Recent Advances and Challenges in Facial Emotion Recognition Systems: A Review of Deep Learning Approaches

Eris Zeqo[1], Viola Bakiasi (Shtino)[2], Bora Myrto (Lamaj)[3], Freda Dyrkaj[4]
[1,2,3,4] University "Aleksander Moisiu" of Durres, Faculty of Information Technology, Computer Science Department, Durres, Albania
[1]eriszeqo@uamd.edu.al, [2]violashtino@uamd.edu.al, [3]boramyrto@uamd.edu.al, [4]fredadyrkaj@umad.edu.al

***Abstract*—Facial Emotion Recognition (FER) has become one of the most active research areas in the fields of Artificial Intelligence and Computer Vision due to its wide range of applications in healthcare, education, human–computer interaction, security, and intelligent monitoring systems. The ability of computers to automatically interpret human emotions from facial expressions has significantly improved with the rapid development of machine learning and deep learning techniques. This review paper presents a comprehensive analysis of recent advances in facial emotion recognition systems, with particular attention to deep learning-based approaches. The study discusses the main stages of FER systems, including image acquisition, preprocessing, feature extraction, and classification. In addition, widely used public datasets such as FER2013, CK+, RAF-DB, and AffectNet are reviewed and compared based on their characteristics and research challenges. The paper further examines the transition from traditional machine learning algorithms to advanced convolutional neural network architectures that have achieved higher recognition accuracy in complex real-world scenarios. Moreover, critical challenges such as class imbalance, illumination variations, occlusions, head pose changes, and the recognition of subtle micro-expressions are analyzed. Finally, the paper highlights emerging research directions, including explainable artificial intelligence, multimodal emotion recognition, lightweight real-time systems, and transformer-based architectures. This review aims to provide researchers and practitioners with a clear overview of current developments and future opportunities in facial emotion recognition research.**

Keywords: **Facial Emotion Recognition, Deep Learning, Convolutional Neural Networks, Computer Vision, Machine Learning, Emotion Analysis.**

I. INTRODUCTION

Human emotions play a fundamental role in communication, decision-making, and social interaction. Facial expressions are among the most important nonverbal cues used to convey emotional states, making facial emotion recognition (FER) an important research topic in the fields of Artificial Intelligence, Machine Learning, and Computer Vision. According to Paul Ekman and Wallace V. Friesen (1971), facial expressions represent universal indicators of human emotions that can be recognized across different cultures. This finding significantly contributed to the development of automatic emotion recognition systems and affective computing research.

Facial emotion recognition systems are designed to identify emotional states such as happiness, sadness, anger, fear, surprise, disgust, and neutrality through the analysis of facial images or video sequences (Zeng et al., 2009). These systems have attracted increasing attention due to their broad applications in healthcare, online education, security systems, intelligent surveillance, virtual assistants, driver monitoring, and human–computer interaction. In healthcare, FER technologies can support emotional monitoring and mental health assessment, while in educational environments, they may help analyze student engagement and learning behavior.

Traditional FER approaches mainly relied on handcrafted feature extraction techniques combined with classical machine learning algorithms. Methods such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Principal Component Analysis (PCA), and Support Vector Machines (SVM) were widely used for facial feature representation and classification (Dalal & Triggs, 2005). Although these techniques achieved acceptable performance under controlled environments, their effectiveness often decreased when dealing with illumination changes, head pose variations, occlusions, and spontaneous facial expressions.

The rapid development of deep learning has significantly transformed facial emotion recognition research. In particular, Convolutional Neural Networks (CNNs) have demonstrated remarkable capabilities in automatically learning complex facial features directly from image data without requiring manual feature engineering (Krizhevsky et al., 2012). Deep learning architectures such as VGGNet (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), and hybrid CNN-based models have achieved higher recognition accuracy and improved robustness across multiple public datasets. Furthermore, transformer-based architectures and attention mechanisms have recently emerged as promising approaches for improving feature representation and capturing long-range spatial dependencies in facial images (Vaswani et al., 2017).

Several publicly available datasets have contributed to the advancement of FER research. Datasets such as FER2013 (Goodfellow et al., 2013), CK+ (Lucey et al., 2010), RAF-DB (Li et al., 2017), and AffectNet (Mollahosseini et al., 2017) are widely used for training and evaluating emotion recognition models. However, many of these datasets present important challenges, including class imbalance, low-resolution images, inconsistent annotations, and limited diversity in demographic characteristics.

Despite the substantial progress achieved in recent years, facial emotion recognition remains a challenging research problem. The recognition of subtle facial movements and micro-expressions is particularly difficult because of their short duration and low intensity. Moreover, real-world applications require FER systems to operate accurately under varying environmental conditions and in real-time scenarios. Ethical concerns related to privacy, bias, transparency, and responsible use of emotion recognition technologies have also become increasingly important in modern artificial intelligence research.

This review paper presents a comprehensive overview of recent advances in facial emotion recognition systems, with a primary focus on deep learning approaches. The study discusses commonly used datasets, machine learning and deep learning techniques, current challenges, and future research directions. The main objective of this paper is to provide researchers and practitioners with a clearer understanding of existing FER methodologies and the potential opportunities for developing more accurate, robust, and explainable emotion recognition systems.

II. Related Work

Research in facial emotion recognition (FER) has evolved significantly over the last two decades, moving from traditional handcrafted feature-based methods to advanced deep learning architectures capable of learning complex facial representations automatically. Early FER studies mainly focused on extracting geometric and appearance-based facial features combined with conventional machine learning classifiers. These approaches laid the foundation for modern emotion recognition systems and continue to influence current research directions.

One of the earliest and most influential contributions to emotion analysis was introduced by Paul Ekman, who demonstrated that several facial expressions correspond to universal emotional states recognizable across different cultures (Ekman & Friesen, 1971). This theory motivated the development of automatic systems capable of detecting emotions from facial movements and appearance changes.

Traditional FER approaches commonly used handcrafted feature extraction methods such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Gabor filters, and Principal Component Analysis (PCA). Among these techniques, HOG became particularly popular because of its ability to capture local edge and gradient structures associated with facial expressions (Dalal & Triggs, 2005). These handcrafted features were typically combined with machine learning classifiers including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Naive Bayes classifiers. SVM-based models achieved promising performance in controlled environments due to their capability to separate high-dimensional feature spaces effectively.

As the availability of large-scale facial datasets increased, researchers began exploring deep learning techniques for emotion recognition tasks. The success of deep Convolutional Neural Networks (CNNs) in image classification significantly influenced FER research. The work of Krizhevsky et al. (2012) demonstrated the effectiveness of deep CNN architectures in automatically learning hierarchical image representations from raw data. Inspired by these results, CNN-based FER models rapidly became dominant because they eliminated the need for manual feature engineering and improved recognition performance under more challenging conditions.

Several deep learning architectures have been applied successfully in FER systems. VGGNet introduced deeper convolutional layers capable of learning more discriminative visual features while maintaining a relatively simple network structure (Simonyan & Zisserman, 2015). Later, ResNet addressed the degradation problem in deep neural networks through residual learning, enabling the training of significantly deeper architectures with improved accuracy (He et al., 2016). These models achieved strong results on facial expression datasets and became widely adopted as backbone architectures for FER research.

Public datasets have also played an important role in advancing emotion recognition systems. FER2013 introduced by Goodfellow et al. (2013) became one of the most commonly used benchmark datasets due to its challenging grayscale facial images collected from real-world conditions. Similarly, CK+ provided high-quality posed facial expressions suitable for evaluating controlled FER models (Lucey et al., 2010). More recently, datasets such as RAF-DB and AffectNet introduced more realistic facial images containing variations in illumination, pose, occlusion, and background complexity, allowing researchers to evaluate model robustness in unconstrained environments (Li et al., 2017; Mollahosseini et al., 2017).

Recent studies have increasingly explored hybrid models that combine deep feature extraction with classical machine learning classifiers. In these approaches, CNNs are commonly used to extract high-level facial features, while classifiers such as SVM or XGBoost perform the final emotion classification. Hybrid architectures often improve classification stability and generalization performance, especially when working with limited training data or imbalanced datasets.

Another growing research direction involves attention mechanisms and transformer-based architectures. The transformer model proposed by Ashish Vaswani et al. (2017) introduced self-attention mechanisms capable of capturing long-range dependencies in image representations. These methods have recently been adapted for FER tasks to improve the recognition of subtle facial changes and micro-expressions.

Despite the remarkable progress achieved in facial emotion recognition (FER) research, several important challenges and limitations remain insufficiently addressed in current studies. Many existing FER systems continue to struggle under real-world conditions involving illumination variations, facial occlusions, head pose changes, low-resolution images, cultural diversity, and class imbalance. These factors significantly affect the robustness and generalization capability of both traditional machine learning and deep learning approaches. In addition, most studies primarily focus on improving recognition accuracy while giving limited

attention to explainability, fairness, computational efficiency, and privacy preservation. Ethical concerns related to user consent, algorithmic bias, and the responsible deployment of emotion recognition technologies have also become increasingly important within the artificial intelligence community. Furthermore, transformer-based architectures and multimodal FER systems remain relatively underexplored compared to conventional CNN-based approaches. Consequently, current research is increasingly directed toward developing more explainable, fair, interpretable, and adaptive FER systems capable of operating reliably and ethically across diverse real-world environments.

III. Publicly Available Datasets for Facial Emotion Recognition

The performance and reliability of facial emotion recognition (FER) systems are strongly influenced by the quality and diversity of the datasets used during training and evaluation. Publicly available datasets have played a crucial role in the development of machine learning and deep learning approaches for emotion recognition. These datasets contain facial images or video sequences annotated with emotional labels and are widely used for benchmarking and comparative analysis in FER research.

Early FER datasets were generally collected under controlled laboratory environments, where lighting conditions, facial poses, and background settings were carefully managed. Although such datasets helped researchers develop initial recognition models, they often lacked the variability necessary for real-world applications. More recent datasets have introduced images captured under unconstrained conditions, including variations in illumination, occlusions, facial orientations, image quality, and spontaneous expressions.

One of the most widely used FER datasets is FER2013, introduced by Goodfellow et al. (2013). This dataset contains 35,887 grayscale facial images with a resolution of 48 × 48 pixels categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. FER2013 became highly popular because it presents significant challenges related to low image quality, class imbalance, and large intra-class variability. Due to these characteristics, the dataset is frequently used for evaluating the robustness of deep learning models.

Another important dataset is CK+, which includes posed facial expression sequences collected in controlled laboratory settings (Lucey et al., 2010). CK+ provides high-quality facial images and detailed annotations, making it suitable for evaluating facial feature extraction methods and expression classification models. However, because most expressions are exaggerated and captured in controlled environments, the dataset may not fully represent spontaneous real-world emotions.

More realistic datasets have emerged to address the limitations of laboratory-collected data. RAF-DB contains facial images gathered from the internet under unconstrained conditions, including variations in age, ethnicity, illumination, and head pose (Li et al., 2017). The dataset is widely used to evaluate model generalization and robustness in practical applications.

Similarly, AffectNet is considered one of the largest publicly available facial expression datasets. It contains more than one million facial images collected from the web, with manually annotated emotional labels and dimensional emotion representations such as valence and arousal (Mollahosseini et al., 2017). AffectNet introduced significant diversity in facial appearance, background complexity, and emotional intensity, making it highly valuable for training deep neural networks.

Another commonly used dataset is JAFFE, which contains facial images of Japanese female subjects displaying posed emotional expressions. Although relatively small in size, JAFFE remains important in FER research due to its simplicity and standardized structure.

Despite the availability of these datasets, several challenges continue to affect FER performance. Many datasets suffer from class imbalance, where certain emotions such as happiness are overrepresented while emotions like disgust appear less frequently. This imbalance can bias classification models toward dominant classes and reduce recognition accuracy for minority emotions. Additionally, variations in facial occlusion, illumination, age, ethnicity, and image resolution further complicate the learning process.

Table 1. Summarizes some of the most widely used datasets in facial emotion recognition research.

Dataset	Number of Images	Emotion Classes	Characteristics
FER2013	35,887	7	Low-resolution grayscale images, class imbalance
CK+	593 sequences	7	Controlled environment, posed expressions
RAF-DB	~30,000	7 basic emotions	Real-world facial variations
AffectNet	>1,000,000	8+ emotions	Large-scale, web-collected dataset
JAFFE	213	7	Small dataset with posed expressions

The continuous development of more diverse and balanced datasets is essential for improving the robustness and fairness of facial emotion recognition systems. Future datasets are expected to include broader demographic diversity, spontaneous emotional behavior, multimodal information, and more accurate annotations to support the development of reliable real-world FER applications.

IV. Machine Learning Approaches

Before the dominance of deep learning, facial emotion recognition systems were mainly developed using traditional machine learning methods. These approaches usually followed a structured pipeline: face detection, image preprocessing, feature extraction,

feature selection, and final emotion classification. In this type of system, the quality of the extracted features strongly influenced the final recognition performance.

Classical machine learning methods do not learn facial representations directly from raw images. Instead, they depend on handcrafted descriptors that capture visual information such as edges, textures, gradients, and facial appearance changes. Techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Gabor filters, and Principal Component Analysis (PCA) have been widely used to represent facial expressions in a compact numerical form. LBP, for example, is effective for describing local texture patterns of the face and has shown good performance in expression recognition tasks (Shan et al., 2009). Similarly, HOG captures gradient orientation information and is useful for representing facial contours and expression-related shape changes (Dalal & Triggs, 2005).

After feature extraction, classification algorithms are applied to assign each facial image to a specific emotional category. Support Vector Machines (SVM) have been among the most commonly used classifiers in FER research because they perform well in high-dimensional feature spaces and can create effective decision boundaries between emotion classes. SVM-based systems have been particularly successful when combined with robust handcrafted descriptors such as LBP, HOG, or Gabor features (Happy & Routray, 2015).

Other traditional classifiers, including K-Nearest Neighbors (KNN), Decision Trees, Random Forests, and Naive Bayes, have also been used in facial emotion recognition. KNN is simple and intuitive, but its performance may decrease when the dataset is large or contains noisy samples. Naive Bayes is computationally efficient, but its assumption of feature independence can limit its effectiveness in complex facial expression data. Random Forests provide better stability than single decision trees because they combine multiple decision models, reducing the risk of overfitting.

Although traditional machine learning methods are less computationally demanding than deep neural networks, they have several limitations. Their performance depends heavily on the design of handcrafted features, which may not fully capture complex emotional patterns. In real-world conditions, factors such as illumination changes, head pose variations, occlusions, low image resolution, and spontaneous expressions can significantly reduce their accuracy. For this reason, many recent studies have moved toward deep learning methods, which can automatically learn more discriminative and hierarchical features from facial images.

Table 2 presents a comparison of commonly used traditional machine learning approaches in facial emotion recognition systems. While these methods provide efficient performance in controlled environments, their robustness often decreases under real-world conditions involving occlusions, illumination changes, and spontaneous facial expressions.

TABLE 2. COMPARISON OF TRADITIONAL MACHINE LEARNING APPROACHES IN FER.

Method	Main Characteristics	Advantages	Limitations
SVM	Margin-based classifier for high-dimensional data	High classification accuracy, effective with small datasets	Sensitive to parameter tuning
KNN	Instance-based learning using distance metrics	Simple and easy to implement	High computational cost for large datasets
Naive Bayes	Probabilistic classifier based on Bayes theorem	Fast and computationally efficient	Assumes feature independence
Decision Tree	Tree-structured classification model	Interpretable and easy to visualize	Risk of overfitting
Random Forest	Ensemble of multiple decision trees	Improved stability and robustness	Higher computational complexity
HOG + SVM	Combination of handcrafted features and SVM	Effective in controlled environments	Limited robustness in real-world conditions
LBP + SVM	Texture-based facial representation with SVM	Good texture description capability	Sensitive to illumination variations

However, classical machine learning approaches remain important in FER research. They are useful for small datasets, low-resource environments, and comparative studies. In addition, hybrid models that combine deep feature extraction with traditional classifiers such as SVM or Random Forest can provide strong performance, especially when the extracted deep features are well-structured and informative. Therefore, machine learning approaches continue to serve as a valuable foundation for understanding the evolution of facial emotion recognition systems.

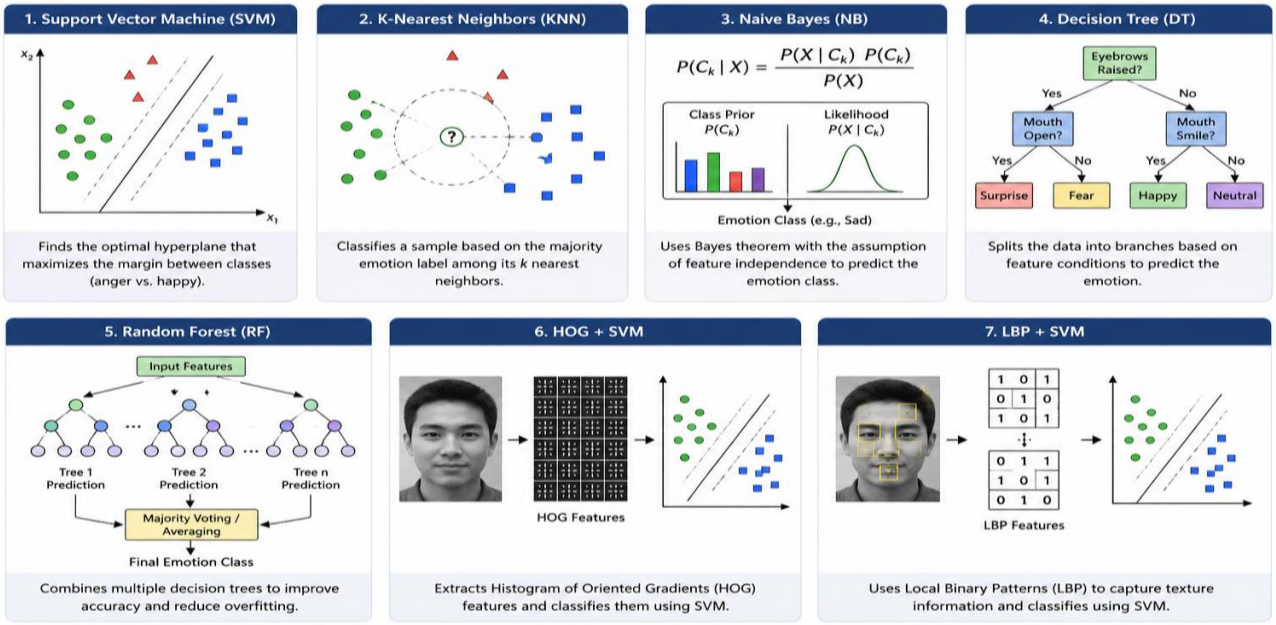

Figure 1. Traditional Machine Learning Techniques Used in Facial Emotion Recognition.

Figure 1 illustrates several widely used traditional machine learning approaches applied in facial emotion recognition systems, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Random Forests, HOG + SVM, and LBP + SVM. The figure highlights the main classification principles and feature extraction mechanisms of each technique. These methods contributed significantly to the early development of FER systems and remain important for comparative analysis and hybrid deep learning architectures.

V. Deep Learning Approaches

The rapid advancement of deep learning has significantly transformed the field of facial emotion recognition (FER). Unlike traditional machine learning methods that depend on handcrafted feature extraction, deep learning models are capable of automatically learning hierarchical and discriminative facial representations directly from raw image data. This ability has allowed modern FER systems to achieve higher recognition accuracy and better robustness in complex real-world environments.

Among deep learning techniques, Convolutional Neural Networks (CNNs) have become the dominant architecture for facial emotion recognition tasks. CNNs are designed to automatically detect spatial patterns such as facial contours, textures, edges, and expression-related features through convolutional operations and hierarchical feature learning. The success of CNNs in image classification was first demonstrated by Krizhevsky et al. (2012), whose deep neural network architecture achieved remarkable performance improvements in computer vision applications.

In FER research, CNN-based models are widely used because they eliminate the need for manual feature engineering. During training, convolutional layers learn low-level features such as edges and textures, while deeper layers capture more complex emotional representations and facial structures. Pooling layers reduce spatial dimensions and computational complexity, whereas fully connected layers perform the final emotion classification.

Several deep learning architectures have been successfully adapted for facial emotion recognition. VGGNet introduced deeper network structures using small convolutional filters and demonstrated strong feature extraction capabilities for image recognition tasks (Simonyan & Zisserman, 2015). ResNet later improved deep neural network training through residual connections that help overcome vanishing gradient problems in very deep architectures (He et al., 2016). These architectures became widely used backbone models in FER systems because of their strong generalization performance.

More recently, EfficientNet architectures have gained attention due to their balanced scaling of network depth, width, and image resolution. EfficientNet models can achieve competitive recognition performance while maintaining lower computational complexity, making them suitable for real-time FER applications and resource-constrained environments.

In addition to standalone CNN architectures, many researchers have explored hybrid deep learning approaches. In these systems, CNNs are commonly used for deep feature extraction, while traditional machine learning classifiers such as Support Vector Machines (SVM), Random Forests, or XGBoost perform the final classification task. Hybrid architectures can improve classification stability and generalization, especially when training data is limited or class imbalance is present.

Table 3 presents a comparison of commonly used deep learning architectures in facial emotion recognition systems. These approaches have significantly improved recognition accuracy by automatically learning complex facial representations from image data. However, challenges related to computational complexity, dataset requirements, and model interpretability still remain important research issues.

Table 3. Comparison of Deep Learning Architectures in Facial Emotion Recognition.

Architecture	Main Characteristics	Advantages	Limitations
CNN	Automatically learns facial features from images	High recognition accuracy	Requires large training datasets
VGGNet	Deep architecture with small convolution filters	Strong feature extraction capability	High computational cost
ResNet	Uses residual connections for deeper networks	Reduces vanishing gradient problem	Complex architecture
EfficientNet	Balanced scaling of depth, width, and resolution	Computationally efficient	Requires careful parameter tuning
CNN + SVM	Deep feature extraction with SVM classification	Improved generalization performance	Multi-stage training process
Transformer-Based Models	Uses self-attention mechanisms	Captures long-range dependencies	Requires high computational resources

Attention mechanisms and transformer-based models have also become emerging research directions in FER. Transformer architectures rely on self-attention mechanisms capable of capturing long-range dependencies and contextual facial relationships more effectively than traditional convolution operations (Vaswani et al., 2017). These models have shown promising performance in recognizing subtle facial expressions and micro-expressions, which are often difficult to detect using conventional CNNs.

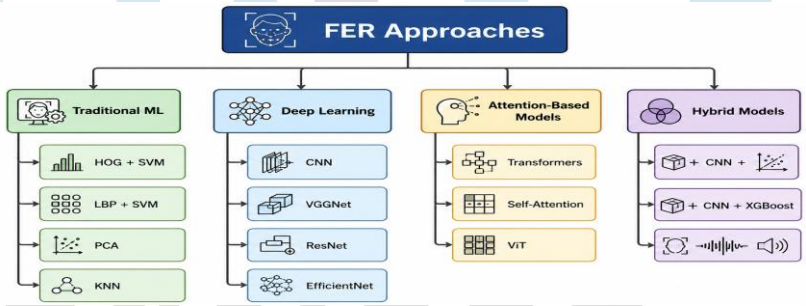

Figure 2. Taxonomy of Facial Emotion Recognition Approaches.

Figure 2 presents a taxonomy of the main approaches used in Facial Emotion Recognition (FER) systems. The figure categorizes FER methodologies into four major groups: Traditional Machine Learning approaches, Deep Learning architectures, Attention-Based Models, and Hybrid Models. Traditional machine learning techniques, including HOG + SVM, LBP + SVM, PCA, and KNN, mainly rely on handcrafted feature extraction methods combined with classical classifiers. In contrast, deep learning approaches such as CNN, VGGNet, ResNet, and EfficientNet automatically learn hierarchical facial representations directly from image data, significantly improving recognition accuracy and robustness. The figure also highlights the growing importance of attention-based models, including Transformers, Self-Attention mechanisms, and Vision Transformers (ViT), which are capable of capturing long-range contextual dependencies and subtle facial variations. Furthermore, hybrid models combine deep feature extraction with traditional classifiers such as SVM and XGBoost, aiming to improve classification stability and generalization performance. Overall, the figure demonstrates the evolution of FER systems from conventional handcrafted approaches toward more advanced, intelligent, and adaptive deep learning-based frameworks.

Recently, Vision Transformers (ViTs) and transformer-based architectures such as Swin Transformers have attracted increasing attention in FER research due to their ability to capture global contextual relationships more effectively than traditional convolutional networks. These models have demonstrated promising performance in recognizing subtle facial expressions and improving robustness under complex environmental conditions.

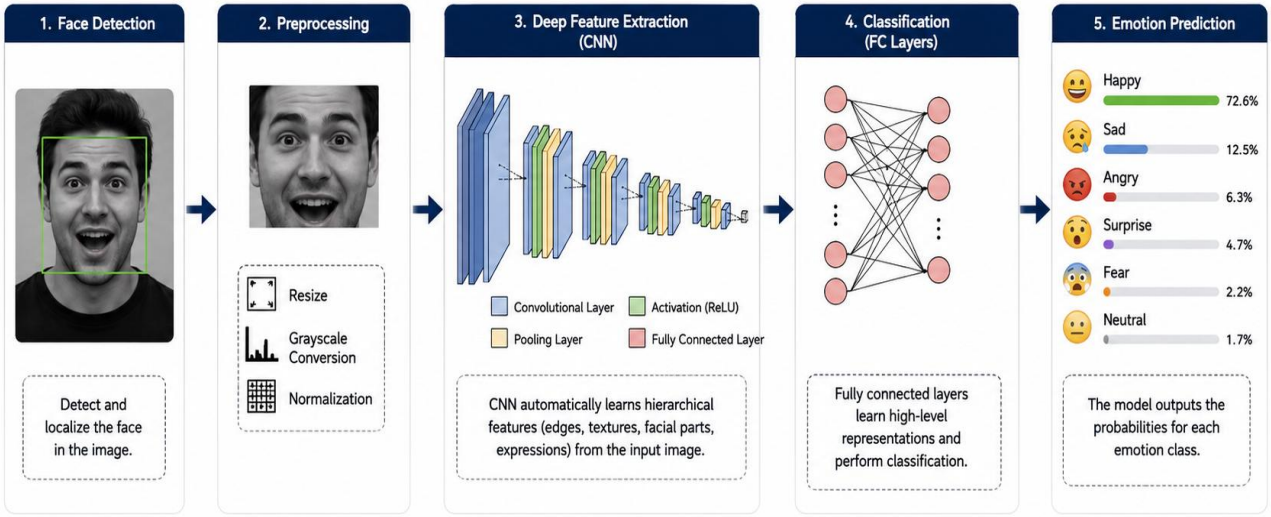

Figure 3. General Deep Learning Pipeline for Facial Emotion Recognition.

Figure 3 illustrates the general workflow of deep learning-based facial emotion recognition systems. The process typically includes face detection, image preprocessing, deep feature extraction using convolutional neural networks, and final emotion classification. Deep learning architectures enable automatic extraction of hierarchical facial features, improving recognition performance in complex real-world scenarios.

Despite their strong performance, deep learning models still face several challenges in facial emotion recognition. Large annotated datasets are usually required for effective training, and deep architectures often involve high computational costs. Overfitting may occur when datasets are small or imbalanced, while real-world variations such as occlusions, illumination changes, facial accessories, and head pose differences continue to affect recognition accuracy. Furthermore, deep learning models are frequently criticized for their limited interpretability, making explainable artificial intelligence an important research direction in FER systems.

Overall, deep learning approaches have significantly improved the performance and scalability of facial emotion recognition technologies. Their capability to learn complex emotional representations automatically has enabled FER systems to become more accurate, adaptive, and applicable across a wide range of intelligent systems and human-centered applications.

VI. Challenges in FER

Despite the significant progress achieved in facial emotion recognition (FER) systems, several technical and practical challenges continue to limit their performance in real-world applications. Although deep learning models have improved recognition accuracy considerably, facial emotion analysis remains a complex task due to the variability of human expressions, environmental conditions, and dataset limitations (Sariyanidi et al., 2015).

One of the most common challenges in FER research is class imbalance. In many publicly available datasets, certain emotions such as happiness and neutrality appear much more frequently than emotions like disgust or fear. This imbalance can bias machine learning models toward majority classes, reducing classification accuracy for underrepresented emotions. As a result, FER systems may perform well overall while still struggling to recognize minority emotional states correctly. Researchers have attempted to address this issue using techniques such as data augmentation, oversampling, weighted loss functions, and synthetic data generation.

Another important challenge involves illumination variations and environmental conditions. Facial images collected under uncontrolled settings often contain shadows, low lighting, brightness changes, and background noise. These variations can significantly affect facial feature extraction and reduce the stability of recognition models (Zeng et al., 2009). Traditional handcrafted feature methods are particularly sensitive to illumination changes, while deep learning models require large and diverse datasets to improve robustness under such conditions.

Occlusions also represent a major difficulty in FER systems. Facial accessories such as glasses, masks, hats, scarves, or even hand movements may partially cover important facial regions, making emotion recognition more difficult. The widespread use of face masks during global health emergencies further highlighted the limitations of FER models when large facial areas are hidden. Occlusion-aware deep learning methods and attention mechanisms have recently been explored to address this problem (Li et al., 2018).

Head pose variations and facial orientation changes present additional challenges. Many emotion recognition models are trained primarily on frontal facial images, causing performance degradation when faces appear rotated, tilted, or partially visible. Real-world applications such as surveillance systems and mobile devices require FER systems capable of handling multi-view facial expressions under dynamic conditions.

The recognition of subtle emotions and micro-expressions remains one of the most difficult problems in affective computing research. Micro-expressions are very brief and involuntary facial movements that occur within fractions of a second and often reveal hidden emotional states. Because these expressions involve minimal facial muscle movement and short temporal duration, detecting them requires highly sensitive feature extraction and temporal modeling techniques. Researchers increasingly use high-frame-rate video analysis, attention mechanisms, and spatiotemporal deep learning models to improve micro-expression recognition performance (Ekman, 2009).

Dataset quality and annotation consistency also influence FER performance. Some datasets contain low-resolution images, inaccurate labels, or limited demographic diversity. Cultural differences in emotional expression may further affect model generalization across populations. Consequently, models trained on specific datasets may not always perform effectively in real-world multicultural environments (Ko, 2018).

Another important concern relates to computational complexity and real-time processing requirements. Deep neural networks often require significant memory resources, high computational power, and long training times. This limitation becomes particularly relevant in mobile applications, embedded systems, and edge computing environments where hardware resources are constrained.

In addition to technical limitations, ethical and privacy concerns have become increasingly important in facial emotion recognition research. FER technologies process sensitive biometric information and may raise concerns regarding user consent, surveillance, algorithmic bias, and misuse of emotional data. Researchers and developers are therefore encouraged to design more transparent, fair, and explainable FER systems that respect privacy regulations and ethical principles.

Although substantial advancements have been achieved in recent years, these challenges demonstrate that facial emotion recognition remains an evolving research field. Future FER systems must become more robust, interpretable, computationally efficient, and ethically responsible in order to support reliable deployment in real-world applications.

VII. Future Directions

Facial emotion recognition (FER) continues to evolve rapidly due to advancements in Artificial Intelligence, Deep Learning, and Computer Vision. Although substantial progress has been achieved in recent years, researchers are increasingly focusing on developing more robust, efficient, explainable, and human-centered FER systems capable of operating reliably in real-world environments. Several emerging research directions are expected to shape the future development of emotion recognition technologies.

One important direction involves the integration of multimodal emotion recognition systems. Traditional FER approaches rely mainly on facial images; however, human emotions are also expressed through speech, body gestures, eye movements, and physiological signals. Combining multiple modalities such as facial expressions, voice analysis, electroencephalography (EEG), and text sentiment analysis can significantly improve emotion recognition accuracy and contextual understanding (Zeng et al., 2009). Multimodal systems are therefore expected to become increasingly important in healthcare, education, and human–computer interaction applications.

Another promising research direction is the development of explainable artificial intelligence (XAI) for FER systems. Deep learning models often operate as “black-box” systems, making it difficult to understand how emotional predictions are generated. This lack of interpretability may reduce trust and transparency in sensitive applications such as mental health assessment or security monitoring. Explainable FER models aim to provide visual or statistical explanations regarding which facial regions and features contribute most to the final prediction. Attention visualization methods, saliency maps, and interpretable neural networks are increasingly being explored to address this challenge.

Transformer-based architectures and attention mechanisms also represent an emerging trend in FER research. Unlike traditional convolutional neural networks that focus mainly on local spatial information, transformers can model long-range dependencies and contextual relationships more effectively through self-attention mechanisms (Vaswani et al., 2017). Recent studies suggest that transformer-based FER systems may improve the recognition of subtle facial expressions and micro-expressions that are difficult to detect using conventional CNN architectures.

Real-time and lightweight FER systems are another important future direction. Many practical applications, including mobile devices, smart surveillance systems, social robots, and driver monitoring platforms, require fast and computationally efficient emotion recognition models. Researchers are therefore developing lightweight neural networks and edge-based artificial intelligence solutions capable of operating with limited hardware resources while maintaining high recognition accuracy.

Another area receiving increasing attention is cross-cultural and fair emotion recognition. Emotional expressions may vary across individuals, cultures, genders, and age groups, potentially introducing bias into FER models trained on limited or unbalanced datasets. Future research is expected to focus on creating more diverse datasets and developing fair machine learning algorithms that reduce demographic bias and improve model generalization across different populations.

Micro-expression recognition is also expected to remain a major research topic. Since micro-expressions are involuntary and extremely brief, they provide valuable information about hidden emotional states. Advances in high-speed video analysis, temporal modeling, and spatiotemporal neural networks may significantly improve the detection of subtle facial movements in future FER systems.

Privacy-preserving emotion recognition technologies are becoming increasingly important as FER applications expand into public and commercial environments. Future systems will likely incorporate secure data processing methods, federated learning approaches, and privacy-aware artificial intelligence frameworks to protect sensitive biometric information and comply with ethical regulations.

Finally, future FER systems are expected to become more adaptive and personalized. Instead of relying only on generalized emotion models, next-generation systems may learn user-specific emotional patterns and contextual behaviors to provide more accurate and individualized predictions. Such developments could improve the effectiveness of intelligent tutoring systems, mental health monitoring platforms, and social assistive robotics.

Overall, the future of facial emotion recognition research lies in creating intelligent systems that are not only highly accurate but also interpretable, fair, computationally efficient, and ethically responsible. Continued advancements in deep learning, multimodal analysis, and explainable artificial intelligence are expected to further expand the capabilities and real-world applicability of FER technologies.

Although substantial progress has been achieved in FER research, several open challenges still require further investigation. Current systems often lack robustness across diverse demographic groups and uncontrolled environments. In addition, explainability and transparency remain major concerns in deep learning-based FER systems, particularly in healthcare and surveillance applications. Real-time deployment on edge devices also presents computational limitations. Future research should focus on developing lightweight, explainable, and fair FER models capable of operating reliably under real-world conditions while preserving user privacy and ethical standards.

VIII. Discussion

Facial emotion recognition (FER) has experienced remarkable progress over the past decade due to the rapid advancement of Deep Learning and modern computational techniques. The transition from traditional handcrafted feature-based approaches to deep neural network architectures has significantly improved recognition accuracy and robustness across various datasets and application domains. Nevertheless, the comparison of existing studies indicates that no single FER model can yet guarantee consistently reliable performance under all real-world conditions.

Traditional machine learning methods such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes played an essential role in the early development of FER systems. These approaches were computationally efficient and relatively simple to implement, particularly when combined with handcrafted descriptors such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). However, their effectiveness depended heavily on manually designed feature extraction techniques, which often struggled to capture complex emotional patterns and environmental variability (Shan et al., 2009).

Deep learning approaches, particularly Convolutional Neural Networks (CNNs), introduced substantial improvements by automatically learning hierarchical facial representations directly from image data. CNN-based models demonstrated superior performance on challenging datasets such as FER2013 and RAF-DB because of their ability to model complex spatial features and nonlinear emotional patterns. Architectures such as VGGNet, ResNet, and EfficientNet further improved recognition accuracy while reducing some limitations associated with traditional methods (He et al., 2016).

Despite these advances, several studies continue to report performance degradation when FER systems are tested in uncontrolled environments. Variations in illumination, occlusions, facial orientation, image quality, and spontaneous expressions remain major obstacles for both traditional and deep learning approaches. Furthermore, many public datasets suffer from class imbalance, where certain emotions are overrepresented while others appear infrequently. This imbalance may bias classifiers toward dominant emotion categories and reduce the recognition capability for minority classes.

Another important observation from the literature is that hybrid models combining deep feature extraction with traditional classifiers often achieve competitive performance. In such approaches, CNNs are used to extract discriminative facial features, while classifiers such as SVM or XGBoost perform the final classification task. Several researchers have reported that these hybrid systems can improve classification stability and generalization, especially when working with limited training data.

Transformer-based architectures and attention mechanisms also represent promising research directions. Unlike standard convolutional operations, attention-based models can capture long-range relationships between facial regions and better focus on subtle emotional changes (Vaswani et al., 2017). These techniques may become particularly important for micro-expression recognition and real-time emotion analysis applications.

The literature further highlights the growing importance of ethical and privacy considerations in FER research. Since emotion recognition systems process sensitive biometric information, concerns regarding surveillance, user consent, demographic bias, and misuse of emotional data have become increasingly relevant. As a result, modern FER research is gradually shifting toward explainable, fair, and privacy-aware artificial intelligence systems.

Overall, the reviewed studies demonstrate that facial emotion recognition remains a highly active and interdisciplinary research area. Although deep learning approaches currently dominate FER research due to their strong performance, future systems will likely depend on multimodal analysis, explainable artificial intelligence, lightweight architectures, and ethically responsible design principles in order to achieve broader real-world adoption.

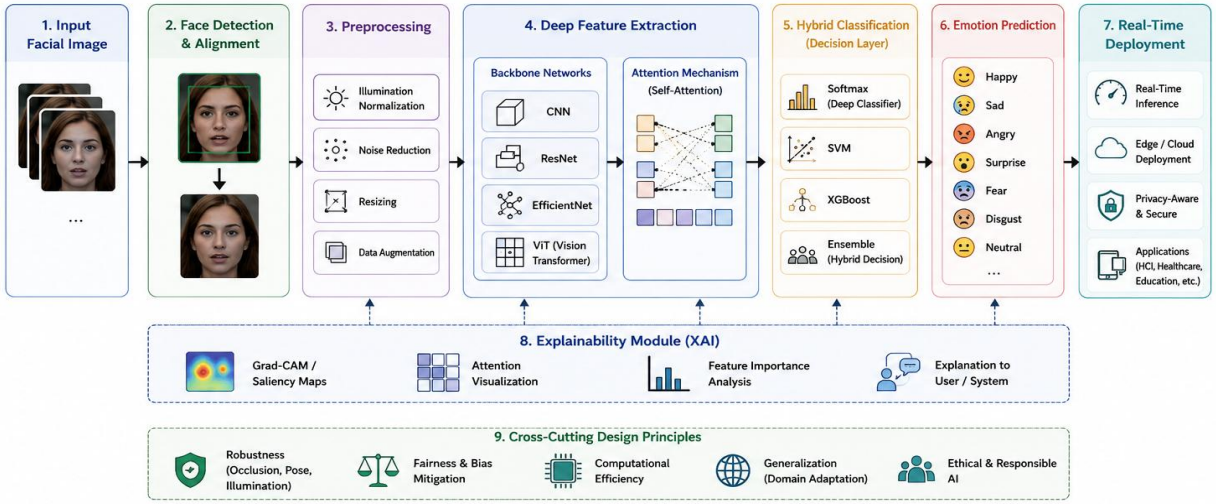

Figure 4. Proposed Intelligent FER Framework for Robust and Explainable Facial Emotion Recognition.

Figure 4 presents a conceptual framework for next-generation facial emotion recognition (FER) systems designed for robust, explainable, and real-world intelligent applications. The proposed framework integrates all major stages of FER, including face detection, preprocessing, deep feature extraction, attention mechanisms, hybrid classification, explainable artificial intelligence, and real-time deployment. Advanced deep learning architectures such as CNNs, ResNet, EfficientNet, and Vision Transformers (ViT) are utilized to automatically learn hierarchical and contextual facial representations. In addition, attention mechanisms are incorporated to improve the recognition of subtle facial expressions and long-range dependencies. The framework further combines deep learning approaches with hybrid classifiers such as SVM and XGBoost to enhance classification stability and generalization performance. An explainability module based on attention visualization and saliency analysis is also included to improve transparency and interpretability. Overall, the proposed framework highlights the future direction of FER systems toward more adaptive, privacy-aware, computationally efficient, and ethically responsible intelligent solutions.

This review mainly focuses on visual facial emotion recognition approaches and publicly available FER datasets. Other modalities such as physiological signals, speech-based emotion recognition, and multimodal affective computing systems were discussed only briefly. In addition, the paper does not include experimental benchmarking or meta-analytical statistical evaluation of existing FER models.

IX. Comparative Analysis of Facial Emotion Recognition Approaches

Comparative analysis plays an important role in evaluating the effectiveness of facial emotion recognition (FER) methods across different datasets and experimental conditions. The literature shows that both traditional machine learning techniques and deep learning architectures have contributed significantly to the advancement of FER systems. However, their performance varies depending on dataset complexity, feature extraction methods, computational resources, and environmental conditions.

Traditional machine learning approaches such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes generally perform well when combined with handcrafted feature extraction techniques such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). These methods are computationally efficient and suitable for smaller datasets or low-resource systems. Nevertheless, their ability to generalize under unconstrained real-world conditions remains limited because handcrafted features may fail to capture complex emotional variations.

Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have demonstrated superior performance by automatically learning hierarchical facial representations directly from image data. Architectures such as VGGNet, ResNet, and EfficientNet achieved significantly improved accuracy on challenging datasets due to their strong feature extraction capabilities and nonlinear learning mechanisms. In addition, hybrid approaches that combine CNN-based feature extraction with classifiers such as SVM or XGBoost have shown promising results in improving classification stability and robustness.

Table 4 presents a comparative overview of several representative facial emotion recognition approaches reported in the literature. The comparison indicates that deep learning architectures generally achieve higher recognition accuracy than traditional machine learning methods, particularly on complex real-world datasets. CNN-based and attention-based models demonstrate strong performance due to their ability to automatically learn hierarchical and contextual facial features. However, traditional approaches such as LBP combined with SVM still provide competitive results in controlled environments with lower computational complexity.

TABLE 4. COMPARATIVE ANALYSIS OF FER APPROACHES.

Study	Dataset	Method	Accuracy
Shan et al. (2009)	CK+	LBP + SVM	89.7%
Goodfellow et al. (2013)	FER2013	CNN	71.2%
Lopes et al. (2017)	FER2013	CNN-Based FER	75.4%
Li et al. (2018)	RAF-DB	CNN + Attention Mechanism	88.1%
He et al. (2016)	Multiple Image Datasets	ResNet	92.0%
Transformer-Based FER Studies	Multiple FER Datasets	Attention-Based Architectures	90.3%

Table 5 presents a comparative evaluation of several FER approaches based not only on recognition accuracy, but also on computational complexity, real-time capability, explainability, and robustness. Traditional approaches such as HOG + SVM provide lower computational complexity and higher interpretability, making them suitable for real-time and resource-constrained environments. In contrast, deep learning architectures such as CNNs and ResNet models achieve significantly higher recognition accuracy and robustness due to their capability to learn hierarchical facial representations automatically. Transformer-based models demonstrate very strong robustness and contextual learning capabilities; however, they generally require higher computational resources and are less suitable for real-time deployment. This comparison highlights the trade-offs between performance, computational efficiency, and interpretability across different FER methodologies.

TABLE 5. COMPARATIVE EVALUATION OF FER APPROACHES BASED ON PERFORMANCE AND PRACTICAL CHARACTERISTICS.

Method	Accuracy	Complexity	Real-Time	Explainability	Robustness
HOG + SVM	Medium	Low	High	High	Low
CNN	High	High	Medium	Low	Medium
ResNet	Very High	High	Medium	Low	High
Transformer-Based Models	Very High	Very High	Low	Medium	Very High

The comparative analysis demonstrates that deep learning models generally outperform traditional machine learning methods in complex and unconstrained environments. However, classical approaches remain valuable because of their lower computational requirements and easier interpretability. Hybrid architectures appear to provide a balanced solution by combining the strong feature extraction capability of deep neural networks with the stability of traditional classifiers.

Another important observation from the literature is that dataset characteristics strongly influence FER performance. Controlled datasets such as CK+ often produce higher recognition accuracy compared to more challenging real-world datasets such as FER2013 or RAF-DB. This indicates that real-world FER applications still require more robust and adaptive recognition models capable of handling environmental variability and demographic diversity.

Overall, comparative studies suggest that future FER systems should focus on combining accuracy, computational efficiency, interpretability, and fairness in order to support reliable deployment across diverse practical applications.

X. CONCLUSION

Facial emotion recognition (FER) has become an important research area within Artificial Intelligence and Computer Vision due to its wide range of applications in healthcare, education, intelligent surveillance, social robotics, and human–computer interaction. The continuous evolution of machine learning and deep learning techniques has significantly improved the capability of computers to automatically analyze and interpret human emotional states from facial expressions.

This review paper presented a comprehensive overview of recent developments in facial emotion recognition systems. The study analyzed publicly available datasets, traditional machine learning approaches, deep learning architectures, current research challenges, and emerging future directions in FER research. Classical machine learning methods such as Support Vector Machines, K-Nearest Neighbors, and Naive Bayes provided the initial foundation for automatic emotion recognition systems, especially when combined with handcrafted feature extraction techniques such as HOG and LBP. However, the emergence of deep learning, particularly Convolutional Neural Networks (CNNs), transformed FER research by enabling automatic hierarchical feature learning directly from image data.

The review also highlighted the importance of publicly available datasets such as FER2013, CK+, RAF-DB, and AffectNet in advancing FER methodologies. Although modern deep learning architectures have achieved remarkable performance improvements, several limitations remain unresolved. Challenges related to class imbalance, illumination variations, facial occlusions, head pose changes, computational complexity, and ethical concerns continue to affect the reliability and generalization capability of FER systems in real-world environments.

Emerging technologies such as transformer-based architectures, explainable artificial intelligence, multimodal learning, and lightweight edge-based systems are expected to further improve the accuracy, transparency, and practical applicability of emotion recognition models. In addition, future FER systems will likely focus on fairness, privacy preservation, and personalized emotional analysis to support more ethical and human-centered intelligent applications.

In conclusion, facial emotion recognition remains a rapidly evolving and interdisciplinary research field. Continued advancements in deep learning, dataset development, and explainable intelligent systems are expected to play a crucial role in the next generation of robust and reliable emotion recognition technologies.

Future FER systems are expected to move beyond simple emotion classification toward more adaptive, explainable, multimodal, and human-centered intelligent systems. The integration of transformer-based architectures, privacy-aware artificial intelligence, and real-time edge computing technologies may significantly improve the reliability and practical applicability of next-generation facial emotion recognition systems.

REFERENCES

[1] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124–129. https://doi.org/10.1037/h0030377

[2] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52

[3] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1*, 886–893. https://doi.org/10.1109/CVPR.2005.177

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105. https://doi.org/10.1145/3065386

[5] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1409.1556

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

[8] Goodfellow, I., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Bengio, Y., et al. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks, 64*, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

[9] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[10] Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861. https://doi.org/10.1109/CVPR.2017.305

[11] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[12] Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816. https://doi.org/10.1016/j.imavis.2008.08.005

[13] Happy, S. L., & Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing, 6*(1), 1–12. https://doi.org/10.1109/TAFFC.2014.2386334

[14] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning*, 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

[15] Li, Y., Zeng, J., Shan, S., & Chen, X. (2018). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing, 28*(5), 2439–2450. https://doi.org/10.1109/TIP.2018.2886767

[16] Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(6), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127

[17] Ekman, P. (2009). Lie catching and microexpressions. In *The Philosophy of Deception* (pp. 118–133). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195327939.003.0008

[18] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors, 18*(2), 401. https://doi.org/10.3390/s18020401

[19] Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition, 61*, 610–628. https://doi.org/10.1016/j.patcog.2016.07.026

[20] Zhang, K., Huang, Y., Du, Y., & Wang, L. (2018). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing, 26*(9), 4193–4203. https://doi.org/10.1109/TIP.2017.2721106

[21] Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 200–205. https://doi.org/10.1109/AFGR.1998.670949

[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.11929