STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2025, pp 0–28.
Published online in International Academic Press (www.IAPress.org)

Facial Expression Recognition: A Survey of Techniques, Datasets, and Real-World Challenges

Mohamed A. Abdeldayem [1,*], Hesham F. A. Hamed [1], Amr M. Nagy [2,3]

[1] *Department of Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Egypt*
[2] *Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Egypt*
[3] *Department of Computer Science, Faculty of Computer Science, Benha National University, Egypt*

Abstract Facial expressions are a powerful nonverbal communication tool that can convey emotions, thoughts, and intentions, enhancing the richness and effectiveness of human interaction. Facial Expression Recognition (FER) has gained increasing attention due to its applications in education, healthcare, marketing, and security. In this survey, we examine the key techniques and approaches employed in FER, focusing on three main categories: traditional machine learning, deep learning, and hybrid methods. We review traditional pipelines involving image preprocessing, feature extraction, and classification, along with deep learning methods such as convolutional neural networks (CNNs), transfer learning, attention mechanisms, and optimized loss functions. Furthermore, the study provides a comprehensive examination of existing research and available datasets related to emotion recognition. We also summarize the best-performing methods used with the most common datasets. In addition, the survey addresses the technical challenges of emotion recognition in real-world scenarios, such as variations in illumination, occlusion, and population diversity. The survey highlights state-of-the-art FER models, comparing their accuracy, efficiency, and limitations. Ultimately, this work serves as a comprehensive starting point for researchers, offering insights into current FER trends and guiding the development of more robust and accurate recognition systems.

Keywords Facial expression recognition, Machine learning, Deep learning, Transfer learning, Attention mechanism, Hybrid techniques, Survey.

DOI: 10.19139/soic-2310-5070-2789

1. Introduction

Facial expression recognition (FER) is essential in nonverbal communication and remains fundamental in interpersonal interactions [1], [2]. Additionally, facial expressions and their interpretations may differ between cultures, but a general understanding of some expressions enhances understanding among people from different cultural backgrounds. These expressions can be classified into eight distinct categories: anger, happiness, neutrality, contempt, disgust, fear, sadness, and surprise. The number of research papers on FER is continually increasing, according to the Scopus database, which makes the topic more appealing to researchers, as illustrated in Figure 1.

Vision-based FER has emerged as a formidable tool for emotion assessment in a wide range of practical applications. In [3], for example, counseling psychologists evaluate a patient's psychosocial state and formulate therapeutic strategies by consistently monitoring nonverbal cues, such as facial expressions. In the realm of retail sales, the analysis of client facial expression data is employed to determine the necessity of having a human sales representative present [4], [5]. Other significant application areas include social robotics and facial expression

*Correspondence to: Mohamed A. Abdeldayem (Email: mohamed-abdeldaym@eru.edu.eg). Artificial Intelligence, Faculty of Artificial Intelligence, Egyptian Russian University, Badr, 11829, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

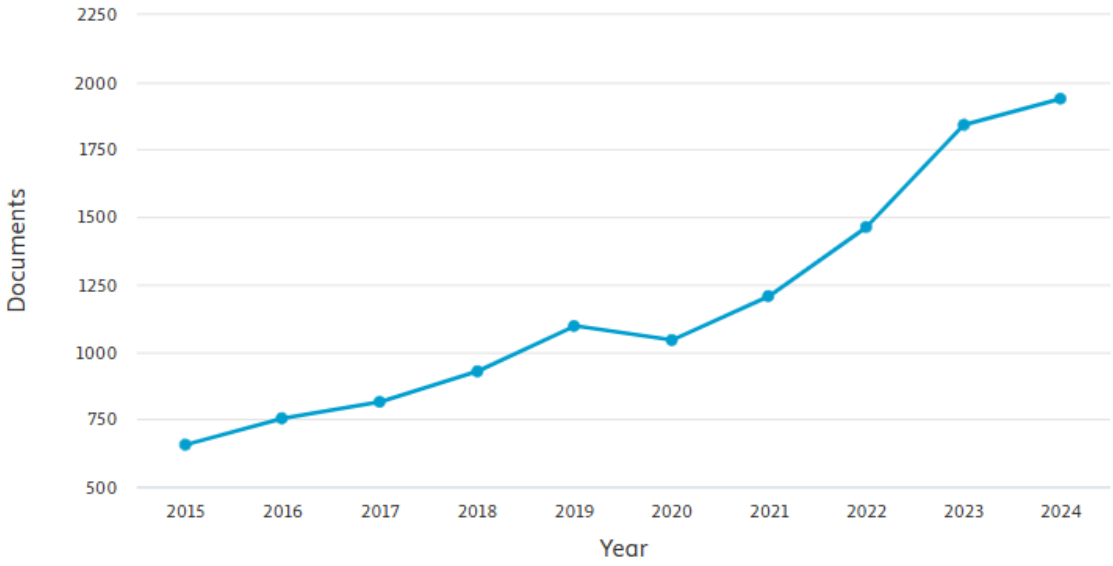

Figure 1. Number of research papers from 2015 to 2024.

synthesis, particularly in human interaction systems, such as feedback for improved e-learning and driver fatigue monitoring.

The advancement of computer technology has prompted improvements in various domains, including artificial intelligence and machine learning. Consequently, various techniques have been employed for the identification of facial expressions. These techniques utilize a substantial amount of valuable information during communication to train for effective facial expression recognition (FER) [6]. The two fundamental phases of classical FER include emotion identification and feature extraction. Moreover, traditional machine learning FER methods typically use human-defined features, employing local binary pattern (LBP) to extract features from images [7], and Histogram of Oriented Gradient (HOG). In [8], HOG was applied to the JAFFE dataset, achieving an accuracy of 92.97%. In [9], the authors utilized LBP with different datasets, achieving accuracies of 99.66%, 74.23%, 89.53%, and 88.20% on the CK+, FER2013, FERPLUS, and RAF-DB datasets, respectively.

Recent advancements in the domains of machine learning and deep learning have enabled the development of increasingly sophisticated algorithms for FER, which can achieve notable precision in classifying fundamental facial expressions, particularly in demanding real-world scenarios [10]. Advances in FER have been made possible by advanced neural network architectures, including CNNs, attention mechanisms, and transfer learning, which skillfully manage variations in facial expressions influenced by factors such as lighting, head poses, occlusions, and focus on parts of images when training [11], [12]. In contrast, features can be automatically recognized through deep neural networks which recently dominated most of the research for facial expression [13], [14]. Despite deep learning models showing better performance, intricate deep learning architectures come with substantial hardware demands, thereby constraining their utility on mobile and embedded systems [15], [16]. In [17], authors used CNNs to extract features to interpret and respond to human facial emotions. FER has made significant progress using CNNs to classify emotions. Moreover, the use of transfer learning has greatly enhanced the power of facial emotion recognition models. In [18], transfer learning was used through DenseNet-161 achieve the best accuracy: 96.51% on KDEF and 99.52% on JAFFE,and with Resnet-50 on FER2013 and RAF-DB, achieving an accuracy of 88.13% and 86.72% [3].

This survey aims to explore the challenges of facial image recognition by reviewing previous research and identifying key difficulties in the field. The main contributions of this paper can be summarized as follows:

- **A comprehensive overview of FER techniques:** We categorize and examine the main methodologies utilized in facial expression recognition, such as traditional machine learning methods, deep learning

architectures (including CNNs, attention mechanisms, and transfer learning), and hybrid techniques, with a focus on their strengths and limitations.
- **Analysis of FER Datasets:** We present a comprehensive evaluation of the most frequently utilized FER datasets, highlighting their attributes, challenges (including occlusion, mislabeling, and class imbalance), and their impact on model performance.
- **Identification of key challenges:** We identify important technical and practical challenges faced in real-world FER applications, such as illumination variation, data bias, ethical gaps, expression similarity, intra-class and inter-class variability, and ambiguity in facial images.
- **Survey of cutting-edge models and solutions:** We review recent advances in FER, including loss function enhancements (such as AdaReg, Ad-Corre, and PDLS), attention mechanisms, data augmentation strategies (including MixAugment), and ensemble approaches such as ESR, and explore how these techniques address existing limitations.

The rest of this paper is organized as follows: Section 2 discusses the challenges, Section 3 presents the methodologies, Section 4 reviews the datasets, Section 5 explores the applications, Section 6 provides evaluation and analysis, Section 7 outlines future work, and finally, Section 8 concludes the paper. By evaluating the current state of FER research, we offer recommendations to guide researchers in developing more accurate, efficient, and generalizable FER systems. In summary, Table 5 provides a comparison of the most recent studies addressing various FER solutions.

2. Challenges

2.1. Dataset

Despite significant progress in facial expression recognition (FER), the field continues to face complex challenges arising from variations in facial appearance, similarity, and ambiguity of expressions, compounded by environmental factors and dataset limitations. Datasets collected in the wild often suffer from occlusion issues such as low lighting, headwear, glasses, or hair covering parts of the face which hinder accurate feature extraction and model generalization [19], [20], [21]. Some datasets even contain mislabeled or incomplete samples, such as FER2013, which includes images without faces that degrade model performance [22]. Additional challenges stem from ambiguous labeling, large intra-class variability, and class imbalance, all of which contribute to inconsistent accuracy and model instability [23], [24]. Dataset specific limitations further exacerbate these issues for instance, AffectNet initially contained inconsistent image sizes (later standardized to 224×224) [20], while AFEW suffered from a scarcity of usable video samples, leading to the creation of SFEW [25]. Underlying many of these challenges are representational biases embedded in training datasets. Narrow demographic coverage, such as JAFFE's exclusive focus on ten Japanese women [26], biases models toward specific ethnic and gender groups, while AffectNet, despite its scale, overrepresents Western and young adult faces and underrepresents older, non-binary, and neurodiverse individuals [27]. Such demographic imbalances can cause models to perform 20–30% better on Caucasian faces than on Black or Asian faces, and up to 15% worse on older adults due to age-related facial muscle differences [28], [29]. Environmental and cultural factors, including lighting, occlusion, and expression norms, further reinforce these biases, reflecting historical data collection practices that prioritized convenience over inclusivity and ultimately limiting model fairness and robustness across diverse populations.

2.2. Model

Additionally, the process of identifying facial expressions raises many methodological challenges [30], such as the use of augmentation strategies to overcome overfitting problems on unconstrained datasets [31], and attention mechanisms that pose difficulties in capturing facial images and obtaining global information from them [32], [33], [12]. Performance deteriorates due to the extraction of low-level features [34]. Furthermore, learning from video frames is difficult due to frame instability [35]. Traditional methods, such as ensemble learning, require more computational resources and involve replication [23]. A new data augmentation strategy called MixAugment

has been proposed to overcome these challenges. It relies on the Mixup technique to augment data, enhancing data diversity through constrained transformations applied to the original dataset [36]. The attention mechanism, instead of focusing solely on low-level features, integrates two stages: the initial stage of low-level feature extraction and a later stage of high-level semantic representation [34]. To distinguish the different effects of facial expressions from the lower and upper face regions, a combination of channel attention and spatial self-attention mechanisms is used [37]. Ensemble Shared Representations (ESR) have been introduced to address challenges in ensemble learning. ESR encompasses various networks dedicated to the learning process, focusing on the execution of low- and mid-level discriminative learning within convolutional layers [23]. FER remains a promising area for researchers, as many issues and challenges still exist regarding both the datasets used and the methods applied.

2.3. Similarity and class imbalance

The most common issues facing existing FER models are high expression similarity and class imbalance. Previous studies have primarily focused on increasing discriminative ability through deep CNNs. However, these architectures require significant memory resources and incur high computational costs. Using deep neural networks remains challenging due to intra-class variability, inter-class similarity in facial images, and the ambiguity of certain images. Discriminative power and class imbalance handling have been improved by introducing an adaptive supervised objective known as AdaReg Loss [38]. Additionally, the discriminative capability of embedded feature vectors has been enhanced through the application of deep metric learning methods. These challenges can further be addressed by Adaptive Correlation Loss (Ad-Corre), which improves FER performance under difficult conditions [39].

2.4. Ethical Gaps

Beyond technical challenges, FER deployment raises serious ethical concerns, especially regarding privacy and potential misuse. The survey addresses these issues only briefly, leaving critical discussions underexplored. Facial recognition technologies, including FER, collect highly sensitive biometric data, such as unique emotional signatures. This information can be exploited without consent, particularly in surveillance contexts. Such non-consensual monitoring undermines individual autonomy and raises serious ethical concerns. In hiring scenarios, emotion-sensing tools pose a substantial risk of discriminatory outcomes. Algorithms may misinterpret “neutral” expressions as indicators of disinterest, thereby disadvantaging candidates from cultures where subdued emotional displays are normative. Such misinterpretations can reinforce occupational segregation along racial and gender lines, perpetuating systemic inequities in employment practices. Privacy invasions are further intensified in workplace emotion tracking, where employees often report fears of constant surveillance. Such monitoring can lead to psychological stress associated with “performing” emotions and increase the risk of biased inferences, such as misinterpreting anxiety as incompetence [40]. Moreover, unreliable FER performance in non-ideal conditions, such as poor video quality, increases the likelihood of false positives in high-stakes security applications. This raises the risk of wrongful profiling and other serious consequences. Ethical frameworks must therefore prioritize consent, transparency, and systematic bias audits. However, current practices frequently fall short of these standards, as evidenced by the proliferation of unregulated commercial tools [41].

3. Methods

Facial Expression Recognition techniques can be broadly classified into three categories: Machine Learning (ML), Deep Learning (DL), and Hybrid approaches. Each method follows a different processing pipeline, as illustrated in Figure 2.

3.1. Machine Learning Techniques

Early FER approaches involve several steps, including preprocessing, feature extraction, feature selection, and classification using machine learning techniques [42]. Preprocessing is one of the primary processes for handling

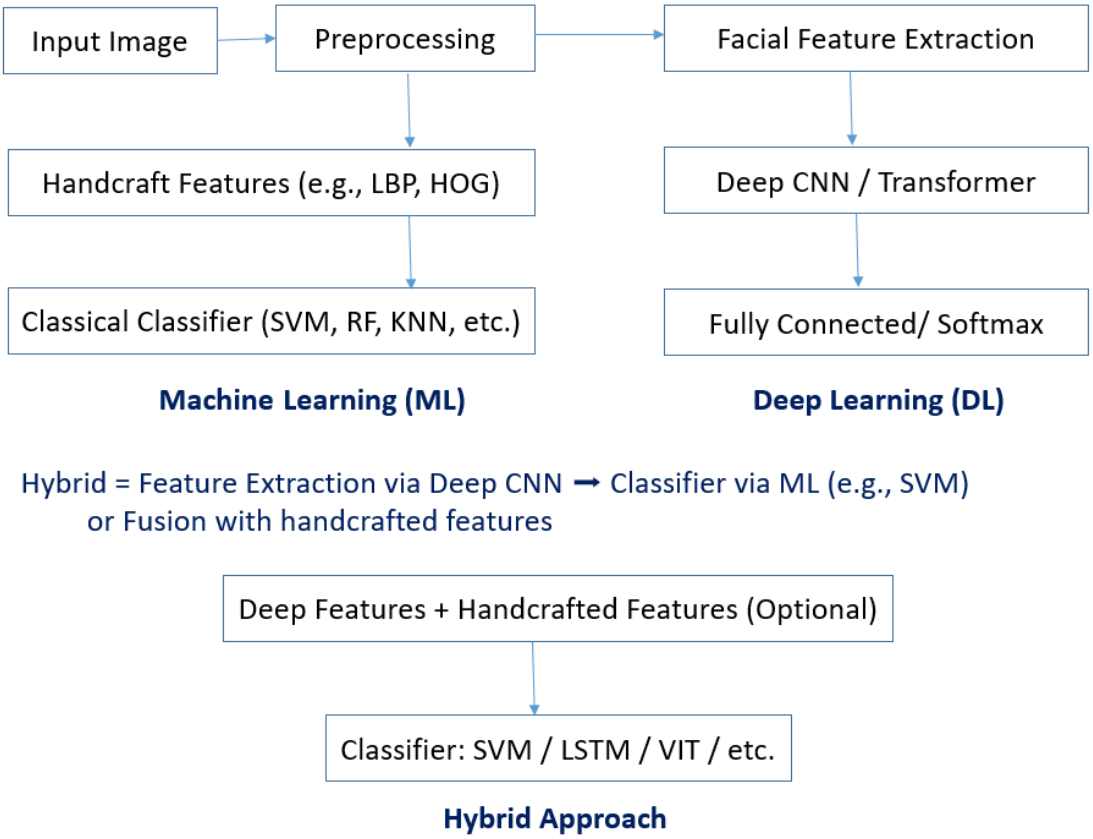

Figure 2. Pipeline Diagram: ML vs. DL vs. Hybrid Techniques for FER

images and improving image quality. These operations include reducing noise [43], converting images to binary or grayscale, adjusting pixel brightness [44], resizing images [45], reducing the effect of illumination using histogram equalization and image sharpening [7], and managing the background, which affects classification accuracy [46]. Feature extraction is the most critical stage in object recognition within an image. In [7], LBP was used to extract features from images, achieving accuracies of 98.61% and 97.62% for the Cohn-Kanade and JAFFE databases, respectively. In [47], a combination of PCA and LBP was used to select the best features, achieving an accuracy of up to 93.75%. In [48], a combination of HOG and LBP achieved accuracies of 92.22%, 97.70%, 96.02%, and 91.30% on the MMI, JAFFE, LNMIIT, and Cohn-Kanade datasets, respectively. Classical machine learning algorithms involve identifying handcrafted features to represent facial expressions [49]. In [50], facial expressions were classified using active learning algorithms and SVM, achieving an accuracy of 92.26% on the Cohn-Kanade dataset. Other papers have used SVM after feature extraction, such as [51], KNN [52], [48], random forest [8], and decision tree [53]. In [54], SVM was applied to the MMI and JAFFE datasets, achieving accuracies of 99.02% and 98.44%, respectively. Additionally, in [55], an ensemble learning strategy was used to combine the outputs of SVM, random forest, and logistic regression to improve accuracy through majority voting, enhancing stability and prediction performance.

After revising the section, we have expanded the discussion to include an analytical comparison with traditional machine learning techniques. Specifically, we highlight methods that rely on handcrafted features such as LBP [7, 47] and HOG [48]. These methods achieved high accuracy on classical datasets like Cohn-Kanade, JAFFE, and LNMIIT, often ranging between 90–99%, depending on the specific feature design and classifier used. HOG generally outperforms LBP alone, as its gradient-based descriptors capture local shape and edge information more effectively under varying illumination conditions [48], [42]. However, these improvements come with notable

computational trade-offs. Traditional approaches require extensive preprocessing—such as color conversion, illumination correction, and background normalization along with feature selection to address dimensionality issues. Classifiers like SVM, KNN, and Random Forest also experience increased time and memory complexity with high-dimensional feature spaces. Overall, while traditional techniques perform well on controlled datasets, they lack robustness when faced with real-world variations such as occlusion, lighting, and pose. This limitation highlights the motivation for adopting more advanced deep learning-based FER models [49].

3.2. Deep Learning Techniques

Deep learning has recently made significant strides, enabling it to deliver highly accurate results in FER. It has revolutionized FER by providing powerful tools to automatically learn features from large datasets. CNNs have become the cornerstone of modern FER systems due to their ability to automatically learn hierarchical features from raw image data. Modern techniques commonly involve fine-tuning pre-trained models that were initially trained on extensive datasets. Transfer learning, where a model trained on one task is adapted for another, is a common technique in this field.

In this part, we will present the most commonly used methods in deep learning applications to recognize facial expressions. Figure 3 illustrates the architectural evolution of facial expression recognition (FER) models, highlighting the progressive shift from traditional convolutional networks to attention-based and transformer-driven frameworks. Early CNN architectures, such as FERNet [45] and DCNN [56], laid the foundation for automated facial feature extraction. However, their reliance on local receptive fields limited their capacity to capture broader contextual information. The introduction of transfer learning enabled deeper architectures such as ResNet [17] and MobileNet [1] to leverage large-scale pretraining, significantly enhancing their ability to generalize across diverse FER datasets. The subsequent integration of attention mechanisms, such as CBAM [32], further improved model performance by directing computational focus toward salient facial regions, thereby increasing robustness against variations in occlusion, pose, and illumination. Most recently, transformer-based architectures, such as ViT [34] and POSTER++ [14] have redefined feature representation through global self-attention, enabling comprehensive modeling of spatial dependencies and emotional context. This shift marks a conceptual transition from localized convolutional learning to globally contextualized perception in FER systems.

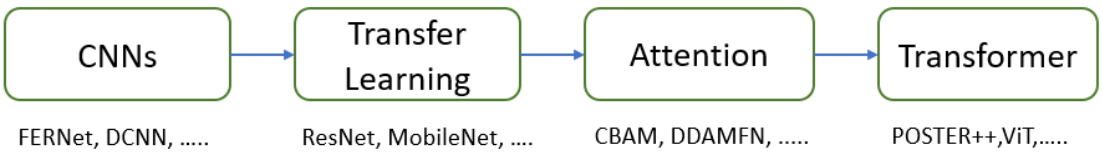

Figure 3. The progression of FER architectures from traditional CNNs to advanced transformer-based models.

3.2.1. Convolutional Neural Network: CNNs are among the most powerful algorithms for extracting image features [11], [6]. Many studies in this field have used CNNs to analyze FER processes from numerous images in order to learn facial expression features [56], [57], [31], [58], [59]. In [60], the authors proposed a FER approach for masked faces, combining low-light image enhancement with upper-face feature analysis using a CNN. The model achieved an accuracy of 69.3% on the AffectNet dataset for 8 facial expression classes. In [61], a model was built using a CNN that includes two main components: EfficientFace, used for recognizing facial expressions, and the Label Distribution Generator (LDG), which creates a label distribution that functions as the ground truth for training EfficientFace. In this context, a deep DWConv convolution, global average pooling, and a fully connected network were combined to build a model capable of classifying facial expressions. In [45], a model called FERNet was developed using a deep convolutional neural network (DCNN), which included multiple blocks with convolutional layers and sublayers. The model achieved an accuracy of approximately 69.57% on the FER2013 dataset. Convolutional networks have also been used to generate Ensembles with Shared Representations (ESRs), which have been shown to be scalable and efficient for processing large-scale facial expression datasets. The ESRs

were trained using subject-independent ten-fold cross-validation on the CK dataset [23]. The Facial Motion Prior Networks (FMPN) framework was proposed to expedite facial expression recognition. This framework includes two components: a Facial Motion Generator (FMG) and a Classification Network (CN). The FMG is responsible for generating the facial motion mask by integrating prior domain knowledge, while the CN is a deep learning-based classification network trained using the cross-entropy loss function [16].

3.2.2. Transfer Learning: Pre-trained models are employed to address novel, closely related problems that exhibit limited data availability and low complexity through transfer learning [10], [62]. In [18], transfer learning based on a deep CNN model was used to recognize facial expressions and extract optimal features. By optimizing pre-existing models with data specific to a particular domain, researchers have successfully achieved superior performance levels by leveraging the intricate feature representations learned by these models from extensive datasets [63]. Numerous studies have trained self-constructed networks from scratch or improved performance by fine-tuning established models such as MobileNetV1 [1], [22], ResNet-18 [17], [54], GoogleNet [64], and VGG-Face [65], and have utilized various deep learning frameworks, including Inception-V3 [6], Inception-ResNet-v2 [66], VGG19, and ResNet-50 [67], [68], [3], to detect interaction in e-learning environments by recognizing facial expressions [67]. In [69], a single FER neural network was used, which was pre-trained for facial recognition and fine-tuned on AffectNet static images using a powerful optimization technique. The model achieved an accuracy of 66.34% for 7 classes and 63.03% for 8 classes. For the other two datasets, VGAF and AFEW, the model achieved accuracies of 70.23% and 59.27%, respectively. In summary, the use of transfer learning for facial expression recognition has played an effective role in achieving state-of-the-art results across different datasets.

3.2.3. Attention Mechanism: Attention mechanisms work deeply within images to enhance feature extraction and have proven to be powerful tools for processing facial expressions. Convolutional Block Attention Mechanisms (CBAM) are among the most successful and effective techniques in FER, as they selectively focus on important regions within facial images to improve recognition accuracy [32], [70]. In [71], an attention mechanism was proposed to operate across pyramid levels to identify the most critical facial regions, achieving accuracies of 63.77% on AffectNet-7, 60.68% on AffectNet-8, 88.98% on RAF-DB, and 89.75% on FER+. This mechanism also aids in selecting the most informative metrics to obtain discriminative expression representations. Similarly, [13], [68] employed different non-overlapping regional attention mechanisms to extract features from distant facial regions. Additionally, a convolutional attention module was combined with residual network optimization to improve FER accuracy [72]. In this context, attention weights associated with the importance of features were calculated using intermediate spatial feature maps within a CNN, serving as a contextual framework. The integration of attention-based hybrid models has effectively addressed the limitations of traditional convolutional filters, such as limited inductive bias and receptive field size [73].

In [33], a Region Attention Network (RAN) was developed to emphasize the importance of facial regions, especially under occlusion and expression variation. This model performed well on datasets such as FERPlus, AffectNet, RAF-DB, and SFEW. In [74], the Deep Attentive Center Loss (DACL) method was introduced to enhance feature discrimination by adaptively selecting the most relevant features using spatial feature maps within the CNN. Another study, [75], proposed the Dual-Direction Attention Mixed Feature Network (DDAMFN), which utilizes convolution kernels of varying sizes to extract diverse spatial information and integrates coordinate attention layers to model long-range dependencies. DDAMFN generates feature maps in vertical and horizontal directions using multiple Dual-Direction Attention (DDA) heads and reshapes them to predict expression categories via a fully connected layer.

Attention mechanisms are frequently built on convolutional neural networks and often support both low-level feature learning and high-level semantic representation. In [34], two types of attention mechanisms were introduced: grid-wise attention, designed to capture interdependencies in local features, and a visual transformer-based attention method for global semantic understanding using visual semantic tokens. In [14], POSTER++ improved upon its predecessor by introducing cross-fusion, a two-stream architecture, and multi-scale feature extraction. These enhancements replaced the original window-based mutual attention with conventional

mutual attention, removed the image-to-landmark branch, and fused image features with landmark-based multi-scale features, achieving 67.49%, 63.77%, and 67.49% accuracy on AffectNet-7, AffectNet-8, and RAF-DB, respectively. Lightweight attention-based FER models have also emerged. For instance, a MobileNet-based model in [15] incorporated an attention module with dropout to mitigate overfitting and was evaluated on FERPlus and RAF-DB using MobileNetV1, V2, and V3. In [1], PAtt-Lite, a lightweight patch-based attention network based on MobileNetV1, used a patch extraction block to enhance local feature representation. Recently, attention mechanisms have continued to play a critical role in FER, especially when combined with CNNs and transfer learning strategies.

3.2.4. Loss Function: A loss function within the realm of deep learning serves as a measure of the performance of a neural network model on a given dataset. It quantifies the disparity between the predicted outcomes and the actual outputs generated by the model. The primary objective of neural networks is to minimize the loss function, thereby enhancing the model's predictive accuracy. In [76], a novel adaptive supervised objective known as Adaptive Regularization (AdaReg) Loss is presented, which recalibrates class significance coefficients to mitigate class imbalance and enhance the robustness of expression representations. The standard AdaReg is employed to optimize disparities among facial expressions within a high-dimensional space during CNN training. In [38], a standard term is formulated according to the expression presented in Equation 1; however, it is noted that this standard is susceptible to imbalanced label distributions.

$$L_{\text{AdaReg}} = \frac{1}{\lambda \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} \alpha_i ||e_i - e_j||_2} \tag{1}$$

Where λ is a parameter used to weight the importance coefficients; e_i and e_j represent the embeddings of categories i and j, respectively, N is the total number of categories; and α_i is the importance coefficient for category i

In [39], another loss function, which constitutes the adaptive correlation loss (Ad-Corre) for FER in the wild, consists of feature discrimination, average discrimination, and embedded discrimination. Deep metric learning (DML) is applied in the paper's Xception network architecture to enhance the discriminative capabilities of the learned embedded features. The network is guided by the Ad-Corre loss to produce embedded feature vectors with low between-class sample correlation and high within-class sample correlation. The model is then trained by combining the proposed Ad-Corre loss with the cross-entropy loss.

$$\text{COR}(X,Y) = \frac{\sum_{k=1}^{d}(X_k - \bar{x})(Y_k - \bar{y})}{\sqrt{\sum_{k=1}^{d}(X_k - \bar{x})^2} \times \sqrt{\sum_{k=1}^{d}(Y_k - \bar{y})^2}} \tag{2}$$

According to equation 2, the correlation between two vectors X and Y is equal to 1 when they are identical and -1 when they are uncorrelated. Here $\bar{x}$ and $\bar{y}$ represent the mean values of vectors X and Y, respectively.

In-the-wild FER involves challenges related to pose, orientation, and input accuracy. To enhance low-resolution inputs, the authors in [20] proposed a Pyramid with Super-Resolution (PSR) network. They also addressed expression ambiguity by using the Prior Distribution Label Smoothing (PDLS) loss function, as specified in equation 3. The PDLS loss function incorporates the initial study's understanding of expression confusion in facial expression recognition (FER) tasks by examining specific confusion patterns between different emotion labels. This function is used with the VGG16 backbone network, employing input images at different scales from the AffectNet-7, AffectNet-8, FER+, and RAF-DB datasets, achieving accuracies of 63.77%, 60.68%, 89.75%, and 88.98%, respectively.

$$\text{PDLS} = -\sum_{c \in C} ((t_c \cdot \alpha + d_{kc} \cdot (1 - \alpha)) \cdot \log(\sigma(z_c))) \tag{3}$$

Where c represents each class in the FER task; t_c is the true label for class c; α is the weight parameter for the true label; d_{kc} provides prior information regarding the ambiguity of each expression in the FER task; and $\sigma(zc)$ represents the softmax output for class c.

In [77], a new loss function was introduced, known as the weighted clustering loss function, as shown in equation 4, specifically designed to address imbalanced datasets. This loss function was implemented during the fine-tuning phase by establishing a class center for each expression category and assigning weights based on the proportional representation of each expression in the overall image count. This approach enhanced intra-class compactness and inter-class separability [78].

$$L_{\text{loss}} = \frac{1}{2}\sum_{i=1}^{n} W_{y_i} \frac{||x_i - c_{y_i}||_2^2}{\left(\sum_{\substack{j=1 \\ j\neq i}}^{k} ||c_j - c_{y_i}||_2^2\right) + \gamma} \tag{4}$$

Where n is the number of samples in the dataset; W_{y_i} is the weight associated with the true class label y_i; $C_{y_i}$is the centroid of the true class y_i; k is the total number of classes; λ is the regularization parameter; and x_i represents the input features of the i^{th} sample.

3.2.5. Efficient and Lightweight Architectures for Edge Devices: Deep CNNs and attention-based models frequently deliver high accuracy. However, their use in mobile and embedded systems is limited by significant computational and memory demands. FER applications in mobile learning apps, driver monitoring, and social robots demand lightweight, efficient models. The remainder of this subsection summarizes lightweight architectures and practical model optimization techniques that are particularly relevant to FER.

Lightweight architectures.

- **MobileNet (V1/V2/V3):** Depthwise separable convolutions help reduce the parameter count, making them suitable for mobile FER applications. Using this approach, MobileNet achieved 97.9% accuracy on the FER2013 dataset [1].
- **A-MobileNet:** Incorporates attention to improve discriminability while keeping a compact design, validated on RAF-DB and FERPlus [15].
- **Lightweight hybrids for attention:** PAtt-Lite incorporates patch- or symbol-based attention modules into MobileNet-like backbones. This design has been shown to enhance feature localization while introducing only minor increases in the number of parameters [22].
- **Specialized tiny models**: Task-specific tiny CNNs are designed from scratch with fewer layers and narrower channels. They are particularly suited for strict-latency applications such as driver monitoring and real-time classroom feedback [79].

Model optimization techniques. Beyond selecting a compact architecture, three main families of compression and optimization methods are commonly applied, particularly when targeting edge devices:

- **Pruning.** Remove redundant weights/filters (unstructured or structured). Structured pruning (filter/channel removal) generally yields speedups on real hardware unstructured pruning reduces the overall parameter count. However, it often requires sparse kernels or specialized runtimes to achieve noticeable latency gains. Pruning is typically followed by fine-tuning to recover accuracy. In [80], pruning optimization was applied to VGG and ResNet models on the CK+ and JAFFE datasets. The results demonstrated reduced redundancy while preserving overall performance.
- **Quantization.** Reducing numeric precision, for example through 8-bit integer quantization, is a common technique to lower memory requirements. It also accelerates inference on resource-constrained devices. In the context of FER, demonstrates that quantization can effectively reduce memory usage and increase inference speed without significant accuracy degradation [81]. Complementary work on MobileNet architectures further supports these findings: A Quantization-Friendly Separable Convolution for MobileNets proposes structural modifications to enhance robustness under quantization. It shows that 8-bit inference achieves accuracy comparable to full-precision models [82]. Similarly, Memory-Driven Mixed Low Precision Quantization investigates INT4/INT8 quantization on MobileNets deployed to microcontrollers. It reports

favorable trade-offs between latency and accuracy [83]. Together, these studies highlight quantization as an effective strategy for enabling efficient FER on mobile and embedded platforms.

- **Knowledge Distillation (KD).** The authors in [84] use KD to create a model of less than 1 MB for FER, achieving 1851 FPS on a CPU. In [85], an online KD approach is proposed to improve FER accuracy without introducing additional parameters. However, [86] demonstrates KD-based lightweight models for FER on FER2013 and RAF-DB, showing both speed and memory gains.

Recent research in facial expression recognition (FER) highlights not only advancements in architectural innovation but also the growing emphasis on computational efficiency for deployment in resource constrained environments. Traditional CNNs, while lightweight, often compromise representational capacity [45], whereas attention based and transformer driven architectures deliver superior accuracy but demand substantially greater computational resources, including higher FLOPs, memory usage, and inference time [34], [14]. Hybrid architectures that combine efficient backbones such as MobileNet with lightweight attention modules have emerged as a promising compromise, achieving a balanced trade-off between performance and computational cost [1], [15], [22]. Consistent performance patterns have been observed across prior studies: attention mechanisms such as HAM [19] and POSTER++ [14] enhance conventional CNNs, improving robustness under challenging conditions such as occlusion and pose variation, and yielding 2–5% accuracy gains on datasets like RAF-DB and AffectNet. However, these enhancements often increase computational complexity by 20–40%, limiting their practicality for real-time or edge based deployment. Lightweight models such as MobileNetV1 [22] achieve a favorable balance between accuracy and efficiency—reaching 97.9% on FER2013 and 95.05% on RAF-DB though their generalization tends to decline on larger datasets such as AffectNet. To maintain competitive performance, fine-tuning, attention integration, and advanced loss functions are often employed. In particular, Ad-Corre [39] and DACL [74] enhance feature separability and robustness against noisy or ambiguous labels by mitigating intra-class similarity and inter-class overlap. Similarly, multi-scale and region-aware networks such as MM-Net [87] and FMPN [16] demonstrate greater stability under occlusion and pose variation, emphasizing the importance of hierarchical and localized feature learning in real-world applications. Overall, while attention and transformer-based architectures continue to achieve state-of-the-art performance in FER [14], [34], their computational demands constrain deployment in low-power or real-time scenarios. Consequently, lightweight CNN architectures remain highly attractive for efficiency-driven FER systems, where scalability must optimize not only accuracy but also energy consumption and real-time responsiveness [82], [83].

3.3. Hybrid Techniques

FER is a complex task that often requires balancing feature extraction, representation learning, and classification. Hybrid techniques, which combine multiple methodologies, have emerged as powerful tools to address these challenges by integrating traditional machine learning with deep learning methods to improve recognition accuracy and robustness. These techniques leverage the advantages of different strengths in computational paradigms to enhance accuracy, robustness, and generalizability. In [54], a pre-trained ResNet18 model with a triple loss function (TLF) was used to extract features, which were then input into an SVM classification model on the MMI dataset, achieving 99.02% accuracy, and 98.44% accuracy on the JAFFE dataset. Similarly, [88] developed a model that integrates a pre-trained EfficientNetB0 network with additional CNN layers, achieving a top accuracy of 74.39% on the FER2013 dataset and outperforming several state-of-the-art approaches. One popular approach combines CNNs and SVMs: CNNs are used for feature extraction, capturing subtle patterns in facial expressions, while SVMs serve as classifiers, distinguishing between different emotional states.

Additionally, hybrid FER models have included Vision Transformers (ViTs). Using a dual-stream structure to collect both local and global facial features, a study presented a hybrid local attention module in a ViT framework, which improved performance on the RAF-DB, FERPlus, and AffectNet datasets [89]. In [90], state-of-the-art results on the FER2013, FER+, and AffectNet datasets were achieved by combining CNN-derived features with features extracted using the Bag-of-Visual-Words (BOVW) model within a local learning framework, along with a local SVM for classification. In multimedia applications, CNNs combined with Deep Belief Networks (DBNs) were proposed by [91], achieving high recognition rates of 98.14%, 95.29%, and 98.86% on the JAFFE, KDEF, and RaFD datasets, respectively, demonstrating the model's strong generalization capability. In [92], a CNN model with

long short-term memory (LSTM) was built, where features were extracted by a bidirectional LSTM for the CK+ and an in-house dataset. A deep learning approach based on Spatial Transformer Networks (STNs) was presented in [93] for handling issues with high variability datasets like FER-2013. By focusing on important facial areas, this model increased the accuracy of feature extraction and classification.

In [94], a hybrid model was created that combined CNN-based facial motion flow classification with SVM-based geometric displacement classification, achieving 99.69% accuracy on the CK+ dataset and 94.69% on the BU4D dataset. In [7], [48], SVM and KNN were used for classification to recognize facial expressions using two features: HOG and LBP, on several image databases. In [95], a model was proposed that combines DCNNs for feature extraction and Haar Cascades for real-time facial detection, resulting in enhanced classification performance on the FER-2013 dataset. Hybrid techniques are not widely used in FER and may be somewhat limited, perhaps due to the large number of features that have a high similarity rate, making other algorithms that focus on details and parts of the images more important, such as attention mechanisms [13], [74], DeepCNN [60], [45], [96], and transfer learning [68], [22], [67], followed by artificial neural networks. Table 1 shows a typical heat map of accuracy values, simplified for datasets with different methods.

Table 1. Accuracy Comparison Across Methods and Datasets

Dataset	**ML (HOG+SVM)**	**DL (MobileNetV1)**	**DL (POSTER++)**	**Hybrid (CNN+SVM)**
FER2013	90.16%	**97.9%**	—	75.42%
AffectNet-7	—	66.97% (HAM)	**67.49%**	65.07% (ViT+Hybrid)
AffectNet-8	—	**69.3%** (DeepCNN)	63.77%	62.78%
RAF-DB	91.30%	**95.05%**	92.21%	90.45%
JAFFE	**99.66%**	—	—	98.44%
CK+	98.61%	**100%** (MobileNetV1)	—	99.04%

Hybrid techniques in facial expression recognition (FER) leverage features extracted from pre-trained deep models such as ResNet18 and EfficientNet to enhance recognition accuracy and robustness. These features are subsequently used to train advanced classifiers such as SVM or incorporated into fusion architectures, including CNN combined with Bag-of-Visual-Words (BOVW) frameworks [54], [88], [90]. Recent advancements also integrate local and global attention mechanisms within Vision Transformer (ViT)-based frameworks to improve spatial representation and recognition stability under challenging conditions [89]. Empirical results demonstrate the strong potential of such hybrid systems: for instance, ResNet18 coupled with SVM achieved accuracies of 99.02% and 98.44% on the MMI and JAFFE datasets, respectively [54], while CNN–SVM models combining geometric motion and appearance-based features achieved 94–99% on CK+ [94]. Comparable improvements have been reported on large-scale datasets such as FER2013 and RAF-DB through CNN–attention fusion strategies and the inclusion of additional network layers [88]. The primary advantage of hybrid techniques lies in their enhanced robustness and generalization capability. By integrating complementary feature extraction and classification paradigms, these models become less sensitive to variations in lighting, pose, occlusion, and background complexity, thereby achieving more stable performance in real-world scenarios. However, this improvement often comes at the cost of increased computational complexity and higher resource requirements during both training and inference, especially when multiple components such as CNNs, Transformers, local attention modules, and fusion units are combined. Moreover, handling large feature sets may introduce redundancy or high inter-feature correlation, potentially diminishing performance if not properly optimized. Analytically, hybrid models provide a balanced trade-off between the localized precision of CNN-based architectures and the global contextual understanding enabled by Transformer-based designs. While Transformer only FER models achieve strong performance with end-to-end learning, they typically incur substantial computational overhead. Conversely, pure CNNs are computationally efficient but often less capable of handling complex spatial dependencies. Hybrid architectures bridge this gap, combining the efficiency of CNNs with the representational strength of attention mechanisms. This balance makes them particularly suitable for semi-constrained real-world FER applications, where both accuracy and computational efficiency are critical considerations [13], [74], [68], [22], [67].

4. Datasets

In recent years, the number of datasets dedicated to facial expression recognition has significantly increased, driven by the growing demand for emotion-aware applications. This section presents a comprehensive overview of the most widely used datasets, highlighting their characteristics, structure, and relevance in FER research.

4.1. Large-Scale Image Datasets

AffectNet
It contains approximately 400,000 images that have been manually annotated for the purpose of recognizing eight specific facial expressions. Along with the corresponding valence and arousal levels, these expressions include neutral, surprise, disgust, happy, angry, sad, fear, and contempt. This dataset is categorized into two main branches: AffectNet-7 and AffectNet-8, the latter incorporating the "contempt" category. AffectNet-7 encompasses a total of 287,401 images distributed across seven classes, with 283,901 images assigned to the training set and 3,500 images to the test set. Furthermore, AffectNet-8 includes contempt data, resulting in an increase in the number of training and testing samples to 287,568 and 4,000 images, respectively [27].
RAF-DB
(Real-world Affective Faces Database) includes 29,672 images with seven basic facial expressions. The images depict people of different ages, nationalities, and head positions, including occlusions such as head tilt, glasses, or hair, and processed images such as filters and special effects [97].
FER2013
It contains over 30,000 RGB facial photos of various expressions, each limited to 48 by 48 pixels. The primary labels in FER2013 fall into seven categories: sad, disgust, surprise, fear, happiness, and neutral. The disgust expression has the fewest images (600), compared to nearly 5,000 examples for each of the other labels [98].
FER Plus
It includes eight expressions derived from the original FER dataset. It exhibits class imbalance, with 9,030 neutral images in the training set and 1,102 in the test set. Distaste has the fewest images, with just 107 in the training set and 15 in the test set. The contempt emotion contains a comparable number of images—115 in the training set and 13 in the test set. Compared to the other five emotions, disgust, contempt, and fear have fewer visual representations. This is typical in natural conversation, where people are generally happy or in a neutral state, and rarely express contempt, disgust, or fear [99].
FERG
It's a repository comprising stylized characters with annotated facial expressions. It includes 55,767 images of six stylized faces, all generated using MAYA software. The images for each character are categorized into seven distinct types of expressions [100].
JAFFE
It focuses on Japanese females and includes seven facial expressions. It contains 213 images of these expressions from 10 different Japanese female subjects. These images were annotated with average semantic ratings for each facial expression by 60 annotators [26].
KDEF
created under laboratory conditions, was originally designed for use in psychological and medical research. It comprises images from 70 actors captured at five different angles and is annotated with six fundamental facial expressions, along with a neutral expression. In the field of basic emotion recognition, various extensively used and large-scale facial expression datasets obtained from online sources have emerged in recent years to support the training of deep neural networks [101].
CAER-S
It contains 65,983 images, categorized into seven facial expressions. These images are divided into 68% for training and 32% for testing. The dataset was collected from social media to support the study and development of facial expression recognition systems [102].

4.2. Video and Dynamic Datasets

Cohn-Kanade (CK+)
It contains 593 videos of facial expressions from 123 people, aged 18 to 50, both male and female, with seven expressions. These were captured at 30 frames per second, with frame dimensions of 640 × 490 or 640 × 480 pixels. This dataset includes 327 videos, each classified into one of the seven facial expressions. The CK+ database is widely used for facial expression classification and is considered one of the most widespread and accessible [103].

BU-3DFE
The database from Binghamton University comprises 606 facial expression sequences obtained from 100 individuals. It includes six basic facial expressions (disgust, fear, happiness, sadness, surprise, and anger) displayed by each participant, induced in different ways and with varying levels of intensity. Like the Multi-PIE dataset, this collection is commonly applied for the examination of three-dimensional facial expressions from multiple viewpoints [104].

DFEW
It comprises more than 16,000 video clips sourced from a multitude of films. These video segments exhibit a range of complex challenges in real-world scenarios, including severe lighting variations, occlusions, and unpredictable changes in facial orientation [105].

DFER
The dataset was made publicly available in 2022 by a research team affiliated with Fudan University, encompassing 38,935 videos and seven distinct emotional categories, thus establishing its status as the most extensive dataset for dynamic expression recognition currently available. Each video in the dataset is meticulously annotated by 30 professional annotators, ensuring a high level of data quality. The dataset is categorized into four groups based on varying scene environments, which include daily social interactions, weekly presentations, intense interactive displays, and unusual scenarios. Moreover, it can be further broken down into 22 more distinct scenarios. Owing to its substantial sample size, each category contains approximately 10,000 video samples. Typically, 80% of the dataset is used for training, while the remaining 20% is used for validation and the calculation of model evaluation metrics [106].

MAFW
It's a comprehensive, multimodal, complex affective dataset designed for recognizing dynamic facial expressions in natural settings. This dataset comprises 10,045 video-audio segments, each labeled with a compound emotional classification. Additionally, a brief description of the subjects' emotional responses in each segment is provided [107].

SFEW
Static frames were carefully selected from the AFEW database using face point clustering techniques to identify crucial frames. The most common version, SFEW 2.0, is used in the SReco sub-challenge benchmarking dataset in EmotiW 2015. Three separate sets make up the SFEW 2.0 dataset: Test (372 samples), Val (436 samples), and Train (958 samples). Each image in this dataset is assigned one of seven different expressions—neutral, happiness, sadness, surprise, disgust, fear, or anger. The labels for the training and validation sets are publicly available to all participants, whereas the labels for the test set are hidden by the challenge organizers [108].

4.3. Specialized or Multi-condition Datasets

Multi-PIE (Multi Pose, Illumination, Expressions)
It comprises facial images of 337 individuals captured under varying conditions of pose, illumination, and facial expressions. There are 15 distinct perspectives within the pose range, depicting profiles of the face in a side-to-side manner. The alterations in illumination were simulated through the utilization of 19 different flashlights positioned at various points within the room [109].

MMI
It's a database dedicated to the expression of the six fundamental emotions. Among these are common expressions found in the MMI facial expression database, which also contains expressions for every Action Unit (AU)

currently identified, along with a number of action descriptors that include one AU enabled in the Facial Action Coding System (FACS). Recently, recordings of natural expressions have also been added. The database contains more than 2,900 high-resolution videos and still images covering 75 subjects. It indicates whether the AU is in a neutral, onset, peak, or offset phase for each frame and is partially encoded at the frame level. Understanding the facial expressions and audio clips from the scene sequences in the videos provides a complete explanation for the existence of AUs in movies (event coding) [110].

The facial expression database used in ***Oulu-CASIA*** NIRVIS includes six distinct expressions, captured by having participants sit on a chair in a room and fixate their position in front of the camera. The distance between the camera and the subject's face is approximately 60 cm. A group of 80 individuals aged between 23 and 58 years were captured. A large majority of the participants were male (73.8%). Participants were asked to mimic a specific facial expression based on a given example shown through image sequences. Recordings were captured at 25 fps with a resolution of 320 × 240 pixels [111].

Radboud Faces Database (RaFD)

It comprises 1,608 images from 67 subjects in a laboratory-controlled setting, featuring three distinct gaze directions: frontal, leftward, and rightward, each associated with eight expressions linked to every image [112].

Building on the previous discussion of FER challenges, a comprehensive understanding of dataset characteristics is essential for improving the reliability, fairness, and generalizability of facial expression recognition systems. In table 2 provides an expanded overview of widely used FER datasets, several critical aspects affecting their real-world applicability warrant closer examination. A major limitation across many datasets is the lack of standardized quality indicators, including labeling reliability, inter-annotator agreement, and detailed demographic distribution. For instance, FER2013 contains numerous mislabeled samples that distort training dynamics and reported accuracy, while AffectNet and RAF-DB offer limited demographic transparency, making fairness and generalization assessment difficult. Another concern involves dataset licensing and usage restrictions, as many large-scale FER datasets impose unclear or restrictive terms that hinder their use in commercial or privacy-sensitive contexts, with few studies addressing their impact on reproducibility and deployment. Dynamic datasets such as DFER, DFEW, and MAFW introduce additional challenges related to temporal alignment, annotation sparsity, and frame-level consistency, affecting the reliability of sequence-based models like CNN–LSTM and transformers in fine-grained temporal analysis. Moreover, dataset suitability often depends on research objectives, yet this is rarely specified; smaller lab-controlled datasets like JAFFE and CK+ enable benchmarking under consistent conditions but lack ecological validity, whereas large-scale in-the-wild datasets provide greater diversity at the expense of increased noise and bias. Therefore, establishing a structured evaluation framework that includes demographic reporting, annotation quality, licensing transparency, and temporal labeling rigor would help researchers select datasets more strategically, ultimately improving model robustness, fairness, and ethical accountability.

Table 2. Expanded comparison of facial expression recognition datasets including demographic details, Annotation, and recommended use cases.

Dataset	Samples	# Exp.	Demographic	Labeling / Annotation	Recommended Use Case
AffectNet	∼400K images	7/8	Broad age/ethnicity diversity (web-sourced). Exact distribution not reported.	Manually annotated (valence–arousal + categorical).	Large-scale in-the-wild training, loss robustness, label-noise studies.
RAF-DB	29,672 images	7	Mix of ages/nationalities; includes head pose/occlusion. Exact distribution not reported.	Manual annotation (crowd/expert).	Robustness to occlusion and head pose; demographic-diversity experiments.

Dataset	Samples	# Exp.	Demographic	Labeling / Annotation	Recommended Use Case
FER2013	∼30K (48×48)	7	Web-sourced; no demographic breakdown reported.	7-class labels; contains noise and mislabels.	Low-res FER, label noise handling, lightweight models.
FERPlus	32,298 images	8	Same as FER2013; class imbalance noted.	Relabeled with label distributions.	Loss-function studies (label smoothing, PDLS, Ad-Corre).
JAFFE	213 images (10 subjects)	7	Japanese female subjects only.	Semantic ratings averaged over 60 annotators.	Controlled, culture-specific studies.
KDEF	∼4,900 images	7	70 actors, lab setting. No ethnicity breakdown.	Posed expressions in controlled angles.	Model validation under controlled pose/illumination.
CK+	593 video sequences (123 subjects)	7–8	Subjects aged 18–50, mixed genders.	Onset/peak/offset annotated; AU labels available.	Temporal FER, AU studies, onset–offset modeling.
BU-3DFE	606 sequences (100 subjects)	6	100 subjects, multiple intensities.	Lab-captured 3D sequences, posed.	3D / multi-view FER, expression intensity analysis.
MMI	>2,900 videos and stills (75 subjects)	6 (+AUs)	75 subjects, lab setting.	Frame-level AU onset/peak/offset annotations.	AU detection, biomechanics of expression.
SFEW 2.0	958 train / 436 val / 372 test	7	Frames extracted from AFEW movies (in-the-wild).	Clustered frame selection; test labels hidden in EmotiW challenge.	Occlusion / in-the-wild robustness; benchmark testing.
DFEW	∼16K video clips	7	Movie clips, varied conditions. Demographics not reported.	Clip-level annotations; in-the-wild.	Large-scale temporal FER, shot-change robustness.
DFER	38,935 videos	7	4 scene categories (social, presentation, interaction, unusual).	Annotated by 30 professional annotators.	Benchmark for dynamic FER, CNN+LSTM/3D CNN training.
MAFW	10,045 video–audio segments	11	In-the-wild multimodal dataset. Demographics not reported.	Segment-level compound labels + short text descriptions.	Multimodal FER (video+audio), compound emotion recognition.

In recent years, the number of datasets dedicated to FER has increased significantly. While these datasets provide a basis for training and evaluating FER models, each has fundamental limitations that affect generalisation, reliability, and fairness. Table 3 illustrates the problems with the dataset and Table 4 helping researchers select the appropriate data for facial expression recognition.

Table 3. Overview of Common FER Datasets and Their Limitations

Dataset	**Type**	**Limitations**
AffectNet	Large-scale in-the-wild (static)	Severe class imbalance, ambiguous "contempt", noisy annotations
RAF-DB	In-the-wild (static)	Class imbalance, subtle emotion ambiguity
FER2013	In-the-wild (static)	Many mislabeled/non-face samples, very low resolution
FERPlus	In-the-wild (static)	Better labels than FER2013 but still imbalanced
FERG	Synthetic (stylized)	Cartoon-like faces, lacks realism
JAFFE	Lab-controlled (static)	Very small, only Japanese women (demographic bias)
KDEF	Lab-controlled (static)	Posed, limited diversity (Swedish actors)
CAER-S	Social media (static)	Noisy images, variable quality
CK+ (Cohn-Kanade)	Lab-controlled (video)	Posed, lacks real-world variability
BU-3DFE	3D static	Controlled conditions, limited natural variation
DFEW	In-the-wild (video)	Annotation subjectivity, noisy backgrounds
DFER	In-the-wild (video)	Large-scale, but subjective/contextual ambiguity
MAFW	Multimodal (video+audio)	Complex compound emotions, annotation subjectivity
SFEW 2.0	In-the-wild (static, from AFEW)	Small, test labels hidden (challenge format)
Multi-PIE	Lab-controlled (pose/illumination)	Unrealistic for natural scenarios
MMI	Lab-controlled (video+images)	Mostly posed, AU coding but less natural
Oulu-CASIA	Lab-controlled (video, NIR & VIS)	Small, constrained, mostly male participants
RaFD	Lab-controlled (static)	Posed, limited diversity, artificial setup

Table 4. Extended Dataset Selection Guidelines for FER Research

Research Goal	**Recommended Datasets**	**Limitations**
Real-world robustness	AffectNet, RAF-DB, CAER-S	Imbalanced and noisy labels; ambiguous classes; variable image quality
Controlled baseline experiments	CK+, JAFFE, KDEF, RaFD	Limited diversity, posed expressions, demographic bias
Video/temporal dynamics	DFEW, DFER, MAFW, Oulu-CASIA, MMI, AFEW	Annotation subjectivity, variable quality, mostly constrained setups
Quick lightweight testing	FER2013, FERPlus, SFEW 2.0	Mislabeled/noisy samples, low resolution, small test sets
3D expression analysis	BU-3DFE, Multi-PIE	Controlled lab conditions, unrealistic for in-the-wild deployment
Synthetic/augmented testing	FERG	Cartoon-like synthetic faces, lacks realism for transfer learning
Multimodal emotion analysis	MAFW	Complex compound emotions, subjective annotations
Action Unit / fine-grained studies	MMI	Posed expressions, AU coding but limited natural variability

5. Applications

The significance of employing facial expression recognition (FER) stems from its utilization in a multitude of applications, including the identification of emotions in masked faces for individuals with visual impairments [60], communications, driving [15], and human-computer interaction [14], [113]. Moreover, in education, emotional learning has been explored. For instance, [38] examined how students behaved in an online learning environment by utilizing video facial processing, tracking, and clustering to quickly and concurrently extract facial sequences for every student and predict group-level effects, individual emotions, and student engagement levels. Flip videos do not need to be sent to a distant server or teacher's computer, as the model can perform real-time video processing on a mobile device. By learning from brief clips that show the various emotions and engagement levels of each student, the model may also be used to generate a summary of the course [96], [67], [68], [61]. Facial expressions serve as crucial cues for assessing the efficacy of digital educational platforms and gauging student involvement in virtual learning environments [114]. In medicine, FER has been used for medical diagnosis [115], [116]. The identification of emotions expressed through facial cues holds considerable significance for a range of applications, including psychological profiling, autonomous driving, and public space security [3], as well as mental health and psychology [18], [117]. In security, FER can be applied in autonomous vehicles to identify the emotional states of drivers, passengers, or pedestrians. This application has the potential to enhance safety by alerting the vehicle to possible sources of distraction or hazardous situations, such as driver fatigue [118]. FER is also relevant in virtual reality and social robotics [16], as well as in emotion analysis and affective computing [119], [17]. In marketing, cameras equipped with artificial intelligence in shopping malls can analyze the immediate emotional responses of patrons, offering valuable applications in the field [49]. Data from customer facial expressions are utilized in retail sales to determine whether a human sales assistant is needed [4], [5].

6. Evaluation and Analysis

This section compares the majority of previous research using various algorithms to recognize facial expressions across a range of datasets. As a result of this comparison, the algorithms that achieved the highest accuracy in facial expression recognition are highlighted. For instance, a deep CNN achieved an accuracy of 69.3% on the AffectNet-8 dataset [60], while a model called POSTER++, based on a pre-trained Vision Transformer, achieved 67.49% accuracy on the AffectNet-7 dataset [14]. MobileNetV1 was used on the FER2013 [1] and RAF-DB [22] datasets, achieving accuracies of 97.9% and 95.05%, respectively. FER2013 remains one of the most widely used datasets in facial expression recognition. Table 5 summarizes recent research on facial expression recognition using various datasets that include both images and videos.

Table 5. Comparison between FER approaches.

Facial Expression Recognition Methodologies based on Machine learning				
Methods	**year**	**dataset**	**Accuracy**	**Code**
HOG + RF [8]	[2024]	JAFFE	7 class : 92.97%	————
SVM [54]	[2023]	MMI JAFFE	7 class : 98.44% 7 class : 98.44%	————
(LBP + SVM) (Viola Jones + SVM) [48]	[2021]	CK+ LFW CK+ LFW	7 class : 92.26% 7 class : 94.67% 7 class : 97.69% 7 class : 98.88%	————
HOG + SVM [7]	[2020]	JAFFE Cohn-Kanade	7 class : 97.62% 7 class : 98.61%	————

Continued on next page

Table 5 – continued from previous page

Methods	year	dataset	Accuracy	Code
(HOG and LBP) + (SVM and KNN) [48]	[2020]	LNMIIT MMI CK JAFFE	7 class : 96.02% 7 class : 92.22% 7 class : 91.30% 7 class : 97.70%	————
Decision Tree KNN [53]	[2020]	self-collected self-collected	6 class : 98.03% 6 class : 97.21%	————
WLD + SVM [120]	[2019]	CK+ MMI SFEW	6 class : 96.00% 7 class : 98.62% 7 class : 50.20%	————
SVM [51]	[2008]	Cohn-Kanade	7 class : 84.55%	————
Facial Expression Recognition Methodologies based on Deep learning				
Methods	**year**	**dataset**	**Accuracy**	**Code**
MM-Net [87]	[2024]	AffectNet RAF-DB FERPlus	7 class : 65.05% 7 class : 84.90% 7 class : 89.34%	————
Resnet-18 Swin-T [121]	[2024]	AffectNet AffectNet RAF-DB FERPlus	8 class : 62.16% 7 class : 65.73% 7 class : 89.77% 7 class : 92.31%	code1
Inception- Resnet-v2 [66]	[2024]	AffectNet SFEW Multi-PIE BU-3DFE	7 class : 62.7% 7 class : 58.2% 6 class : 93.4% 6 class : 88.3%	————
DFER [122]	[2024]	DFEW AFEW FERV39k MAFW	7 class : 68.73% 7 class : 55.00% 7 class : 47.80% 7 class : 47.44%	————
HAM [19]	[2024]	AffectNet AffectNet RAF-DB CAER-S	8 class : 63.82% 7 class : 66.97% 7 class : 91.92% 7 class : 92.86%	————
MobileNetV1 [1]	[2024]	FER2013	7 class : 97.9%	————
DAN [13]	[2023]	AffectNet AffectNet RAF-DB SFEW 2.0	8 class : 62.09% 7 class : 65.69% 7 class : 89.70% 7 class : 53.18%	code2
ResNet-18 [17]	[2023]	AffectNet RAF-DB	7 class : 63.03% 7 class : 85.69%	————
deep CNN [60]	[2023]	AffectNet FER-2013 DISFA	8 class : 69.3% —— ——	————
POSTER++ [14]	[2023]	AffectNet AffectNet RAF-DB	8 class : 63.77% 7 class : 67.49% 7 class : 92.21%	code3
ResNet-50 [68]	[2023]	RAF-DB FER2013	7 class : 86.72% 7 class : 88.13 %	————

Continued on next page

Table 5 – continued from previous page

Methods	year	dataset	Accuracy	Code
MobileNetV1 [22]	[2023]	RAF-DB CK+ FER2013 FERPLUS	7 class : 95.05% 7 class : 100.0% 7 class : 92.50% 8 class : 95.55%	code4
Resnet-50 [67]	[2023]	FER-2013 CK+ RAF-DB Own	7 class : 73.4% 7 class : 89.56% 7 class : 76.72% 8 class : 90.83%	————
Ad-Corre ResNet50 [39]	[2022]	AffectNet RAF-DB FER-2013	7 class : 63.36% 7 class : 86.96% 7 class : 72.03%	code5
Meta-Face2Exp [31]	[2022]	AffectNet RAF-DB	7 class : 64.23% 7 class : 88.58%	————
FERNet [45]	[2022]	FER2013	7 class : 69.57%	————
EfficientNet -B2&-B0 [69]	[2022]	AffectNet AffectNet VGAF AFEW	7 class : 66.34% 8 class : 63.03% 70.23% 59.27	code6
STSN/KTN [38]	[2021]	AffectNet AffectNet RAF-DB RAF-DB FERPlus FERPlus	63.03% STSN 63.97% KTN 87.52% STSN 88.07% KTN 89.66% STSN 90.49% KTN	————
(MA-Net) ResNet18 [21]	[2021]	AffectNet AffectNet RAF-DB SFEW CAER-S	7 class : 64.53% 8 class : 60.29% 7 class : 88.40% 59.40% 88.42%	code7
A-MobilNet [15]	[2021]	FERPLUS RAF-DB	8 class : 88.11% 7 class : 84.49%	————
DenseNet161 [18]	[2021]	KDEF JAFFE	7 class : 96.51% 7 class : 99.52%	————
DeepEmotion [32]	[2021]	CK+ Fer2013 FERG JAFFE	7 class : 98% 7 class : 70.02% 7 class : 99.3% 7 class : 92.8%	code8
RUL [24]	[2021]	FER2013 RAF-DB	7 class : 73.75% 7 class : 88.98%	————
FER-VT [34]	[2021]	RAF-DB FER+ CK+ CK+ CK+	7 class : 88.26% 8 class : 90.04% 6 class : 100.0% 7 class : 100.0% 8 class : 99.46%	————
		AffectNet	8 class : 59.89%	

Continued on next page

Table 5 – continued from previous page

Methods	year	dataset	Accuracy	Code
EfficientFace [61]	[2021]	AffectNet RAF-DB CAER-S	7 class : 63.70% 7 class : 88.36% 7 class : 85.87%	code9
DCNN-VC [55]	[2021]	UIBVFED FERG CK+ JAFFE TFEID	7 class : 98.85% 7 class : 99.96% 7 class : 99.04% 7 class : 99.57% 7 class : 99.31%	————
DACL [74]	[2021]	Affectnet RAF-DB	7 class : 65.20% 7 class : 87.78%	————
PSR (VGG16) [20]	[2020]	AffectNet AffectNet RAF-DB FER+	7 class : 63.77% 8 class : 60.68% 7 class : 88.98% 8 class : 89.75%	code10
MFMP [123]	[2020]	AffectNet	7 class : 58.93%	————
(ESR-9) [23]	[2020]	AffectNet FER+ CK+	8 class : 59.3% 8 class : 87.15% 8 class : 89.4%	code11
Inception-v3 [6]	[2020]	JAFFE FER2013	7 class : 99.66% 7 class : 90.16%	————
FMPN [16]	[2019]	AffectNet MMI CK+	7 class : 61.52% 6 class : 82.74% 8 class : 98.06%	code12
VGG ResNet Inception Deep CNN [96]	[2016]	Fer2013 Fer2013 Fer2013 Fer2013	7 class : 72.7% 7 class : 72.4% 7 class : 71.6% 7 class : 75.2%	code13
Facial Expression Recognition Methodologies based on Hybrid Techniques				
Methods	**year**	**dataset**	**Accuracy**	**Code**
hybrid local attention + VIT [89]	[2024]	RAF-DB FERPlus AffectNet	7 class : 90.45% 7 class : 90.13% 7 class : 65.07%	————
LSGTNet [124]	[2024]	FERV39K DFEW MMI Oulu-casia	7 class : 51.31% 7 class : 72.34% 6 class : 88.61% 6 class : 91.88%	————
ResNet18 + SVM [54]	[2023]	AffectNet FER2013 MMI JAFFE	8 class : 62.78% 7 class : 74.64% 6 class : 99.02% 7 class : 98.44%	————
CNN + DBN [91]	[2023]	JaFFE KDEF RaFD	7 class : 98.14% 7 class : 95.29% 7 class : 98.86%	————
CNN + LSTM [92]	[2023]	CK+ in-house	5 class : 84.87% 4 class : 92.84%	————
CNN + SVM [94]	[2022]	CK+ BU4D	7 class : 99.69% 6 class : 94.69%	————

Continued on next page

Table 5 – continued from previous page

Methods	year	dataset	Accuracy	Code
CNN + Harr Cascade [95]	[2022]	FER2013	7 class : 70.04%	————
EfficientNetB0 + CNN [88]	[2021]	FER2013	7 class : 74.39%	————
DCNN-VC [55]	[2021]	UIBVFED FERG CK+ JAFFE TFEID	7 class : 98.85% 7 class : 99.96% 7 class : 99.04% 7 class : 99.57% 7 class : 99.31%	————
CNN and BOVW + local SVM In [90]	[2019]	AffectNet AffectNet FER2013 FER+	8 class : 59.58% 7 class : 63.31% 7 class : 75.42% 7 class : 87.76%	————

This survey not only gathers facial expression recognition techniques but also critically examines the conceptual frameworks that explain their evolution, advantages, and disadvantages. The analysis reveals the methodological and cognitive foundations of FER models, going beyond technical comparisons. At its core, FER is a multidisciplinary challenge situated at the intersection of medical diagnosis, human-computer interaction, emotional education, and marketing. Traditional machine learning methods, such as those based on LBP, HOG, SVM, DT, RF, and KNN, reflect a feature-engineering paradigm grounded in the assumption that facial expressions can be manually decomposed into salient local features. While interpretable and lightweight, this approach is limited in handling in-the-wild variations such as low light, occlusions, and pose changes, revealing its conceptual limitations in real-world scenarios. Figure 4 shows some of the most commonly used datasets—AffectNet-8, AffectNet-7, RAF-DB, and FER2013—along with the best-performing models: DeepCNN, POSTER++, and MobileNetV1, which achieved accuracies of 69.3%, 67.49%, 95.05%, and 97.9%, respectively.

The representation learning paradigm, which hierarchically abstracts features from raw pixels, forms the foundation of deep learning-based techniques, including CNNs and attention-based networks. In [18], the authors used the DeepCNN model with the KDEF and JAFFE datasets, achieving accuracies of 96.51% and 99.52%, respectively. This model utilized transfer learning, which enhances the accuracy of emotion recognition, especially when using pre-trained DCNN models. A key benefit of this method is its capacity to recognize images taken from different angles, making it more suitable for practical applications. However, failure in learning transfer requires fine-tuning of model parameters, and the architecture of DNNs is relatively complex, potentially requiring significant computing resources for training. In [45], a simplified DeepCNN with five layers was used to reduce complexity. However, the model's ability to generalize to other datasets may be limited, as it was trained solely on the FER2013 dataset. Additionally, the model was trained on the AffectNet dataset in a different study [60], one of the most crucial datasets for FER. The model was applied to mask-covered images—an especially relevant application given the COVID-19 pandemic—achieving an accuracy of 69.3%. Nevertheless, the method may struggle to recognize complex emotions that rely on movements of the lower part of the face, indicating that further improvements are needed to enhance its accuracy and generalizability.

Moreover, context-aware modeling is added to FER through transfer learning, attention mechanisms, and hybrid architectures. According to these techniques, which are based on cognitive theories of visual saliency, certain facial features—such as the mouth and eyes—contribute disproportionately to emotional inference.

In [15], the A-MobileNet model, which uses the MobileNetV1 backbone, features a lightweight design due to its reduced parameter count, making it suitable for mobile applications where computational resources are limited. Improved feature extraction using an attention mechanism that focuses on critical facial regions—including the eyes, mouth, and forehead, which are important for expression recognition—enhances recognition accuracy without significantly increasing computational cost. A-MobileNet performs more accurately on the RAF-DB and FERPlus datasets than MobileNetV1 and MobileNetV2, validating the effectiveness of the modifications. However, it struggles to distinguish between expressions with high similarity, leading to sensitivity to intra-class expression similarity (e.g., fear and surprise), despite the use of attention mechanisms. This issue is not unique to this model; others such as [87] also struggle with expressions that exhibit high similarity across categories, suggesting potential limitations in separating these labels. To manage class similarity by reducing intra-class differences and maximizing inter-class differences, the authors of [125], [39], and [76] introduced novel loss functions. In [126], a quadratic cross-similarity network is proposed to handle cross-class similarity. This network uses cross-similarity attention mechanisms to identify salient features, thereby significantly reducing cross-class similarity in FER tasks.

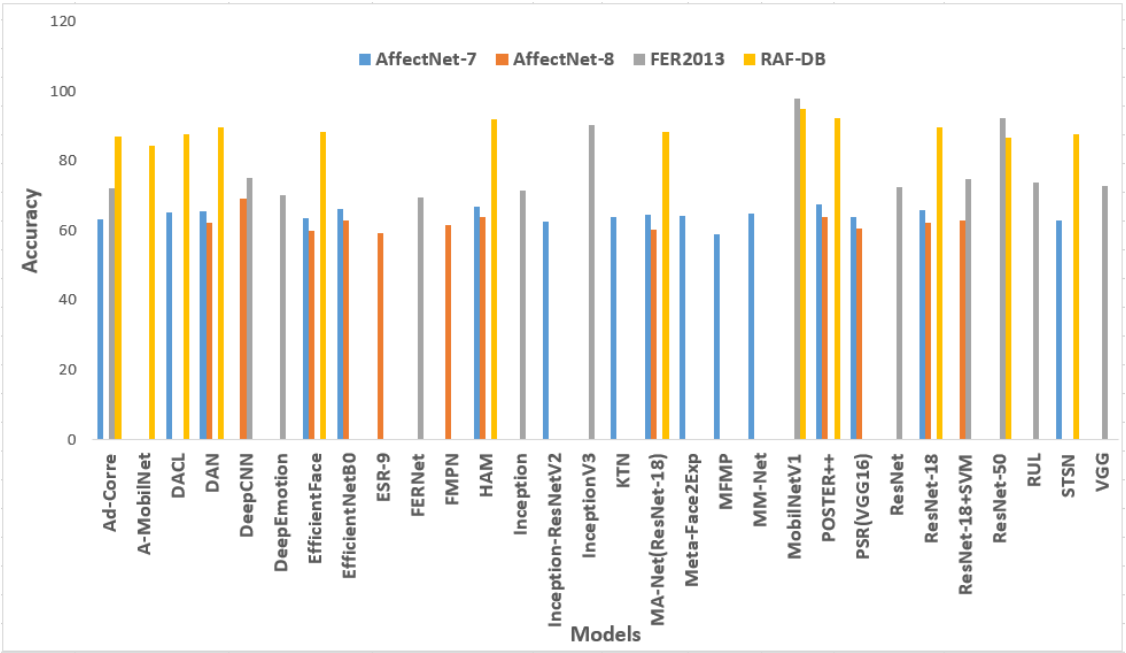

Figure 4. The most used datasets with the best models

FER also includes challenges such as pose variation and occlusion. In [66], the authors presented a method that addresses pose variation by introducing adversarial learning. This model generates pose-invariant feature representations, improving performance in diverse real-world scenarios. For robustness against occlusion and pose changes, [87] introduced the Multi-Granularity and Multi-Scale Network (MM-Net). MM-Net is a deep learning framework that captures facial features at multiple granularities (local and global) and scales, allowing the model to integrate both fine-grained and holistic information. By concentrating on different regions of the face, the network simplifies feature extraction through direct incorporation into the main architecture. Its advantages include improved robustness to variations in pose and occlusion while maintaining an efficient design. Experimental results show that MM-Net outperforms most state-of-the-art methods on datasets such as RAF-DB, FERPlus, and AffectNet, demonstrating its effectiveness across diverse FER benchmarks. In [19], augmented functionalities are introduced through the Hierarchical Attention Module (HAM). HAM is designed to apply attention at multiple levels of feature representation. This mechanism enables the model to emphasize salient features while suppressing irrelevant or redundant information. This advancement leads to improved feature integration, allowing the model to reliably differentiate between local and global facial emotion cues. To optimize spatial feature extraction, [122] employs a Graph Convolutional Network (GCN) model in conjunction with facial key points, which substantially enhances the model's resilience to occlusion and ambient noise. It effectively captures temporal dynamics through subjective attention mechanisms, making it suitable for video sequences characterized by evolving expressions over time. Additionally, it improves accuracy across heterogeneous datasets including DFEW, AFEW, FERV39k, and MAFW, demonstrating the model's generalizability and adaptability to a wide range of expression datasets.

Class imbalance is a significant issue in FER, where the distribution of facial expression classes is often highly skewed. Real-world facial expression datasets such as AffectNet, FER2013, and RAF-DB typically exhibit unbalanced data distributions. Some expressions, such as 'happiness' or 'neutrality', are overrepresented, while others, such as 'disgust' or 'contempt', are significantly underrepresented. This imbalance can severely impact model performance by biasing categorization toward dominant classes, resulting in poor recognition accuracy for minority expressions [27], [98], [97]. The imbalance is compounded by the inherent difficulty in collecting diverse and balanced expression datasets. For example, emotions such as 'fear' or 'disgust' occur less frequently in natural environments compared to 'happiness' [127]. To mitigate class imbalance, ensemble techniques such as boosting and bagging have been used on multiple models trained on balanced subsets of the data [128]. Using focal loss to address class imbalance has proven effective in FER tasks. By down-weighting well-classified examples and emphasizing hard, minority-class samples, focal loss significantly improves performance on imbalanced datasets such as FER2013. This approach leads to substantial gains in minority-class F1-scores. Similarly, adaptive regularization (AdaReg) loss combined with category re-weighting enhances the model's ability to discriminate rare expressions. This effectiveness has been demonstrated in comprehensive deep FER studies [129]. Additionally, generative adversarial networks (GANs) can be used to create high-quality artificial samples, while momentum self-supervised learning

techniques offer potential solutions to effectively address class imbalance in FER models [130]. Combining these methods in hybrid pipelines such as focal loss with GAN boosting can lead to the development of more robust and fair FER models. Future research should systematically evaluate these approaches across diverse and imbalanced datasets. Such evaluations are essential for ensuring ethical progress and enhancing the generalizability of FER systems [131]. In [132], the model effectively mitigates annotation ambiguity and class imbalance by computing class distributions and adaptively combining them with attention weights. To address the misclassification of minority classes resulting from class imbalance, focal loss was used to focus on examples that are difficult to classify [133].

Finally, loss functions play an important and subtle role in facial expression recognition. While traditional loss functions aim to minimize classification errors, they frequently overlook semantic similarities between certain emotional categories such as the closeness of fear and surprise. More advanced techniques, such as Ad-Corre and PDLS Loss, incorporate domain information into the training process. These strategies integrate psychological insights into the model's objectives, enhancing emotional awareness and the interpretability of machine learning models. Importantly, FER research often overlooks cultural, gender, and neurodivergent variables in expression and interpretation. The basic assumption that expressions are universal across populations is being challenged by research in psychology and anthropology. To gain a deeper understanding of FER, it is necessary to integrate culturally situated cognition and contextual emotional theories, moving beyond existing data-driven limitations. Moreover, Bias and fairness in FER are critical for avoiding discriminatory outcomes, ensuring equitable performance across demographic groups, and upholding ethical standards. In-the-wild datasets play a central role in real-world FER applications, it is essential to systematically analyze commonly used datasets to identify and address potential sources of bias. While some studies have partially explored these concerns, a more comprehensive and structured examination remains necessary. FER performance can vary significantly depending on the demographic composition and inherent biases present in the training data. Datasets that lack cultural, demographic, or socio-economic diversity risk perpetuating stereotypes and amplifying algorithmic bias. Additionally, annotation bias such as subjective emotion labeling and low inter-annotator agreement can critically undermine data reliability [134]. In [135], a detailed analysis of widely used in-the-wild datasets, including AffectNet, ExpW, FER2013, and RAF-DB, revealed significant performance disparities across gender, age, and ethnicity. Similarly, [134] demonstrated that systems trained on unbalanced datasets such as DISFA, FER2013, and AffectNet exhibit higher error rates for individuals from underrepresented ethnic and cultural groups. To mitigate these disparities, several debiasing strategies have been proposed. Score Normalization and Fair Score Normalization methods [136] have shown effectiveness in reducing demographic bias through post-processing adjustments of model outputs. Another promising line of research focuses on fairness-aware representation learning. The FADE method [137], for example, enables model training without requiring explicit sensitive attributes for all samples, thereby promoting fairer decision boundaries. It incorporates adversarial debiasing and federated learning techniques to account for population diversity while maintaining competitive performance. Beyond technical interventions, FER deployment also raises broader ethical and legal considerations, including privacy regulations, anti-discrimination laws, and biometric data governance. Researchers and practitioners must ensure compliance with these legal frameworks and establish clear accountability for potential model failures, misclassifications, and resulting harms [138]. Addressing these societal and ethical dimensions is essential to developing FER systems that are not only accurate but also trustworthy, inclusive, and aligned with human rights principles.

7. Future work

Future advancements in facial expression recognition should prioritize privacy, fairness, cultural adaptability, and interpretability. Privacy preservation can be achieved through federated learning and on-device inference, ensuring that sensitive facial data remain local while supporting collaborative model development and compliance with data protection regulations. Addressing demographic and class imbalances in datasets such as AffectNet and RAF-DB is also essential; synthetic data generation using StyleGAN or diffusion based models can enrich underrepresented emotion categories and promote fairness across age, gender, and ethnic groups. Moreover, since FER performance varies across cultural contexts, interdisciplinary collaboration with psychologists and sociologists is necessary to develop culturally sensitive emotion taxonomies and adaptive classification strategies, thereby enhancing generalizability. Emerging research directions include explainable artificial intelligence (XAI), which enables visualization of attention maps and feature attributions to improve transparency and trust; morphological neural computing, which explores event-driven, brain-inspired architectures for energy efficient, real-time emotion processing on mobile and embedded devices; and multimodal emotion understanding, which integrates facial cues with speech, gestures, and physiological signals to capture richer emotional contexts and enhance robustness in complex real-world environments.

8. Conclusions

Facial expression recognition has progressed from relying on handcrafted features to adopting deep learning approaches. This shift has greatly improved accuracy but also introduced challenges related to complexity, resource requirements, and fairness. Traditional approaches provided interpretability and efficiency but performed poorly in uncontrolled environments. In contrast, deep CNNs and attention-based architectures achieved state-of-the-art results, though they required large datasets and heavy computational resources. Despite progress, unresolved issues such as inter-class similarity, class imbalance, occlusion, cultural diversity, and label ambiguity still limit generalization in real-world applications. The next frontier in FER research goes beyond achieving higher benchmark accuracy. It lies in developing models that are lightweight, resilient to noise and occlusion, fair across demographics, and transparent in their decision-making. Promising directions include loss-function engineering, attention-enhanced networks, hybrid paradigms, and efficiency techniques such as pruning, quantization, and knowledge distillation. More importantly, combining advances in machine learning with insights from psychology and cognitive science can pave the way for context-aware systems. Such systems would be capable of capturing and interpreting the nuanced nature of human affect. Finally, the true impact of FER lies in enabling trustworthy, inclusive, and deployable systems for healthcare, education, safety, marketing, and human–computer interaction. Building the next generation of FER technologies will require more than just technical innovation. It must also prioritize fairness, efficiency, and ethical deployment in dynamic real-world contexts.

List of abbreviations

AdaReg	Adaptive Regular
AffectNet	Affect-in-the-Wild Database
Ad-Corre	Adaptive Correlation
ANN	Artificial Neural Network
AU	Action Unit
BU-3DFE	Binghamton University 3D Facial Expression
CBAM	Convolutional Block Attention Module
CNN	Convolutional Neural Network
CN	Classification Network
CK+	Cohn-Kanade Dataset
CAER-S	Context-aware emotion recognition networks
DFEW	Dynamic Facial Expression in the Wild
DDAMFN	Deep Dual Attention Multi-Fusion Network
DDAN	Dual-Direction Attention Network
DAN	Distract your Attention Network
DACL	Deep Attention Center Loss
DML	Deep Metric Learning
DT	Decision Tree
DFER	Dynamic Facial Expression Recognition
DCNN	Deep Convolutional Neural Network
ESR	Ensemble with Shared Representation
FMPN	Facial Motion Prior Networks
FACS	Facial Action Coding System
FMG	Facial Motion Generator
FER	Facial Expression Recognition
FC	Fully Connected
FERV39k	Facial Expression Recognition Video 39,000
GAP	Global Average Pooling
HAM	Hierarchical Attention Module
HOG	Histogram of Oriented Gradient
JAFFE	Japanese Female Facial Expression
KDEF	Karolinska Directed Emotional Faces
KTN	Knowledgeable Teacher Network
KNN	K-Nearest Neighbors)
LSGTNet	Local Spatial and Global Temporal Network
LDG	Label Distribution Generator
LBP	Local Binary Pattern
MAFW	Multi-Modal Affective in the Wild

Multi-PIE	Multi Pose, Illumination, Expressions
MMI	Matic Media Interface
MFMP	Multi-face Multi-part
MA-Net	Multi-scale and Local Attention Network
MM-Net	Multi-granularity and Multi-scale Network
PSR	Pyramid with Super-Resolution
PCA	principal component analysis
PDLs	Prior Distribution Label Smoothing
RAF-DB	Real-World Affective Database
RaFD	Radboud Faces Database
RUL	Relative Uncertainty Learning
RAN	Region Attention Network
SVM	Support Vector Machine
SFEW	Static Facial Expression in-the-Wild
STN	Spatial Transformer Network
STSN	Self-Taught Student Network
ViT	vision Transformer
VT	Visual Transformer
VC	Version Control

Declarations

Availability of data and materials
The dataset(s) for this paper were obtained from publicly available repositories and journal publications. Although most of the data sets were accessible, some required institutional permissions.

Competing interests
The authors declare no competing interests.

Funding
Not applicable.

Authors' contributions
This review was collaboratively conducted by all authors. MA wrote the main manuscript text and prepared figures and tables. AN has analyzed the results. AN and HH reviewed the manuscripts. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

REFERENCES

1. H. Haq, W. Akram, M. Irshad, A. Kosar, and M. Abid, "Enhanced real-time facial expression recognition using deep learning," *Acadlore Transactions on AI and Machine Learning*, 2024.
2. J. Zhou, S. Zhang, H. Mei, and D. Wang, "A method of facial expression recognition based on gabor and nmf," *Pattern Recognition and Image Analysis*, 2016.
3. B. Li and D. Lima, "Facial expression recognition via resnet-50," *International Journal of Cognitive Computing in Engineering*, 2021.
4. D. Caruelle, P. Shams, A. Gustafsson, and L. Lervik-Olsen, "Affective computing in marketing: practical implications and research opportunities afforded by emotionally intelligent machines," *Marketing Letters*, 2022.
5. P. Weichbroth and W. Sroka, "A note on the affective computing systems and machines: a classification and appraisal," *Procedia Computer Science*, 2022.
6. Y. Wang, Y. Li, Y. Song, and X. Rong, "The influence of the activation function in a convolution neural network model of facial expression recognition," *Applied Sciences*, 2020.
7. S. R. Supta, M. R. Sahriar, M. G. Rashed, D. Das, and R. Yasmin, "An effective facial expression recognition system," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, 2020.
8. V. A. Saeed, "A framework for recognition of facial expression using hog features," *International Journal of Mathematics, Statistics, and Computer Science*, 2024.

9. J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial expression recognition methods in the wild based on fusion feature of attention mechanism and lbp," *Sensors*, 2023.
10. S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE Transactions on Affective Computing*, 2020.
11. S. Singh and F. Nasoz, "Facial expression recognition with convolutional neural networks," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, 2020.
12. J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, 2020.
13. Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *Biomimetics*, 2023.
14. J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster++: a simpler and stronger facial expression recognition network," *ArXiv Preprint ArXiv:2301.12149*, 2023.
15. Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-mobilenet: An approach of facial expression recognition," *Alexandria Engineering Journal*, 2022.
16. Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in *2019 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2019.
17. J. L. Gómez-Sirvent, F. López de la Rosa, M. T. López, and A. Fernández-Caballero, "Facial expression recognition in the wild for low-resolution images using voting residual network," *Electronics*, 2023.
18. M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, 2021.
19. H. Tao and Q. Duan, "Hierarchical attention network with progressive feature fusion for facial expression recognition," *Neural Networks*, 2024.
20. T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, 2020.
21. Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, 2021.
22. J. L. Ngwe, K. M. Lim, C. P. Lee, and T. S. Ong, "Patt-lite: lightweight patch and attention mobilenet for challenging facial expression recognition," *ArXiv Preprint ArXiv:2306.09626*, 2023.
23. H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
24. Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, 2021.
25. V. Kumar, S. Rao, and L. Yu, "Noisy student training using body language dataset improves facial expression recognition," in *European Conference on Computer Vision*, Springer, 2020.
26. M. Lyons, M. Kamachi, and J. Gyoba, "The japanese female facial expression (jaffe) dataset," *The Images Are Provided at No Cost For Non-Commercial Scientific Research Only. If You Agree to The Conditions Listed Below, You May Request Access to Download*, 1998.
27. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, 2017.
28. I. Dominguez-Catena, D. Paternain, and M. Galar, "Metrics for dataset demographic bias: A case study on facial expression recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
29. I. Dominguez-Catena, D. Paternain, and M. Galar, "Dsap: Analyzing bias through demographic comparison of datasets," *Information Fusion*, 2025.
30. C.-L. Kim and B.-G. Kim, "Few-shot learning for facial expression recognition: a comprehensive survey," *Journal of Real-Time Image Processing*, 2023.
31. A. Psaroudakis and D. Kollias, "Mixaugment & mixup: augmentation methods for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
32. S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: facial expression recognition using attentional convolutional network," *Sensors*, 2021.
33. K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, 2020.
34. Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Information Sciences*, 2021.
35. D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019.
36. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: beyond empirical risk minimization," *ArXiv Preprint ArXiv:1710.09412*, 2017.
37. X. Zeng, Q. Wu, S. Zhang, Z. Liu, Q. Zhou, and M. Zhang, "A false trail to follow: differential effects of the facial feedback signals from the upper and lower face on the recognition of micro-expressions," *Frontiers in Psychology*, 2018.
38. H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Transactions on Image Processing*, 2021.
39. A. P. Fard and M. H. Mahoor, "Ad-corre: adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, 2022.
40. D. Almeida, K. Shmarko, and E. Lomas, "The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of us, eu, and uk regulatory frameworks," *AI and Ethics*, 2022.
41. M. P. Cross, A. M. Acevedo, and J. F. Hunter, "A critique of automated approaches to code facial expressions: What do researchers need to know?," *Affective Science*, 2023.

42. M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad, and J. J. Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, 2023.
43. Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: a survey," *Symmetry*, 2019.
44. D. Mungra, A. Agrawal, P. Sharma, S. Tanwar, and M. S. Obaidat, "Pratit: a cnn-based emotion recognition system using histogram equalization and data augmentation," *Multimedia Tools and Applications*, 2020.
45. J. D. Bodapati, U. Srilakshmi, and N. Veeranjaneyulu, "Fernet: a deep cnn architecture for facial expression recognition in the wild," *Journal of the Institution of Engineers (India): Series B*, 2022.
46. H. Wei and Z. Zhang, "A survey of facial expression recognition based on deep learning," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2020.
47. Y. Luo, C.-m. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of pca and lbp with svm," *Optik-International Journal for Light and Electron Optics*, 2013.
48. K. S. Yadav and J. Singha, "Facial expression recognition using modified viola-john's algorithm and knn classifier," *Multimedia Tools and Applications*, 2020.
49. S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations," *Information Fusion*, 2023.
50. L. Yao, Y. Wan, H. Ni, and B. Xu, "Action unit classification for facial expression recognition using active learning and svm," *Multimedia Tools and Applications*, 2021.
51. J.-M. Sun, X.-S. Pei, and S.-S. Zhou, "Facial emotion recognition in modern distant education system using svm," in *2008 International Conference on Machine Learning and Cybernetics*, IEEE, 2008.
52. S. Choi, E.-H. Kim, B. Ahn, and J.-H. Sohn, "Facial emotion recognition using k-nn and svm," *Proceedings of the Korean Society of Ergonomics Conference*, 2012.
53. M. Murugappan, A. Mutawa, S. Sruthi, A. Hassouneh, A. Abdulsalam, S. Jerritta, and R. Ranjana, "Facial expression classification using knn and decision tree classifiers," in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, IEEE, 2020.
54. I. Haider, H.-J. Yang, G.-S. Lee, and S.-H. Kim, "Robust human face emotion classification using triplet-loss-based deep cnn features and svm," *Sensors*, 2023.
55. V. R. R. Chirra, S. R. Uyyala, and V. K. K. Kolli, "Virtual facial expression recognition using deep cnn with ensemble learning," *Journal of Ambient Intelligence and Humanized Computing*, 2021.
56. Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016.
57. D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2017.
58. B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, 2016.
59. B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015.
60. M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, 2023.
61. Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
62. M. Rahimzadeh and A. Attar, "Detecting and counting pistachios based on deep learning," *Iran Journal of Computer Science*, 2022.
63. S. Li, "A brief review of deep learning for facial expression recognition," *Available at SSRN 4318896*, 2023.
64. 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*, 2015.
65. O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 2015.
66. S. Wang, Y. Chang, Q. Li, C. Wang, G. Li, and M. Mao, "Pose-robust personalized facial expression recognition through unsupervised multi-source domain adaptation," *Pattern Recognition*, 2024.
67. S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, 2023.
68. M. Aly, A. Ghallab, and I. S. Fathi, "Enhancing online learning with facial expression recognition system in online learning context using efficient deep learning model," *IEEE Access*, 2023.
69. A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, 2022.
70. X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: improving semantic feature learning in convolutional networks," *ArXiv Preprint ArXiv:1905.09646*, 2019.
71. Y. Liu, J. Peng, J. Zeng, and S. Shan, "Pose-adaptive hierarchical attention network for facial expression recognition," *ArXiv Preprint ArXiv:1905.10059*, 2019.
72. S. Xie, M. Li, S. Liu, and X. Tang, "Resnet with attention mechanism and deformable convolution for facial expression recognition," in *2021 4th International Conference on Information Communication and Signal Processing (ICICSP)*, 2021.
73. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: a survey," *ACM Computing Surveys (CSUR)*, 2022.
74. A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.
75. S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, 2023.

76. Y. Duan, J. Lu, and J. Zhou, "Uniformface: learning deep equidistributed representation for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
77. Q. T. Ngo and S. Yoon, "Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset," *Sensors*, 2020.
78. X. Guo, Y. Zhang, S. Lu, and Z. Lu, "Facial expression recognition: a review," *Multimedia Tools and Applications*, 2024.
79. N. N. Alajlan and D. M. Ibrahim, "Ddd tinyml: a tinyml-based driver drowsiness detection model using deep learning," *Sensors*, 2023.
80. C. Jia, X. Li, R. Qian, and H. Sun, "Facial expression recognition based on pruning optimization technology," *Highlights in Science, Engineering and Technology*, 2023.
81. M. Alam, P. Biswas, M. Rahman, *et al.*, "Light-fer: A lightweight facial emotion recognition," *Sensors*, 2022.
82. B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," *arXiv preprint arXiv:1712.05877*, 2018.
83. J. Lin, C. Gan, and S. Han, "Memory-driven mixed low precision quantization for enabling deep network inference on microcontrollers," *arXiv preprint arXiv:1905.13082*, 2019.
84. I. Cugu, E. Sener, and E. Akbas, "Microexpnet: An extremely small and fast model for expression recognition from face images," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2019.
85. Z. Zheng, C. Rasmussen, and X. Peng, "Student-teacher oneness: A storage-efficient approach that improves facial expression recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
86. A. Murtada, O. Abdelrhman, and T. A. Attia, "Mini-resemotenet: Leveraging knowledge distillation for human-centered design," *arXiv preprint arXiv:2501.18538*, 2025.
87. H. Xia, L. Lu, and S. Song, "Feature fusion of multi-granularity and multi-scale for facial expression recognition," *The Visual Computer*, 2024.
88. W. R. Abdulhussien, N. K. El Abbadi, and A. M. Gaber, "Hybrid deep neural network for facial expressions recognition," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 2021.
89. Y. Tian, J. Zhu, H. Yao, and D. Chen, "Facial expression recognition based on vision transformer with hybrid local attention," *Applied Sciences*, 2024.
90. M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, 2019.
91. A. J. Obaid and H. K. Alrammahi, "An intelligent facial expression recognition system using a hybrid deep convolutional neural network for multimedia applications," *Applied Sciences*, 2023.
92. M. Mohana, P. Subashini, and M. Krishnaveni, "Emotion recognition from facial expression using hybrid cnn–lstm network," *International Journal of Pattern Recognition and Artificial Intelligence*, 2023.
93. N. Khan, A. V. Singh, and R. Agrawal, "Enhanced deep learning hybrid model of cnn based on spatial transformer network for facial expression recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, 2022.
94. J.-C. Kim, M.-H. Kim, H.-E. Suh, M. T. Naseem, and C.-S. Lee, "Hybrid approach for facial expression recognition using convolutional neural networks and svm," *Applied Sciences*, 2022.
95. O. C. Oguine, K. J. Oguine, H. I. Bisallah, and D. Ofuani, "Hybrid facial expression recognition (fer2013) model for real-time emotion classification and prediction," *arXiv preprint arXiv:2206.09509*, 2022.
96. C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *ArXiv Preprint ArXiv:1612.02903*, 2016.
97. S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
98. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: a report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer, 2013.
99. E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of The 18th ACM International Conference on Multimodal Interaction*, 2016.
100. D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, Springer, 2017.
101. D. Lundqvist, A. Flykt, and A. Öhman, "Karolinska directed emotional faces," *PsycTESTS Dataset*, 1998.
102. J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of The IEEE/CVF International Conference On Computer Vision*, 2019.
103. 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, IEEE, 2010.
104. L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, IEEE, 2006.
105. X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of The 28th ACM International Conference on Multimedia*, 2020.
106. Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: a large-scale multi-scene dataset for facial expression recognition in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
107. Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "Mafw: a large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
108. A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015.

109. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.
110. M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, IEEE, 2005.
111. G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, 2011.
112. O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, 2010.
113. Z. Duric, W. D. Gray, R. Heishman, F. Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proceedings of the IEEE*, 2002.
114. W. Maqableh, F. Y. Alzyoud, and J. Zraqou, "The use of facial expressions in measuring students' interaction with distance learning environments during the covid-19 crisis," *Visual Informatics*, 2023.
115. B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, and L. Shapiro, "A facial affect analysis system for autism spectrum disorder," in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019.
116. R. Irani, K. Nasrollahi, M. O. Simon, C. A. Corneanu, S. Escalera, C. Bahnsen, D. H. Lundtoft, T. B. Moeslund, T. L. Pedersen, M.-L. Klitgaard, *et al.*, "Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.
117. S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and Its Applications*, IEEE, 2011.
118. M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, 2018.
119. D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: combating data biases for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
120. A. Mahmood, S. Hussain, K. Iqbal, and W. S. Elkilani, "Recognition of facial expressions under varying conditions using dual-feature fusion," *Mathematical Problems in Engineering*, 2019.
121. Y. Zhang, Y. Li, X. Liu, W. Deng, *et al.*, "Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition," *Advances in Neural Information Processing Systems*, 2024.
122. Z. Huang, Y. Zhu, H. Li, and D. Yang, "Dynamic facial expression recognition based on spatial key-points optimized region feature fusion and temporal self-attention," *Engineering Applications of Artificial Intelligence*, 2024.
123. S. Happy, A. Dantcheva, and F. Bremond, "Expression recognition with deep features extracted from holistic and part-based models," *Image and Vision Computing*, 2021.
124. L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A joint local spatial and global temporal cnn-transformer for dynamic facial expression recognition," *Applied Soft Computing*, 2024.
125. T.-D. Pham, M.-T. Duong, Q.-T. Ho, S. Lee, and M.-C. Hong, "Cnn-based facial expression recognition with simultaneous consideration of inter-class and intra-class variations," *Sensors*, 2023.
126. C. Wang, L. Chen, L. Wang, Z. Li, and X. Lv, "Qcs: Feature refining from quadruplet cross similarity for facial expression recognition," *arXiv preprint arXiv:2411.01988*, 2024.
127. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, 2009.
128. Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on knowledge and data engineering*, 2005.
129. P. Billion Polak, J. D. Prusa, and T. M. Khoshgoftaar, "Low-shot learning and class imbalance: a survey," *Journal of Big Data*, 2024.
130. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, 2014.
131. T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets," *Information*, 2024.
132. S. Liu, Y. Xu, T. Wan, and X. Kui, "Ada-df: An adaptive label distribution fusion network for facial expression recognition," *arXiv preprint arXiv:2404.15714*, 2024.
133. T. Lin, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.
134. M. Mattioli and F. Cabitza, "Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology," *Machine Learning and Knowledge Extraction*, 2024.
135. M. M. Hosseini, A. P. Fard, and M. H. Mahoor, "Faces of fairness: Examining bias in facial expression recognition datasets and models," *arXiv preprint arXiv:2502.11049*, 2025.
136. P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," *Pattern Recognition Letters*, 2020.
137. J. Hong, Z. Zhu, S. Yu, Z. Wang, H. H. Dodge, and J. Zhou, "Federated adversarial debiasing for fair and transferable representations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
138. A. Boutin, L. Lévêque, and S. Desmoulin-Canselier, "On legal and ethical challenges of automatic facial expression recognition: An exploratory study," in *Proceedings of the 2023 ACM international conference on interactive media experiences*, 2023.