.

The shortcut connection enables the block to learn residual functions (the differences from the input), which facilitates optimization and allows for deeper architectures [86].

3.4.5 Experimental Settings

The training approach for the ResEmoteNet-DANN for the labeled source domain and unlabeled target domain based on UDA is discussed in detail in this section. Including the loss functions, optimizer, learning rate scheduler, training procedure, and hyperparameters used. This will be able to highlight their significance and advantages in attaining effective results for CD-FER.

3.4.5.1 Loss Functions

In the proposed framework we used two different loss functions because the model simultaneously optimizes two different learning objectives: (1) accurate facial expression classification using labeled source data, and (2) correctly distinguishing between the two domains (source and target) in order to guide adversarial learning. A single loss function cannot accomplish both goals.

i. Cross-Entropy Loss with Label Smoothing: This loss is applied only to the labeled *source* domain. It trains the classifier to correctly recognize seven emotion categories. This loss function includes label smoothing with a factor

of 0.1. Label smoothing mitigates overconfidence caused by class imbalance in predictions by softening the target distribution. This helps for enhancing generalization [87]. The loss is defined as:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} \left((1-\epsilon) y_c + \frac{\epsilon}{C} \right) \log(p_c) \tag{3.7}$$

where $C = 7$ is the number of emotion classes, y_c is the true label, p_c is the predicted probability, and $\epsilon = 0.1$ is the smoothing parameter.

Role in the proposed solution: This loss function encourages the model to learn highly discriminative expression-related features from the source domain.

ii. Binary Cross-Entropy with Logits Loss: Used to differentiate domains, this loss function aids in adversarial training, as we need to classify features from the source (label 1) and target (label 0) domains. A sigmoid activation combined with binary cross-entropy ensures numerical stability, represented as:

$$\mathcal{L}_{\text{domain}} = -\frac{1}{N} \sum_{i=1}^{N} \left[d_i \log(\sigma(z_i)) + (1-d_i) \log(1-\sigma(z_i)) \right] \tag{3.8}$$

where d_i is the domain label, z_i is the domain discriminator's output, and σ is the sigmoid function [54].

Role in the proposed solution: This ensures that the framework learns to identify whether a feature comes from the source or target domain and promotes domain alignment while enforcing adversarial confusion.

The overall optimization objective combines the two losses as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{domain}} \tag{3.9}$$

3.4.5.2 Optimizer

The model parameters, including both the ResEmoteNet-DANN and the domain discriminator, were optimized using Stochastic Gradient Descent (SGD) with the

following settings:-

i. Learning Rate: A learning rate of 0.005 was chosen empirically. During experimentation, three candidate learning rates were considered, such as 0.01, 0.001, and 0.005. Using a learning rate of 0.01 led fast convergence but was not good in terms of accuracy, whereas 0.001 drastically slowed convergence. The value of 0.005 gave the most stable convergence behavior and thus the best validation accuracy was achieved.

ii. Momentum: We set the momentum to 0.95. This value offers a balanced trade-off: it was sufficient to take advantage of the acceleration and smoothing effects of the past gradient that makes adversarial training more stable but not too high for the model to become completely insensitive to new gradient directions. In fact, the latest studies have demonstrated that with low learning rates, which are commonly used in deep-FER training, the benefit of the momentum is very little, yet the properly calibrated momentum still helps the convergence process by reducing the impact of noisy updates and preventing dramatic swings in weight [88, 89]. Thus 0.95 provides a reasonable compromise between stability and adaptability.

iii. Weight Decay (L2 Regularization): Weight decay was set to 1×10^{-4} to reduce overfitting by penalizing large weight values. This is implemented as L2 regularization, modifying the loss function as:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{total}} + \lambda \sum_{k} \|w_k\|^2 \tag{3.10}$$

where w_k are model parameters and $\lambda = 1 \times 10^{-4}$ controls the strength of regularization [90, 91].

3.4.5.3 Learning Rate Scheduler

A ReduceLROnPlateau scheduler was used to dynamically decrease the learning rate as training progressed. This scheduler adjusts the learning rate upon validation accuracy in the target dataset and reduces the learning rate by a factor of 0.1

after no improvement for 7 epochs.

By decreasing the learning rate when performance plateaus, the scheduler avoids overshooting the loss minimum and allows for finer weight adjustments in later stages of training, which can help converge to a better solution on the target domain [92]. This also makes optimization computationally more efficient by reducing unnecessary iterations with an overly high learning rate that is not aligned with how much the model has learned.

3.4.5.4 Hyperparameters

The following hyperparameters were used to control the training process: Batch Size: 16 (memory-efficient and gradient-stable), Learning Rate: 0.005 (the initial step size that is adjusted by the scheduler), Momentum: 0.95 (to speed up convergence [93]), Weight Decay: 1×10^{-4} , Number of Epochs: 200 (adequate training duration with early stopping to avoid overfitting), and Patience: 25 (early stopping based on the validation accuracy of target dataset).

These values were chosen to optimize for computational efficiency, convergence behavior, and model performance by ensuring that the batch size was practical given GPU memory constraints, the learning rate and momentum facilitated effective optimization, and patience was set so that the model had sufficient time to improve before termination.

Chapter 4

Results and Evaluations

In this chapter, we present the experimental findings of our `ResEmoteNet-DANN` model for FER with domain adaptation from the labeled source domain to the target domain. The `ResEmoteNet-DANN` framework is based on a DANN encoder, SE blocks, and residual connections. This framework is specifically designed to achieve high performance in domains with shifted scenarios, where datasets vary in characteristics such as image quality, lighting, and pose variations [30, 40]. This chapter provide a detailed analysis of the model's training and validation behavior across multiple datasets, highlighting its ability to learn domain-invariant features while maintaining robustness against common FER challenges. Additionally, it discuss the impact of various hyperparameter settings and architectural components on cross-domain generalization.

4.1 Performance Metrics

For the evaluation of ML models, performance metrics serve as essential benchmarks. These metrics aid in the process of model selection and optimization by quantifying elements like predictive accuracy, robustness, and efficiency [27, 28]. These performance metrics serve the purpose of highlighting overfitting and underfitting by measuring the model's performance on previously unseen data.

In this research study, we utilized Accuracy, Precision, Recall, and F1-score as our evaluation metrics [27, 40]. Additionally, these metrics provide a comprehensive view of the model's ability to correctly identify each emotion class and handle class imbalances. By examining multiple metrics together, we can better understand not only overall performance.

4.1.1 Accuracy

Accuracy serves as a key metric for the evaluation of classification models in ML. This indicates the percentage of accurately categorized Examples throughout all classes. It is a simple matrix of the overall assessment of a model. Accuracy is useful for balanced datasets, and it can be deceptive in situations with class imbalance, where models might obtain high scores by predominantly predicting the majority classes [40]. Despite this drawback, accuracy is still commonly used as an important metric.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

4.1.2 Precision

Precision evaluates the reliability of positive predictions by calculating the ratio of true positives to all instances predicted as positive. In situations where false positive results can be expensive, precision becomes more important, like in medical diagnosis or security systems. High precision means that when the model forecasts a positive class (e.g., a certain facial expression), it is probably accurate. However, precision is most insightful when paired with recall [40].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

4.1.3 Recall

Recall is a performance metric that is used to assess a model's capability to correctly identify all relevant positive cases in a data set. Measures the ratio of actual positive instances that are accurately predicted as positive. In scenarios where failing to detect positive cases can lead to serious consequences, such as in medical diagnosis or safety-related applications, recall is particularly important. There is a drawback to recall that it might result in higher number of false positives, high recall suggests that the model is effective at capturing all target classes [27, 40]. However, recall alone does not provide information about the precision of these predictions, so it must be considered alongside other metrics for a complete evaluation. In combination with precision, recall helps in computing the F1-Score, offering a balanced measure of a model's overall performance.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

4.1.4 F_1-Score

The F_1-Score combines precision and recall into a single metric and offers a fair evaluation of model performance. In circumstances where precision and recall have an inverse relationship, F1-Score handles the fairness by taking the harmonic mean of these complementary metrics. F1-Score favors models that preserve both high detection accuracy and prediction reliability. It is especially useful in class-imbalanced datasets, which are typical in FER [30, 40]. It also provides insight into how well the model balances false positives and false negatives, which is crucial for reliable emotion recognition. Additionally, F1-Score allows for meaningful comparisons across different models and datasets, ensuring consistency in evaluation. By emphasizing both precision and recall, it helps guide model improvements and hyperparameter tuning in FER tasks.

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{4.4}$$

4.2 Evaluation Dataset Insights

4.2.1 FER 2013 Data Distribution

This research used FER 2013 as the base dataset in our study and in this section we will highlight the structure of FER2013 dataset. Table 4.1 and Figure 4.1 shows the class distribution of that data set.

TABLE 4.1: FER2013 Dataset Split

Class	Train	Validation	Test
happy	7215	895	879
surprise	3171	415	416
sad	4830	653	594
anger	3995	467	491
disgust	436	56	55
fear	4097	496	528
neutral	4965	607	626

4.2.2 AffectNet

AffectNet used as source labeled dataset in the study and this section highlight the structure. Table 4.2 shows the class distribution of that data set.

TABLE 4.2: AffectNet Dataset Split

Class	Train	Validation	Test
neutral	74,874	3,500	3,500
happy	134,415	5,500	5,500
sad	25,459	1,500	1,500
surprise	14,090	1,500	1,500
fear	6,378	1,500	1,500
disgust	3,803	1,000	1,000
anger	24,882	1,500	1,500

4.2.3 ExpW

ExpW (Expression in-the-Wild) is large-scale dataset collected from Google images search used in cross-domain experiments. The class distribution for the 80/10/10 train/val/test split is presented in Table 4.3.

TABLE 4.3: ExpW dataset Splits

Class	Train	Validation	Test
angry	6,074	759	759
disgust	3,152	394	394
fear	4,378	547	547
happy	17,551	2,194	2,194
sad	5,080	635	635
surprise	7,147	893	893
neutral	9,550	1,194	1,194
Total	**52,932**	**6,616**	**6,616**

4.2.4 JAFFE

The JAFFE (Japanese Female Facial Expression) dataset was used as a target domain. It consists of 213 grayscale images of 10 Japanese female subjects. The distribution is shown in Table 4.4.

TABLE 4.4: JAFFE Dataset Split

Class	Train	Validation	Test
angry	24	3	3
disgust	23	3	3
fear	26	3	3
happy	25	3	3
sad	25	3	3
surprise	24	3	3
neutral	24	3	3
Total	**171**	**21**	**21**

4.2.5 Comparison with EM-UDA: Binary Positive and Negative Emotion Classification

To establish the performance of our model on cross-domain binary facial emotion recognition (FER), we evaluate our findings against the latest EM-UDA. For binary positive/negative emotion classification Both approaches use AffectNet as the labeled source domain and FER2013 as the unlabeled target domain.

EM-UDA uses a VGG-19-based pre-training weights as the backbone coupled with adversarial UDA framework and reports the following performance on FER2013: Accuracy 74.55%, Precision 75.26%, Recall 74.49%, and F1-score 74.87%.

In comparison, our ResEmoteNet-DANN method achieves significantly higher performance on the same cross-domain setting, with Accuracy 80.72%, Precision 75.14%, Recall 94.12%, F1-score 83.56%, and AUROC 91.28%. This represents a gain of over **6 percentage** points in accuracy and nearly **9 points** in F1-score compared to EM-UDA, we are also achieving higher recall.

These results answer the research question that using Modified ResEmoteNet with DANN can enhance the effectiveness of the Model on CD-FER. Figure 4.1 shows the confusion matrices of both FER 2013 and AffectNet on Unseen data.

TABLE 4.5: Binary (Positive/Negative) FER on AffectNet→FER2013 EM-UDA vs. Baselines and Proposed

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Transfer Learning (AffectNet→FER2013)	69.54	68.97	71.61	70.26
EM-UDA (AffectNet→FER2013)	74.55	75.26	74.49	74.87
Supervised Upper Bound (FER2013→FER2013)	82.13	80.67	83.02	81.83
Proposed Model	**80.72**	**75.14**	**94.12**	**83.56**

TABLE 4.6: Backbone Architecture Selection on AffectNet→FER2013 (Binary)

Backbone / Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
DenseNet201	77.20	77.37	76.75	77.06
EfficientNet-B3	79.10	79.12	78.96	79.04
ResNet50	78.60	78.88	78.57	78.73
VGG-16	79.20	78.96	79.28	79.12
VGG-19	79.70	80.32	79.22	79.76
Proposed Model	**80.72**	**75.14**	**94.12**	**83.56**

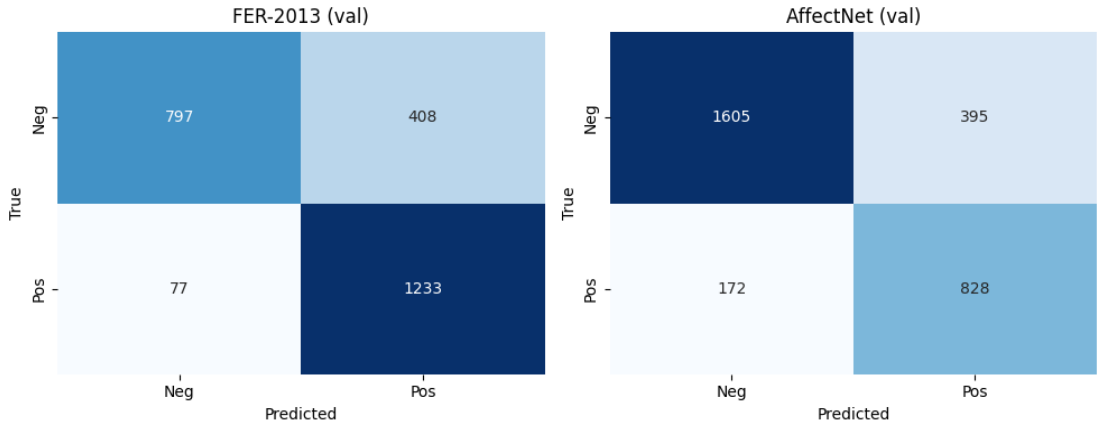

FIGURE 4.1: Confusion matrices of both FER 2013 and AffectNet

4.2.6 Training and Convergence Analysis

We trained the model using FER2013 as the labeled source domain and JAFFE as the unlabeled target domain in order to assess the performance of the suggested ResEmoteNet-DANN model in a cross-domain context. Depending on the accuracy of the validation on FER2013, the training process was terminated early after up to 200 epochs.

A steady improvement was observed during training in both training accuracy and validation accuracy on FER2013, with the model converging within approximately 80 epochs. As JAFFE dataset is very small We used Generative Adversarial Networks (GAN) to generate 2000 new training samples and merged with the orignal JAFFE samples so the model can learn better because of the bigger training data. In terms of cross-domain performance, the model achieved a maximum validation accuracy of 46.16% on the JAFFE dataset. Although there is a clear domain gap (as expected in unsupervised domain adaptation tasks), the model demonstrates the ability to transfer learned representations from FER2013 to JAFFE, yielding significantly better-than-random results on the target domain.

The detailed training information is presented in (Table 4.8).

Overall, these results indicate that the proposed framework can effectively use labeled data from a large-scale source dataset (FER2013) to improve performance

TABLE 4.7: Selected Validation Accuracies (FER2013→JAFFE) Across Epochs. Top Row Representing Production Baseline When Trained on FER2013 Only (No UDA)

No. of Epoch	**Val Acc (FER2013)**	**Val Acc (JAFFE)**
Baseline (FER-only)	—	**0.2503**
1	0.5223	—
5	0.5061	—
10	0.5326	—
20	0.5231	0.3226
29	0.5270	0.3871
40	0.5334	0.4194
50	0.5443	0.3871
55	0.5418	0.4194
60	0.5446	0.4194
69	0.5499	0.3871
70	**0.5504**	0.4194
78	0.5812	**0.4616**
80	0.6037	0.4194
90	0.6481	0.3548
100	**0.6994**	0.3548

on a smaller, unlabeled target domain (JAFFE). This proves the ability of UDA techniques for CD-FER.

4.2.7 Comparison of Baseline and Domain-Adaptive Models: In-domain vs. Cross-domain Accuracy

To rigorously evaluate the effect of domain adaptation, we compared the performance of the baseline ResEmoteNet (trained without domain adaptation). Proposed ResEmoteNet-DANN (with domain-adversarial adaptation) across the

in-domain and cross-domain settings. Table 4.8 summarizes the results.

In the in-domain setting, ResEmoteNet achieves an accuracy of **79%** on FER2013 and **72%** on AffectNet when trained and tested on the same dataset. However, without adaptation, cross-domain accuracy drops sharply: training on FER2013 and testing on AffectNet yields just **32%**, and training on AffectNet and testing on FER2013 gives **38%** accuracy.

By contrast, when using our ResEmoteNet-DANN with domain-adversarial training, cross-domain generalization improves dramatically. In AffectNet→FER2013 setting, our method achieves a best validation accuracy of **56.8%** on FER2013—an absolute improvement of **+18.8 percentage points** over the baseline (no adaptation).

It demonstrates the effectiveness of our domain adaptation strategy in mitigating domain shift and enhancing model robustness.

TABLE 4.8: Comparison of In-Domain and Cross-Domain Accuracy (%) for Baseline and Domain-Adaptive Models

Model	**Train → Test**	**In-domain**	**Cross-domain**
ResEmoteNet (no adaptation)	FER2013 → FER2013	79	–
	AffectNet → AffectNet	72	–
ResEmoteNet (no adaptation)	FER2013 → AffectNet	–	32
	AffectNet → FER2013	–	38
ResEmoteNet-DANN	FER2013 → AffectNet	66	**58**
	AffectNet → FER2013	68	**56.8**

As shown, the proposed ResEmoteNet-DANN substantially narrows the domain gap, delivering significant improvements over the non-adaptive baseline in cross-domain emotion recognition scenarios.

The comparison in Table 4.9 shows that our proposed ResEmoteNet-DANN achieve significant improvement over the baseline and prior domain adaptation methods. When transferring from FER2013 to AffectNet, proposed model reaches 58.04%.

TABLE 4.9: Cross-domain FER Accuracy (%) Comparison on FER2013 and AffectNet as Target Domains.

Method	**FER2013 → AffectNet**	**AffectNet → FER2013**
Source-Only [LA-CMFER [60]]	41.86	42.24
DANN [LA-CMFER [60]]	49.86	51.46
DUML [LA-CMFER [60]]	52.46	56.56
[LA-CMFER] [60]	53.26	57.40
ResEmoteNet-DANN	**58.04**	**56.80**

4.2.8 Comparison on AffectNet to ExpW Adaptation

To further evaluate the efficacy of our proposed ResEmoteNet-DANN framework, we performed cross-domain adaptation experiment from **AffectNet** (source) to **ExpW** (target). This scenario is particularly challenging due to the significant domain shift from the largely controlled, web-collected AffectNet to the fully "in-the-wild" and highly variable ExpW dataset. We compare results against the current state-of-the-art method,AGRA [94], which employs a complex architecture with stacked graph Convolutional Networks for holistic-local adversarial adaptation, resulting in a high-parameter model. The comparative results are presented in Table 4.10.

TABLE 4.10: Cross Domain FER Accuracy (%) Comparison on the Challenging AffectNet → ExpW Adaptation Task

Method	**Accuracy on ExpW (%)**	**Model Complexity**
Source-Only (No Adaptation)	45.21	-
DANN [54]	53.17	Medium
AGRA (ResNet50 Backbone) [94]	**65.03**	**Very High**
ResEmoteNet-DANN (Ours)	**62.90**	**Low**

As shown in Table 4.10, the AGRA framework achieves a superior accuracy of 65.03%, leveraging its extensive multi-scale feature adaptation. Our proposed ResEmoteNet-DANN model attains a competitive accuracy of 62.90% on this difficult task. While this is 2.13 percentage points lower than AGRA, this performance must be interpreted in the context of a significant disparity in model complexity and computational demand.The AGRA architecture consists of stacked intra and inter domain graph convolutional networks and requires millions of parameters,

because of the high model complexity substantial GPU resources for training are required.That making it potentially prohibitive for resource-constrained environments. In contrast, our ResEmoteNet-DANN model utilizes a single, efficient adversarial stream, resulting in a much lighter and computationally efficient architecture. The achieved performance of 62.9% represents a strong result, demonstrating a major improvement of over 17 percentage points over the source-only baseline and nearly 10 points over a standard DANN. This indicates that our method offers an excellent trade-off between performance and efficiency.

Table 4.11 compares the main CD-FER methods, making it easy to see three things at once: how they handle domain shift, their scores on the common transfers (A→F, F→A, A→E, F→J), and their training load. Large, multi-branch models like AGRA and LA-CMFER achieve extra accuracy but require significant computing power. DANN is the classic choice—it offers good value but often falls behind the heavier setups. ResEmoteNet-DANN is simple, quick to train and performs well on multi-class transfers, with a strong advantage on **binary A→F** (80.72% compared to EM-UDA's 74.55%). In summary, we sacrifice a bit of peak accuracy for much lower complexity.

TABLE 4.11: Comparison with cross-domain FER

Method	**Domain shift handling**	**A→F**	**F→A**	**A→E**	**F→J**	**Model complexity (qual.)**
ResEmoteNet (no UDA)	None (in-domain only)	38.0	32.0	–	25.0	Low: lightweight CNN + SE; efficient
DANN [54]	Adversarial GRL feature alignment	51.46	49.86	–	–	Low–Med: simple add-on; stable w/ tuning
LA-CMFER [60]	Dual-level (global/local) alignment; multi-source training	57.40	53.26	–	–	High: twin-branch + consistency; multi-source cost
AGRA [94]	Holistic–local adversarial w/ stacked GCNs	–	–	**65.03**	–	Very High: multi-graph stacks; heavy GPU
EM-UDA (binary)[57]	VGG-19 + adversarial UDA (pos/neg)	**74.55**†	–	–	–	High: VGG-19 backbone; param-heavy
ResEmoteNet-DANN	CNN + GRL (single-stream)	**56.80**	**58.04**	**62.90**	**46.16**	**Low**: single adversarial head; fast/cheap

Notes. A→F: AffectNet→FER2013; F→A: FER2013→AffectNet; A→E: AffectNet→ExpW; F→J: FER2013→JAFFE. "–" = not reported.
† EM-UDA reports **binary** (pos/neg) results on A→F: Acc 74.55%, Prec 75.26%, Rec 74.49%, F1 74.87%. Our binary A→F is **Acc 80.72%, Prec 75.14%, Rec 94.12%, F1 83.56%**.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study introduced a model for recognizing facial emotions across different domains (FER) by utilizing unsupervised domain adaptation. We presented the ResEmoteNet-DANN model, where we merged domain-adversarial training through DANN with deep residual networks and used SE networks for selective most relevant features. Through this framework, we have established that using DANN as a domain-irrelevant feature extractor enhances the accuracy of a model for CD-FER, so this is a reliable solution for the domain shift problem in FER. We also used self-supervised learning (SSL) pretraining for domain-invariant feature extraction coupled with DANN, and it proved that this can even achieve better results than using DANN alone. We conducted thorough experiments on established datasets such as FER2013, AffectNet, JAFFE, and RAF-DB, and during those experiments, our approach showed significant enhancements in cross-domain recognition accuracy when compared to the models that did not use a domain-invariant feature extractor.

Findings indicate that while traditional models perform effectively in in-domain contexts (e.g., 79% on FER2013, 72% on AffectNet), their performance significantly decreases in cross-domain situations due to variations in data distribution.

The ResEmoteNet-DANN framework we proposed successfully reduces this domain gap by attaining a cross-domain accuracy of 56.8% (AffectNet → FER2013) and 58.04% (FER2013 → AffectNet), These performance figures greatly surpass the performance of models that lack domain adaptation techniques. These results emphasize the significance and efficacy of domain-adversarial techniques and self-supervised learning techniques in the development of more generalizable and resilient FER systems.

5.2 Future Work

Despite these advancements, there are some areas where improvements are needed, opening up the way for further exploration:-

i. Deeper Domain Gap Analysis: Although our approach reduces the domain gap, additional efforts are required to comprehend and reduce remaining discrepancies, particularly in scenarios with significant class imbalance of datasets.

ii. Self-supervised and Semi-supervised Learning: Even though we used SSL for the most significant feature extractor, the incorporation of different frameworks of SSL or semi-supervised learning could enhance robustness and provide a comprehensive analysis and comparisons between these techniques. This is more needed in scenarios where labeled data is limited or noisy, especially regarding less-explored emotion classes.

iii. Real-world Evaluation and Deployment: Future research should evaluate model effectiveness in real-world situations to understand the model's effectiveness in real-world settings, where variations in lighting, occlusion, ethnicity, and spontaneous expressions exist. Even for edge devices, efficient frameworks can be researched based on these techniques.

iv. Multimodal and Temporal FER: Expanding the current framework to leverage multimodal data (e.g., audio, text, physiological signals) and temporal

information from video could lead to even more accurate and context-aware emotion recognition.

v. Emotion Intensity and Compound Expressions: Compound expressions involve precise evaluation of emotional intensity in images, involve multiple emotions at the same time, and the identification of complex or blended emotions is an encouraging area for the advancement of future FER systems.

In short, this research represents an important advancement in the development of CD-FER models and also introduces various promising roadmaps for progressing affective computing toward practical application in the real world.

Bibliography

[1] R. Raj and I. Demirkol, "An improved facial emotion recognition system using convolutional neural network for the optimization of human robot interaction," *Scientific Reports*, vol. 15, p. 38940, 2025.

[2] C. C.-C. J.-H. C. Y.-C. Huang, Zi-Yu and Chung, "A study on computer vision for facial emotion recognition," *Scientific Reports*, vol. 13, 2023.

[3] D. A. Sauter, "The nonverbal communication of positive emotions: An emotion family approach," *Psychological Review*, vol. 124, pp. 413–435, 2017.

[4] R. M. Spielman, W. J. Jenkins, and M. D. Lovett, *Psychology 2e*. Houston, TX: OpenStax, second ed., 2020. Chapter 10, Section 4: Emotion.

[5] R. Nandy, K. Nandy, and S. T. Walters, "Relationship between valence and arousal for subjective experience in a real-life setting for supportive housing residents: Results from an ecological momentary assessment study," *JMIR Formative Research*, vol. 7, p. e34989, 2023.

[6] A. M. Ismael, Ö. F. Alçin, K. H. Abdalla, and A. Şengür, "Two-stepped majority voting for efficient EEG-based emotion classification," *Brain Informatics*, vol. 7, pp. 1–12, Dec. 2020.

[7] A. Cîrneanu, D. Popescu, and L. Iordache, "Facial emotion recognition: A survey," *Sensors*, vol. 23, pp. 2456–2478, Mar. 2023.

[8] S. Ghosh, A. Dhall, and N. Sebe, "Automatic detection of driver fatigue using facial cues," *Scientific Reports*, vol. 13, pp. 1–15, Oct. 2023.

[9] N. V. Thanh, "Facial emotion recognition for healthcare applications," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 1345–1356, Mar. 2023.

[10] A. Talaat, "FER-based autism diagnosis via facial expression analysis," *Journal of Autism and Developmental Disorders*, vol. 53, pp. 3012–3025, Aug. 2023.

[11] Y. Hou, Z. Zhang, and C. Li, "Emotion recognition in intelligent classrooms," *IEEE Transactions on Education*, vol. 65, pp. 189–198, May 2022.

[12] J. Wang, H. Liu, and X. Chen, "Real-time student engagement via FER," *Education and Information Technologies*, vol. 25, pp. 2567–2583, July 2020.

[13] S. Geng, L. Meng, and Z. Dou, "FER-enabled IoT system for education," *IEEE Internet of Things Journal*, vol. 9, pp. 8456–8467, June 2022.

[14] R. Srivastava and S. Bag, "Sentiment analysis using FER in marketing," *Journal of Business Research*, vol. 156, pp. 113–125, Feb. 2023.

[15] D. Wright, "Vibraimage for security applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, pp. 4123–4135, July 2023.

[16] M. R. Mohammadi and A. Khodabandeh, "A survey on traditional and deep learning-based FER," *Symmetry*, vol. 11, p. 1189, Sept. 2019.

[17] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, E. Xie, L. Romaszko, B. Xu, Z. Zhang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Proceedings of the International Conference on Neural Information Processing*, (Daegu, South Korea), pp. 117–124, Nov. 2013.

[18] H. Gholamalinezhad and H. Khosravi, "Deep learning in facial expression recognition," *Journal of Visual Communication and Image Representation*, vol. 70, pp. 102–115, July 2020.

[19] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, NV, USA), pp. 770–778, June 2016.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, MA, USA), June 2015.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, (Long Beach, CA, USA), pp. 5998–6008, Dec. 2017.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, May 2021.

[24] R. W. Picard, "Affective computing: Challenges," *International Journal of Human-Computer Studies*, vol. 59, pp. 55–64, July 2003.

[25] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, pp. 807–813, May 2010.

[26] B. Sorscher, R. Geirhos, and S. Hochreiter, "Scaling laws in neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 3012–3023, June 2023.

[27] L. Xue, A. K. Jain, and R. Sukthankar, "Neural scaling for FER," *Pattern Recognition*, vol. 145, pp. 108–120, Jan. 2024.

[28] T. Marion, J. Lee, and P. Domingos, "Large-scale training for FER," *Machine Learning*, vol. 112, pp. 789–805, Mar. 2023.

[29] Y. Zhang, X. Li, T. Tan, and chin chia chen, "Generalization challenges in FER," *Computer Vision and Image Understanding*, vol. 230, pp. 103–115, May 2023.

[30] Z. Zhang, Y. Zhang, and C. C. Chen, "In-the-wild FER performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 956–970, Feb. 2024.

[31] H. Wang, D. Gong, and Z. Li, "Real-world FER dataset evaluation," *Journal of Real-Time Image Processing*, vol. 20, pp. 12–25, Feb. 2023.

[32] S. Jahan, M. S. Islam, and A. K. Jain, "Reevaluating Multi-PIE for modern FER," *IEEE Access*, vol. 11, pp. 45678–45690, May 2023.

[33] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (San Francisco, CA, USA), pp. 94–101, June 2010.

[34] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, NV, USA), pp. 5562–5570, June 2016.

[35] A. Mollahoseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, Jan. 2019.

[36] S. Li, J. D. W. Deng, and chin chia chen, "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*, (Honolulu, HI, USA), pp. 2852–2861, July 2017.

[37] T. Zhang, W. Zheng, and Z. Cui, "Expression in the wild: The ExpW dataset," *IEEE Transactions on Affective Computing*, vol. 9, pp. 456–467, Oct. 2018.

[38] S. Agung, E. Satrio, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using CNN," *Scientific Reports*, vol. 14, p. 14429, June 2024.

[39] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, "Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition," *IEEE Signal Processing Letters*, vol. 32, pp. 491–495, 2025.

[40] P. Kulkarni *et al.*, "Demographic bias as confounding noise in static facial expression recognition," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1123–1145, 2024.

[41] S. Li, W. Deng, and J. Du, "Facial expression recognition with data augmentation and compact feature learning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1957–1961, 2018.

[42] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using CNNs and CRFs," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1521–1529, 2017.

[43] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643–660, June 2001.

[44] A. Author, "Camera angle effects in facial expression recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 500–510, Feb. 2024.

[45] X. Zhang, Y. Li, and Z. Wang, "Impact of image resolution on deep learning-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 15, pp. 123–135, Apr. 2024.

[46] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, p. Article 46, Apr. 2023.

[47] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[48] B. Researcher, "Cultural variations in facial expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1000–1005, June 2024.

[49] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *ArXiv*, 2023. Preprint arXiv:2308.01265.

[50] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9887–9903, 2022.

[51] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *arXiv preprint arXiv:2401.11085v1*, 2024.

[52] Y. Gao, Y. Cai, X. Bi, B. Li, S. Li, and W. Zheng, "Cross-domain facial expression recognition through reliable global–local representation learning and dynamic label weighting," *Electronics*, vol. 12, no. 21, p. 4553, 2023.

[53] Y. Fang, P.-T. Yap, W. Lin, H. Zhu, and M. Liu, "Source-free unsupervised domain adaptation: A survey," *Neural Networks*, 2024.

[54] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, pp. 2097–2030, 2016.

[55] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

[56] Z. Guo, B. Wei, J. Liu, X. Liu, Z. Zhang, and Y. Wang, "USTST: Unsupervised self-training similarity transfer for cross-domain facial expression recognition," *Multimedia Tools and Applications*, vol. 83, oct 2023.

[57] P. R. Jain, S. M. K. Quadri, and A. Khattar, "Em-uda: Emotion detection using unsupervised domain adaptation for classification of facial images," *IEEE Access*, vol. 12, pp. 140262–140276, 2024.

[58] C. C. C. X. Peng *et al.*, "AU-guided unsupervised domain-adaptive facial expression recognition (AdaFER)," *Applied Sciences*, vol. 12, no. 9, p. 4366, 2022.

[59] Y. Zhu, J. Ai, W. Xue, M. Wu, S. Yang, W. Jia, and M. Hu, "Cross-domain facial expression recognition: Bi-directional fusion of active and stable information," *Engineering Applications of Artificial Intelligence*, vol. 149, p. 110357, June 2025. Also referred to as FER-DAS.

[60] Y. Zhang, X. Li, J. Wang, and T. Tan, "Learning with alignments: Tackling the inter- and intra-domain shifts for cross-multidomain facial expression recognition," *arXiv preprint arXiv:2407.05688*, 2024.

[61] B. Malik, A. R. Kashyap, M.-Y. Kan, and S. Poria, "Udapter: Efficient domain adaptation using adapters," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2249–2263, 2023.

[62] A. Abedi, Q. M. J. Wu, N. Zhang, and F. Pourpanah, "Euda: An efficient unsupervised domain adaptation via self-supervised vision transformer," *arXiv preprint arXiv:2407.21311v1*, 2024.

[63] Y. Li *et al.*, "Deep margin-sensitive representation learning for cross-domain facial expression recognition," *IEEE Transactions on Multimedia*, vol. 24, no. 11, pp. 6088–6100, 2022.

[64] T. Liu *et al.*, "Cross-domain facial expression recognition via disentangling identity representation," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1213–1219, 2023.

[65] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[66] T. Kopalidis, G. Petropoulos, and S. Asteriadis, "Advances in facial expression recognition: A survey of methods, benchmarks, models and datasets," *Information*, vol. 15, no. 3, p. 135, 2024.

[67] S. Ullah, J. Ou, Y. Xie, and W. Tian, "Facial expression recognition (fer) survey: A vision, trends, and future perspective," *PeerJ Computer Science*, vol. 10, p. e11157619, 2024.

[68] A. A. Alhussan, M. Khan, and S. M. D. Alkahtani, "A comprehensive survey on facial expression recognition techniques," *Sensors*, vol. 25, no. 12, p. 3832, 2025.

[69] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978. Updated 2002.

[70] Y. Chen, S. Lee, and S. Y. Chen Chia Chen, "Identity-invariant expression learning via adversarial disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15387–15397, 2024.

[71] A. Sharma *et al.*, "Bias mitigation in facial analysis: A survey," *ACM Computing Surveys*, vol. 57, no. 1, p. Article 14, 2023. Early Access: 2024.

[72] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by DE-expression residue learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT), 2018.

[73] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[74] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC) Workshop on Emotion*, pp. 65–70, 2010.

[75] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR)*, pp. 211–216, 2006.

[76] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3438–3446, 2016.

[77] V. S. Amal, S. Suresh, and G. Deepa, "Real-time emotion recognition from facial expressions using convolutional neural network with fer2013 dataset," in *Ubiquitous Intelligent Systems* (P. Karuppusamy, I. Perikos, and F. P. García Márquez, eds.), vol. 243 of *Smart Innovation, Systems and Technologies*, Singapore: Springer, 2022.

[78] A. A. H. Qutub and Y. Atay, "Deep learning approaches for classification of emotion recognition based on facial expressions," *Nexo Revista Científica*, 2023.

[79] Y. El Boudouri and A. Bohi, "Emonext: an adapted convnext for facial emotion recognition," in *2025 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2025.

[80] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, C. Y. S. Xie, S., and C. C. Chea, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

[81] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NeurIPS: Conference on Neural Information Processing Systems)*, vol. 28, 2015.

[82] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," 2024.

[83] Xie *et al.*, "Adaptive global-regional alignment for cross-domain FER," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023.

[84] Y. E. Boudouri, W. Amine Bohi, M. J. Rosato, and C. C. Chen, "Emonext: Spatial-transformer-aided convnext for facial emotion recognition," 2025.

[85] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, (Nara, Japan), pp. 200–205, IEEE, 1998.

[86] Y. He *et al.*, "Residual feature-reutilization inception network," *Signal Processing*, vol. 219, pp. 109–134, 2024.

[87] R. Müller, S. Kornblith, and C. C. C. Geoffrey E. Hinton, "When does label smoothing help?," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] R. Wang, S. Malladi, T. Wang, K. Lyu, and Z. Li, "The marginal value of momentum for small learning rate sgd," *arXiv preprint arXiv:2307.15196*, 2023.

[89] X. Zhou, Z. You, W. Sun, D. Zhao, and S. Yan, "Fractional order stochastic gradient descent method with momentum and energy for deep neural networks," *SSRN*, 2024.

[90] M. Andriushchenko, F. D'Angelo, A. V. Varre, and N. Flammarion, "Why do we need weight decay in modern deep learning?," *ArXiv*, vol. abs/2310.04415, 2023.

[91] L. Pan and C. C. C. Xinyuan Cao, "Towards understanding neural collapse: The effects of batch normalization and weight decay," *ArXiv*, vol. abs/2309.04644, 2023.

[92] I. Zaznov, A. Badii, J. Kunkel, *et al.*, "Adamz: an enhanced optimisation method for neural network training," *Neural Computing and Applications*, vol. 37, pp. 26887–26914, 2025.

[93] X. Zhou, Z. You, W. Sun, D. Zhao, and S. Yan, "Fractional-order stochastic gradient descent method with momentum and energy for deep neural networks," *Neural Networks*, 2025.

[94] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9887–9903, 2022.