information

Article

Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets

Thomas Kopalidis, Vassilios Solachidis *, Nicholas Vretos and Petros Daras

Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece; tkopalid@iti.gr (T.K.); vretos@iti.gr (N.V.); daras@iti.gr (P.D.)
* Correspondence: vsol@iti.gr

Abstract: Recent technological developments have enabled computers to identify and categorize facial expressions to determine a person's emotional state in an image or a video. This process, called "Facial Expression Recognition (FER)", has become one of the most popular research areas in computer vision. In recent times, deep FER systems have primarily concentrated on addressing two significant challenges: the problem of overfitting due to limited training data availability, and the presence of expression-unrelated variations, including illumination, head pose, image resolution, and identity bias. In this paper, a comprehensive survey is provided on deep FER, encompassing algorithms and datasets that offer insights into these intrinsic problems. Initially, this paper presents a detailed timeline showcasing the evolution of methods and datasets in deep facial expression recognition (FER). This timeline illustrates the progression and development of the techniques and data resources used in FER. Then, a comprehensive review of FER methods is introduced, including the basic principles of FER (components such as preprocessing, feature extraction and classification, and methods, etc.) from the pro-deep learning era (traditional methods using handcrafted features, i.e., SVM and HOG, etc.) to the deep learning era. Moreover, a brief introduction is provided related to the benchmark datasets (there are two categories: controlled environments (lab) and uncontrolled environments (in the wild)) used to evaluate different FER methods and a comparison of different FER models. Existing deep neural networks and related training strategies designed for FER, based on static images and dynamic image sequences, are discussed. The remaining challenges and corresponding opportunities in FER and the future directions for designing robust deep FER systems are also pinpointed.

Keywords: facial expression recognition; facial feature extraction; expression classification; deep learning; convolutional neural network; deep belief network; taxonomy; survey; database

check for updates

Citation: Kopalidis, T.; Solachidis, V.; Vretos, N.; Daras, P. Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. *Information* **2024**, *15*, 135. https://doi.org/10.3390/info15030135

Academic Editors: Nikolaos Mitianoudis and Ilias Theodorakopoulos

Received: 16 January 2024
Revised: 8 February 2024
Accepted: 12 February 2024
Published: 28 February 2024

1. Introduction

Various communication elements, such as facial expressions, body movements, and voice, can be employed to identify human emotions. Among these, facial expressions are particularly informative, as they convey information about a person's emotional state and provide insight into their mood, interest, boredom, confusion, stress levels, and conversational signals, such as speech emphasis and syntax. The FER system has generated much interest, leading to its adoption in the computer vision literature for automatic facial expression recognition. For decades, researchers have been studying facial expression recognition. Significant work has been undertaken for comprehensive surveys of past efforts in the field [1–21]. FER is critical in computer vision and machine learning because it investigates ways to analyze and recognize facial muscle movements. The reasons for this renewed interest in facial expressions are numerous in many fields, such as neuroscience [22], biomedical engineering, healthcare, crime detection [23], public safety, education, employment, customer behavior analysis and advertising, sociable robotics, medical treatment, driver fatigue surveillance, and fraud, among others. However, the advances in related research areas such as face detection, face tracking, and face recognition, the current availability of relatively inexpensive processing power, and the rise in deep learning [24] methodologies are of great importance. Furthermore, facial expression recognition is well researched, particularly in the fields of computer vision (CV) and artificial intelligence (AI), with applications in human–computer interaction (HCI) [25,26], virtual reality (VR) [27] and augmented reality technologies [28], advanced driver assistant systems (ADASs) [29], and entertainment [30]. According to a report by Allied Market Research [31], the global emotion recognition market is rapidly growing, valued at $21.7 billion in 2021 and projected to reach $136.2 billion by 2031, with a CAGR of 20.5% from 2022 to 2031, as reported by Allied Market Research.

Facial expression recognition (FER) is a method of detecting human emotions from static images or videos to disclose information about humans' facial expressions or emotional states. In [18], Ekman, and in [32,33], Friesen postulated six primary emotions and discussed their distinct characteristics. People of all races and civilizations share these prototypical emotional displays, often referred to as basic emotions. Thus, to identify human emotions, the key lies in facial expression. In reality, not all facial expressions have significance and can be mapped to emotions, but there are six fundamental emotions (in Figure 1) that are universal [34] and can be represented in the same manner.

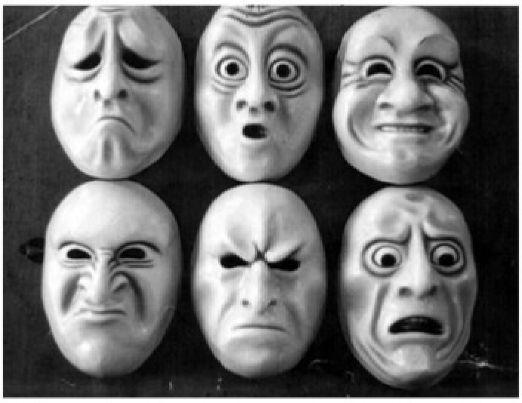


Figure 1. Six basic emotions.

Anger, fear, happiness, disgust, surprise, and sadness are the basic emotions, and one more emotion was added [35], the universal expression of contempt. So, in total, there are seven emotional states. There is evidence for universality in the following seven emotions [36].

Moreover, in facial expression recognition, we can also measure valence and arousal circumplex with the horizontal and vertical axes (Figure 2) describing valence and arousal, respectively (**arousal** is defined as how inclined this individual is to behave based on their emotional state, whereas **valence** is how pleasant or terrible a sensation is [37]). We can see that neutral is in the middle of the axes, as one would assume, because this emotional state has zero arousal and valence.

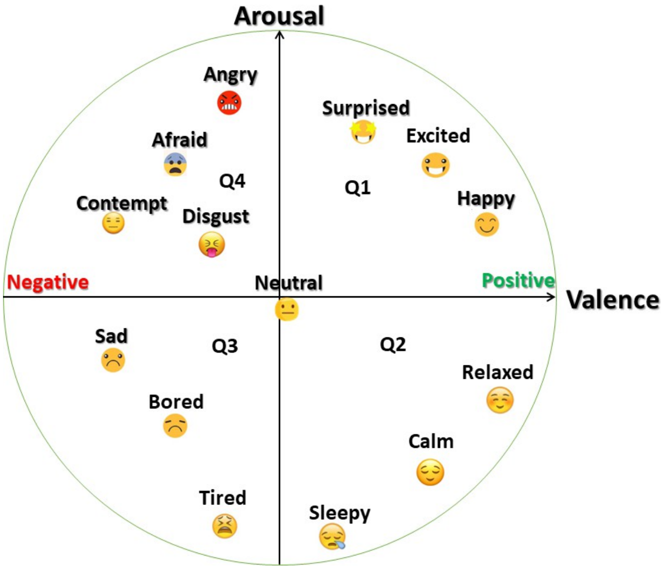


Figure 2. The arousal and valence factors vary from positive to negative. Also, the positive states are in the right hemisphere, and the negative ones are in the left. Emotional states can be easily defined in 4 different quadrants (Q1–Q4), including high arousal–high valence (Q1), high arousal–low valence (Q2), low arousal–low valence (Q3), and low arousal–high valence (Q4), respectively [38].

Our paper's notable contributions are summarized as follows:

- **Taxonomy:** A brand new classification of FER is proposed, as well as for datasets. This taxonomy not only references traditional methodologies, but also incorporates recent advancements in deep learning. Specifically, it integrates techniques such as Generative Adversarial Networks (GANs), graph-based methods, and transformers in the context of FER. Additionally, there is a distinct taxonomy for datasets: the datasets are categorized into images and sequences. These are further divided into controlled and uncontrolled groups. The controlled group includes categories such as movies and lab settings, while the uncontrolled group pertains to the "in the wild" category. To the best of our knowledge, no existing research has addressed this comprehensive classification.
- **Comprehensive review:** This paper contains cutting-edge results from newcomers' research that have not been examined in earlier survey publications [37,39–70].
- **Highlighting top models:** Evaluation results from the most significant methods across different datasets are presented, covering approximately 60 methods.
- **Overview of popular datasets:** A total of 21 of the most commonly used datasets are introduced [71–103].
- **Suggestions and future directions:** Apart from the conclusion, possible future research directions for FER are suggested.

Survey papers will be listed below, detailing both their coverage and their limitations. The papers by Pantic et al. [1,3], Azizan et al. [10], Mehta et al. [11], and Revina et al. [15] provide foundational insights into Facial Expression Recognition (FER), focusing on state-of-the-art techniques but lacking in-depth analyses of advanced deep learning methodologies and dataset evolution. Fasel et al. [2], Tian et al. [4], and Shan et al. [5] discuss facial expression and action recognition from static images, emphasizing the challenges of their time but missing integration with contemporary models and datasets. The studies by Wei et al. [12], Mellouk et al. [13], and Bettadapura et al. [6] review deep learning in FER, notably in industrial electronics, yet overlook a range of deep learning models and comprehensive dataset analyses.

Further, Konar et al. [7], Zhao et al. [8], and Patel et al. [14] offer insights into emotion recognition and feature extraction, providing comprehensive overviews but lacking broad methodological scope and future trend discussions. The surveys by Dang et al. [16], Jampour et al. [17], and Khan et al. [19] address general facial expression recognition and its applications but fall short in analyzing FER evolution and integrating AI techniques comprehensively. Lastly, the works by Canal et al. [18] and Cai et al. [21] cover FER's modality and methodologies, offering a broad view on multimodal aspects but lacking focused discussions on tailored AI techniques for FER, detailed model comparisons, and extensive FER evolution timelines.

In Figure 3, the taxonomy of the paper methods is illustrated. Initially, the evolution timeline is presented; after that, the traditional methods, the deep learning methods, the state-of-the-art core of this paper, and a detailed analysis of datasets in FER are shown.

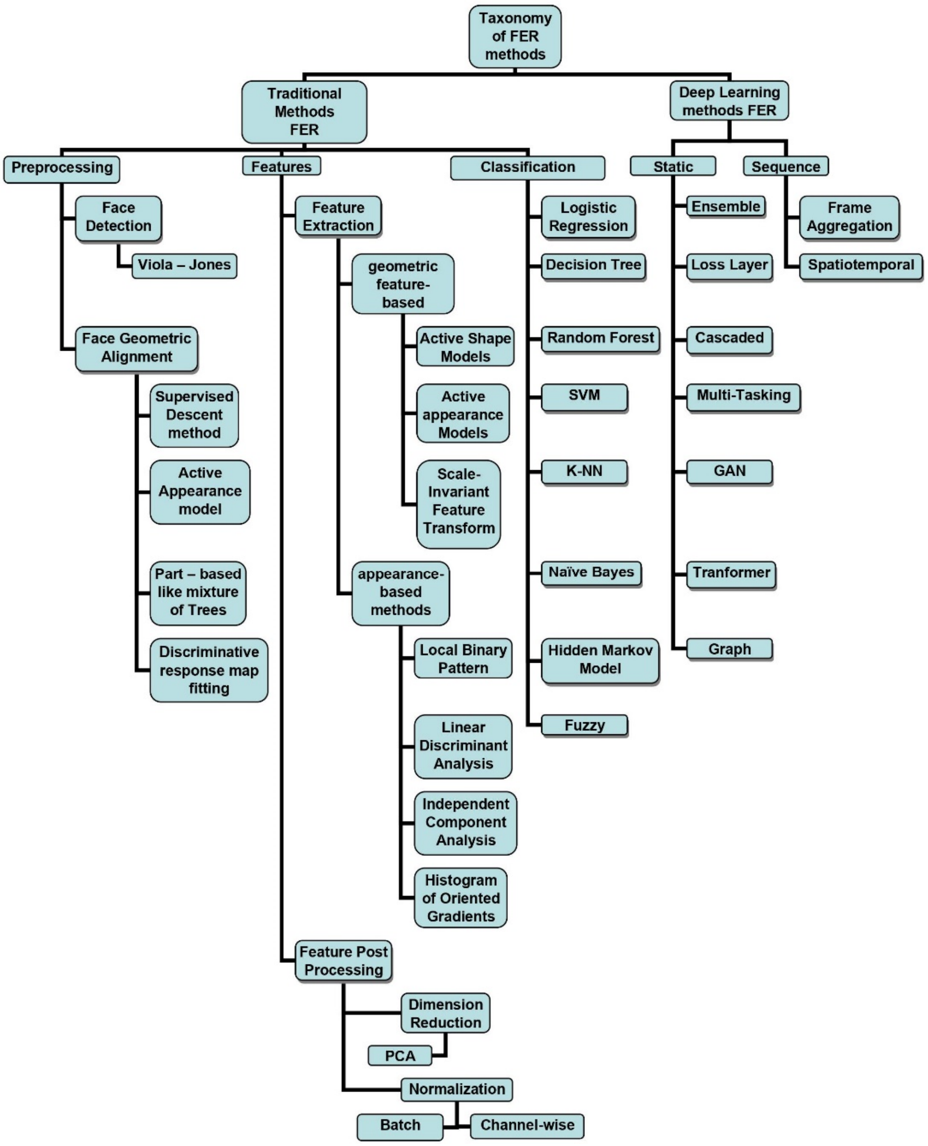


Figure 3. Taxonomy of the FER methods.

Organization of survey: The rest of the paper is organized as follows. ***Section* 2** focuses on the evolution of facial recognition techniques and the most used datasets. ***Section* 3**, discussing the pro-deep learning era, elaborates on the facial expression recognition system, describing its components, including preprocessing, feature extraction, and classification. It also presents relevant background information, supported by examples, to comprehensively understand the topic. ***Section* 4**, the deep learning era section, includes a brief history of deep learning methods such as CNN, RNN, and DBN, etc., with some indicative works on each method. Also, the state-of-the-art is mentioned by many newcomers and novel papers.

In ***Section 5***, an introduction to the most used emotions datasets is provided. Moreover, the nomenclature is provided in this section. ***Section 6*** provides suggestions for future research and concludes this survey.

2. Evolution Timeline

This evolution timeline was made to illustrate the evolution of FER methods and the evolution of the datasets. This timeline encompasses traditional machine learning methods and deep learning methods. Regarding the datasets, they are categorized into four main types. First, the datasets are classified based on their content, distinguishing between images and videos (sequences). Second, they are categorized based on the environment, with distinctions made between controlled environments (lab) and uncontrolled environments (in the wild). So, the categories are IC (image controlled), IU (image uncontrolled), SC (sequence controlled), and SU (sequence uncontrolled). By understanding the timeline of developments in facial expression recognition and the development of datasets, it is possible to provide the reader with valuable insights into the progress and challenges facing this field (FER). This background information lays the foundation for a comprehensive understanding of the FER system.

Facial emotion recognition was initially presented in work proposed by Bassili [104] in 1978. Bassili et al. investigated the influence of facial motion on perceiving faces and emotional expressions. They explored how dynamic changes in facial features, like eye and mouth movements, impact the recognition and interpretation of emotions. The study offers insights into the role of motion in understanding human emotions (six fundamental emotions: happiness, sadness, fear, surprise, anger, and disgust). In Figure 4, the evolution timeline of facial expression recognition methods is illustrated. Numerous FER algorithms have been developed, encompassing both traditional methods and deep learning (DL) algorithms, with many state-of-the-art methods in both categories. Moreover, in Figure 5, a brief overview of the evolution timeline for the datasets is given, in which the terms IC (image controlled), IU (image uncontrolled), SC (sequence controlled), and SU (sequence uncontrolled) refer to different types of datasets commonly used in the context of FER. To clarify each dataset, we will label them as IC, IU, SC, and SU.

Several datasets have been created to train and evaluate FER models. At first, **laboratory-controlled environment (lab)** datasets were created, i.e., Multi-PIE [96], including facial images from various angles and lighting conditions. However, laboratory-controlled datasets have two issues: the fact that there is a lot of repetition and the dataset size. The first issue is that there is much repetition. A basic FER model can perform well in testing in a lab-controlled setting. Still, it cannot deliver a satisfactory result in actual application scenarios, which have noise in many different ways, i.e., variability in input data, background noise, uncertain or ambiguous inputs, and adaptation to dynamic environments. With real-world data, variations arise from distribution, quality, format, and representation differences. Background noise or environmental factors cannot be controlled in the lab. The model must handle incomplete or conflicting data, requiring reasoning and contextual understanding. Real-world scenarios involve dynamic situations, necessitating real-time adaptation to changing conditions, contexts, or user behaviors.

Overall, these “many different ways” of noise and challenges in real-world scenarios demand robustness, adaptability, and the ability to handle uncertainty from AI models, which may not be adequately addressed during lab-controlled testing phases.

The second issue is the dataset size, which cannot reach the prerequisites for deep learning techniques in a laboratory-controlled setting, preventing the successful application of deep learning’s strong feature learning capabilities to FER systems, and, of course, the size of the laboratory datasets is small. To address the two problems outlined above, large FER datasets of hundreds of thousands of diverse images had to be created by gathering and labeling them from the Internet (in the wild). Examples of FER datasets in the wild are AffectNet [83], RAF-DB [94], FER2013 [86], and FER+ [71], etc. As mentioned before, based on content, FER datasets are also categorized into images and sequences (movie). Sequences (or movie) are collections of video clips or frames extracted from movies, TV shows, or other media sources used to train and test deep-learning-based FER models. These datasets typically contain many examples of facial expressions and cover many emotions, making them ideal for training and evaluating FER models. Two well-known examples of this category are AFEW 7.0 [80–82] and SFEW 2.0 [84,85]. The timeline (Figure 5) displays a list of datasets sorted by their year of creation. Next to each dataset name is an acronym (IC, UC, IS, and US) indicating the category to which the dataset belongs.

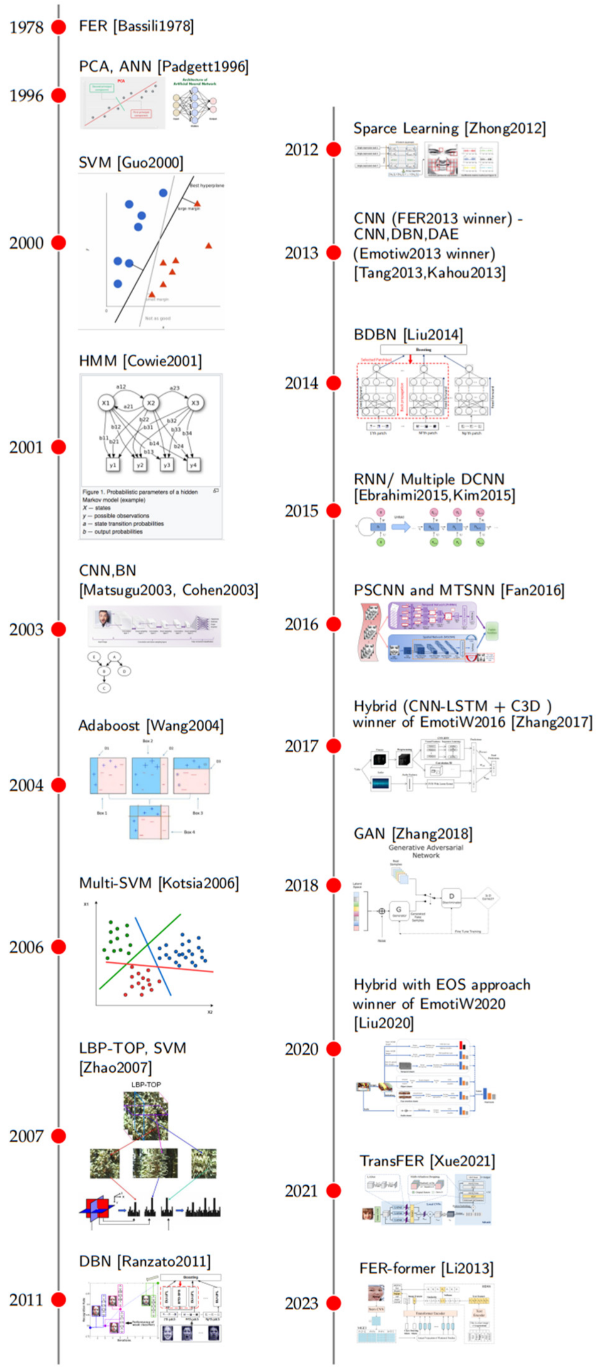


Figure 4. Evolution timeline of methods. List of papers in the image [Bassili1978] [104], [Padgett1996] [105], [Guo2000] [106], [Cowie2001] [25], [Matsugu2003, Cohen2003] [107,108], [Wang2004] [109], [Kotsia2006] [110], [Zhao2007] [111], [Ranzato2011] [112], [Zhong2012] [113], [Tang2013,Kahou2013] [114,115], [Liu2014] [116], [Ebrahimi2015,Kim2015] [117,118], [Fan2016] [119], [Zhang2017] [120], [Zhang2018] [121], [Liu2020] [122], [Xue2021] [64], [Li2023] [62].

3. FER: Facial Expression Recognition

In this section, in the first part, there will be an analysis of the pro-deep learning era in facial expression recognition, describing its components extensively, including preprocessing, feature extraction, and classification. It also presents relevant background information supported by examples to comprehensively understand the topic. Moreover, this is followed by a detailed report on the deep-learning era (data augmentation, a brief history of deep learning, analysis of models [CNN, RNN LSTM, DBN, Autoencoder, GAN, Hybrid, Graphs, and Transformer] with a lot of theory, models, examples, and applications). The next part of this section is the cornerstone of this paper, discussing the state-of-the-art methods with the categorization of static and sequence datasets; then, for each category, there are many more subcategories.

3.1. Pro-Deep Learning Era

In general, FER systems before the advent of deep learning consisted of three components: preprocessing for endeavors such as face detection [123,124], feature extraction [8,125], and expression classification [8]. In the FER approach, the process begins by segmenting an image into two classes: regions containing faces and regions containing non-face elements. Additionally, facial components such as eyes and nose or landmarks are detected within the identified face regions to refine the analysis further. During feature extraction, various temporal and spatial features are extracted from the facial components. Essentially, feature extraction aims to capture facial features related to expressions. After the feature extraction stage, the extracted features are passed to the classification step. In this step, various facial expression classifiers, such as Support Vector Machines (SVM), AdaBoost, Random Forest (RF), or the SoftMax loss layer [126], are utilized. These classifiers leverage the extracted features to predict the facial expression category to which the analyzed face belongs.

FER algorithms are classified into two types based on their input: static images and dynamic sequences. On the one hand, static FER involves the face point position information from a single image's feature representation. On the other hand, dynamic image FER or sequence FER include temporal information with continuous frames [127]. The sole difference between the single static and dynamic sequence facial expression detection tasks is that several face photos must be considered during the feature extraction step in the second phase. In the following paragraphs, we will examine and discuss the three phases of data preprocessing.

3.1.1. Preprocessing

Preprocessing [128] is usually used in the FER system before the feature extraction stage and after the image input. Detecting the face is usually the initial step in preprocessing. Preprocessing is used for face detection and face alignment. It should be noted that the preprocessing stage is divided into two main processes:

Face Detection

Face detection (or face localization) is fundamental in computer vision and pattern recognition. It involves automatically locating human faces in images or video frames. Essentially, the face detection technique returns the bounding box coordinates over the face. Face detection algorithms are crucial in various applications, such as facial recognition, emotion analysis, and biometrics. These algorithms utilize many techniques, such as Haar-like features, deep learning architectures like convolutional neural networks (CNNs), and cascaded classifiers to accurately detect faces in diverse settings and handle variations in pose, illumination, occlusions, and facial expressions. Ongoing research in face detection continues to improve these algorithms' accuracy, robustness, and efficiency, advancing the field and enabling the development of innovative face-related applications.

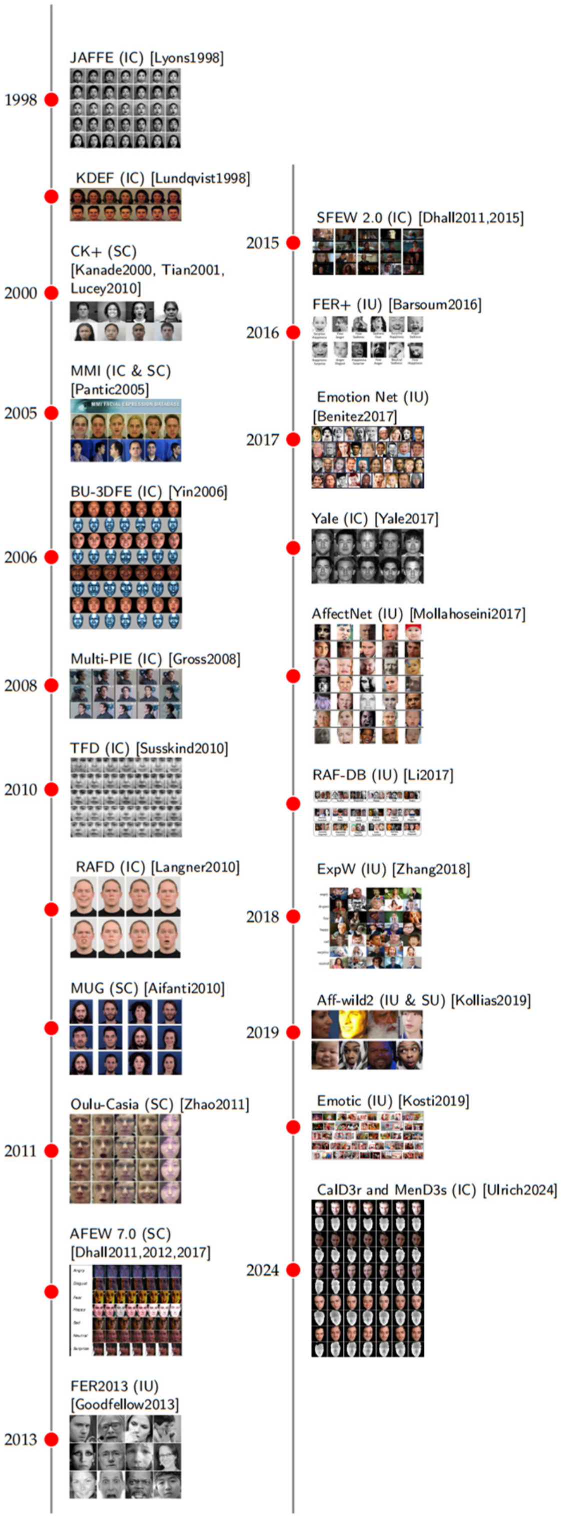


Figure 5. Evolution timeline of datasets (I = image, S = sequence, U = uncontrolled, C = controlled, L = lab). List of papers in the image [Lyons1998] [72], [Lundqvist1998] [90,91], [Kanade2000, Tian2001, Lucey2010] [76–78], [Pantic2005] [97], [Yin2006] [74,75], [Gross2008] [96], [Susskind2010] [87], [Langner2010] [89], [Aifanti2010] [99,100], [Zhao2011] [88], Dhall2011,2012,2017] [80–82], [Goodfellow2013] [86], [Dhall2011,2015] [84,85], [Barsoum2016] [71], [Benitez2017] [93], [Yale2017] [79], [Mollahoseini2017] [83], [Li2017] [94], [Zhang2018] [95], [kollias2019] [103], [Kosti2019] [101,102], [Ulrich2024] [129].

The challenges that face detection has to solve are:

- **Occlusion** [130,131]. Occlusion is when an object or part of it is blocked from view by another object, creating challenges in computer vision tasks like object detection and tracking.
- **Illumination** [132,133]. Illumination refers to lighting conditions affecting object appearance in computer vision. Handling illumination variations is crucial for accurately performing object recognition and image segmentation tasks.
- **Image resolution** [134,135]. Image resolution refers to the level of detail in a digital image. A higher resolution means more pixels, resulting in sharper images, while a lower resolution leads to less detail and potential blurring.
- **Facial expression changes** [136,137]. Facial expression changes refer to variations in facial features that convey different emotions. Analyzing these changes is crucial for understanding human emotions, involving detecting and interpreting facial landmarks and texture patterns.

Face detection is the necessary first step for all facial analysis algorithms (especially the traditional ones), including face alignment, face recognition, face verification, and face parsing. Also, face detection is used in multiple areas, such as content-based image retrieval, video coding, video conferencing, crowd video surveillance, and intelligent human–computer interfaces.

The landscape of face detection brims with a plethora of techniques, each specialized for distinct facets of the process, from initial processing to precise recognition [138]. **Feature-based approaches** are fundamental, with the **Viola–Jones** [139] algorithm being a cornerstone for real-time detection due to its speed and efficiency. While it is a staple in preliminary steps, its precision can wane in diverse facial conditions.

Another category is **Active Shape Models,** which includes **Snakes** [140], **Deformable Template Matching** [141], and **Point Distribution Models** that are also pivotal, providing adaptability against variations in pose and expressions. They are detailed in capturing the structural nuances of faces, adjusting to different shapes within images. Also, **Low-Level Analysis**, which includes assessing **motion**, **color**, **gray**scale information, and **edges**, plays a critical role in differentiating facial features from backgrounds, enabling face detection across varied scenarios.

Finally, **Feature Analysis techniques**, which involve **Feature Searching**, utilize both texture and relational data. The **Gabor Feature** [142] and **Local Binary Patterns (LBP)** [143], supported by **AdaBoost** [109], are particularly effective for their textural and pattern recognition capabilities, crucial for both detecting and recognizing facial features.

In the critical phase of preprocessing, the role of methods like Viola–Jones is indispensable. They efficiently locate faces, laying the groundwork for subsequent stages such as feature extraction, normalization, or recognition. The choice of a preprocessing method is dictated by specific application needs, balancing speed, accuracy, and robustness against various conditions. Viola–Jones, with its harmonious blend of speed and effectiveness, is often the preferred choice in many real-time applications.

Each brings unique strengths to the table, collectively advancing capabilities in face detection. Notably, Viola–Jones is celebrated as the most utilized method for preprocessing, marking its prominence in the domain of face detection methodologies.

Below will be an analysis of face detection methods for preprocessing. The most used method for preprocessing is Viola–Jones.

Viola–Jones

The Viola–Jones algorithm [139] was proposed by Paul Viola and Michael Jones and is one of the most extensively used face detection algorithms, even used by Apple in an improved version on their phone [144]. It is based on using Haar-like features and AdaBoost. Haar-like features are simple rectangular patterns used in the Viola–Jones algorithm for object detection, such as face detection. They capture local contrast variations and are evaluated to distinguish between positive (object) and negative (non-object) examples during training [145]. The Viola–Jones algorithm aims to detect faces from pictures by examining grayscale subregions for specific attributes. It requires a full-frontal view of the face and checks multiple positions and scales to detect faces of various sizes. Although the training period is long, the time required to run the facial detection process is small. The Viola–Jones algorithm has four main steps:

1. Selection of Haar-like features [145].
2. Creation of an integral image (an integral image is computed to enable the fast calculation of Haar-like features over different image regions.)
3. AdaBoost (AdaBoost is a machine learning algorithm that combines multiple weak classifiers to form a strong classifier) [109] training.
4. Creation of classifier cascades.

Face Geometric Alignment

Face alignment [146] is a common preprocessing step in various face detection applications. It minimizes pose, scale, and rotation variations, enabling accurate analysis. This step is essential because it mitigates the variation in face scale and in-plane rotation. Several types of face alignment methods exist, including holistic approaches like Active Appearance Model (AAM) [147], Part-based like mixtures of trees (MoT), Discriminative response map fitting (DRMF) [148], Cascaded regression such as the Supervised Descent Method (SDM) [149], Local binary features [150], and incremental [148] and deep learning such as MTCN [151], etc. Due to its high speed and accuracy, the face alignment method in cascaded regression has become the most common among machine learning methods and cutting-edge approaches for face alignment [152]. Some methods have to combine multiple detectors for better landmark estimation and perhaps for better performance [153].

3.1.2. Feature Extraction

Feature Extraction [8,78,125] is the process of taking an input image, quantifying it according to some algorithm (called a feature extractor or image descriptor), and returning a vector (i.e., a list of numbers) that quantifies the contents of the image. It contains two different components: feature extraction and dimensionality reduction.

Feature Extraction

Feature extraction aims to identify and capture relevant facial features for expression recognition. It is crucial in various applications such as diagnosis, classification, and clustering. According to different types of input images (static or dynamic), there are, respectively, also two types of feature extraction.

For static images, there are two types:

- geometric-feature-based methods
- appearance-based methods

The forms and placements of face components, such as the brows, eyes, nose, mouth, and chin, are described by geometric facial characteristics. Geometric-feature-based approaches seek to extract face characteristics by utilizing the geometric correlations between facial feature locations. On the other hand, appearance-based methods use the whole face or specific regions in a face image to reflect the information in a face image. Some techniques of geometric-feature-based methods are Active Shape Models (ASM) [154], Active appearance Models (AAM) [147], and Scale-Invariant Feature Transform (SIFT) [155,156]. For appearance-based methods, most well-known techniques are LBP (Local Binary Pattern) [143], Weighted Projection-based LBP (WPLBP) [157], Gabor wavelet [143], Histogram of Oriented Gradients (HOG) [158,159], Principal Component Analysis (PCA) [160–162], Facial Action Code (FAC) [163], Linear Discriminant Analysis (LDA) [164], Independent Component Analysis (ICA) [165,166], Supervised Descent Method (SDM) [149], Three Orthogonal Planes (LBP-TOP) [167], and Pyramid Histogram of Visual Words (PHOW) [168], etc. For feature extraction, the Gabor filter, WPLBP, SDM, WLD, and HOG are the most commonly used, but most traditional methods are based on LBP, which gives improved results.

For dynamic images, sequences are represented by deformation and facial muscle movements. There are also two types:

- Optical flow [169]
- Feature point tracking [170]

Optical flow

Optical flow [171] is a technique that characterizes changes in geometry and radiometry within dynamic images. In facial expression recognition (FER), optical flow is utilized to extract expression-related features and calculate pixel-level motion between two frames, typically from a neutral to a peak expression. By tracking features across frames, the optical flow returns a vector describing the movement of pixels from the first to the second frame.

However, this approach is sensitive to noise and occlusions, and its effectiveness relies on the initial selection and tracking of features.

Feature point tracking

Feature point tracking algorithms [170] select specific feature points, often at the corners of the eyes and mouth, to capture deformation information [170]. The objective is to monitor the movement of these chosen feature points among the 15 features based on the well-established facial action coding system [172]. As illustrated in Figure 6, some of the most widely used feature extraction methods are depicted.

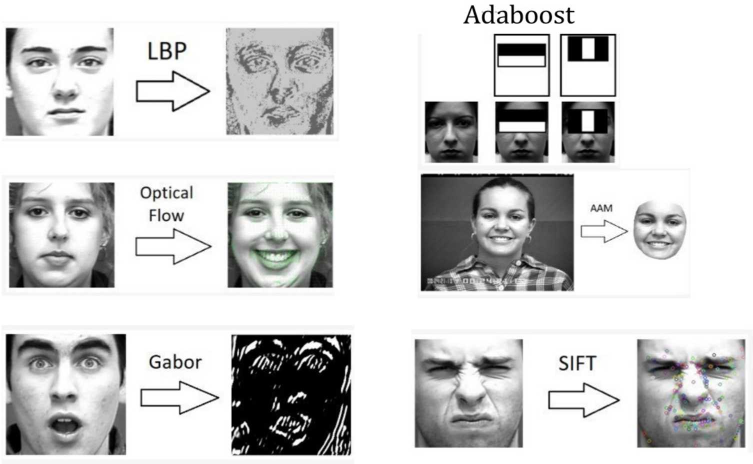


Figure 6. Some feature extraction methods from top to bottom and from the right to the left LBP, Adaboost, Optical Flow, AAM, Gabor, and SIFT [173].

Dimensionality Reduction

Dimensionality reduction is a machine learning (ML) or statistical technique for reducing the number of random variables in a problem by taking a set of principal variables (or reducing the dimensionality of features). These techniques aim to represent the images in a lower-dimensional space while preserving the important information. An example of a dimension reduction method used in image analysis includes Principal Component Analysis (PCA). This technique can help in reducing computational complexity, extracting meaningful features, and improving the efficiency and performance of deep learning models. Certainly, when the features of an image are transformed from an m-dimensional space to a lower-dimensional space with k dimensions: $m > k$, it does not necessarily suggest a deterioration in the image quality. The transformation of image features helps to maintain or even enhance the overall quality of the image. There are pros and cons to applying dimensionality reduction. Some benefits of applying the dimensionality [174] reduction technique to the given dataset are given below:

- The space that is required to store the dataset is reduced.
- Less computation training time is required for reduced dimensions of features.
- Reduced dimensions of features also help to faster visualize the data.
- It removes the redundant features by taking care of multicollinearity.

Below will be an analysis of the Principal Component Analysis, a well-known dimensionality reduction method.

Principal Component Analysis (PCA)

Principal Component Analysis [175] is a well-known facial detection and recognition technique. One widespread implementation of PCA is eigenfaces, which was introduced by Turk and Pentland [176]. Vretos et al. [177] mapped vertices to a new coordinate system using eigenvectors and barycenter, then used SVMs for facial expression recognition with satisfactory results. These methods improve facial detection, recognition accuracy, and robustness.

3.1.3. Classification

The classification stage [8] is a crucial component in the FER system, where the classifier is trained on known datasets to recognize facial expressions in unknown datasets during testing. Although classical approaches for FER have been successful in achieving a high accuracy, they come with high processing costs and are still widely used. The goal of the classifier is to accurately categorize expressions such as happy, sad, surprised, angry, fearful, disgusted, and neutral into their respective labels. There are many classical methods employed in the classification stage, including Linear Discriminant Analysis (LDA) [178], Decision Trees (DT) [179], K-Nearest Neighbor (KNN) [180], Fuzzy [181], Adaboost [109], Hidden Markov Model (HMM) [182], Bayesian Network (BN) [108], Sparse Representation-based Classification (SRC) [113], and Support Vector Machines (SVM) [183,184], among others. An overview of the traditional facial expression recognition stages is depicted in Figure 7.

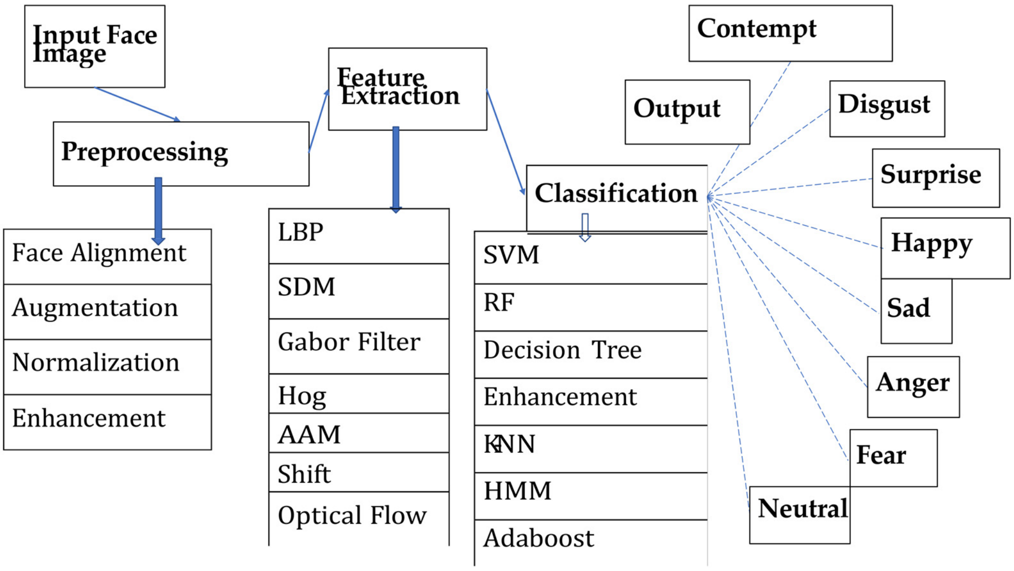


Figure 7. Traditional facial expression recognition stages.

3.2. Deep Learning Era

Nowadays, deep learning (DL) [185] has become one of the hottest research topics in the fields of machine learning (ML) and artificial intelligence (AI), and is considered to be the core technology of the Fourth Industrial Revolution. Also, deep learning's use finds applications primarily in image recognition, natural language processing (NLP) [186], speech recognition software [187], genome engineering, and systems biology [188]. The reasons that more and more people tend to use deep learning models instead of traditional ones are:

(1) Hardware advancements, primarily GPUs, accelerate deep learning. (2) Deep learning can learn from raw data and eliminate the need for manual feature engineering; thus, the models are faster and can capture complex patterns effectively. (3) Moreover, deep learning models learn hierarchical representations of data, capturing intricate relationships and improving performance in tasks like image recognition, natural language processing, and speech recognition. Traditional methods often need help to capture high-level features in these domains. (4) Due to their inherent scalability, deep learning models can handle large-scale datasets and complex problems. Deep learning models can be trained efficiently on massive amounts of data, improving performance. (5) Also, deep learning models are capable of end-to-end learning, meaning they can learn directly from input to output without relying on intermediate steps. This simplifies the overall pipeline. (6) Deep learning models have shown remarkable adaptability across various domains. They have achieved state-of-the-art results in many fields, such as computer vision. (7) Finally, deep learning has witnessed significant advancements over the years, with continuous research and development. New architectures, algorithms, and optimization techniques are constantly being introduced, improving performance and pushing the boundaries of what is possible in different applications.

In addition to end-to-end learning, another alternative is to utilize a CNN or other deep neural network as a feature extraction technique before employing additional independent classifiers, such as a support vector machine (SVM) [189] or decision tree. To enhance the performance of deep learning models, it is crucial to apply data augmentation techniques in comparison to conventional methods.

Data Augmentation [190]

Generally, most publicly accessible FER datasets lack sufficient photos for training. As a result, data augmentation is a critical stage in deep FER, and the primary task is to increase the size of a training dataset. Data augmentation artificially increases the amount of data by generating new data points from the existing data. Furthermore, data augmentation is another significant preprocessing aspect used too much in deep learning.

There are two types of data augmentation techniques: online data augmentation and offline data augmentation. Online augmentation in the training data loader is an excellent way to enlarge the variation in the dataset. However, the augmented data are randomly generated in different ways, i.e., with the GAN model, and the data loader follows when sampling the data. A model may need to be trained for a long time to achieve a high accuracy. Offline augmentation can prevent this and generate a dataset with the required augmentations. Offline augmentation can expand the dataset size when collecting and labeling data is costly or impossible. The most common augmentation (offline augmentation) methods are rotation, flipping, saturation, translation, scaling, cropping, brightness, color augmentation, and contrast, etc. For deep learning using CNNs to produce accurate classification results, many training data are always required. Overfitting (overfitting represents an undesirable behavior in machine learning where a model achieves accurate predictions on its training data but falters when faced with new, unseen data. The model becomes prone to overfitting when the training process spans numerous epochs, and the network's capacity is substantial. This translates to the model performing well on the training set while failing to generalize effectively, leading to a small training error but an extensive validation or test error) becomes a significant issue since our database is so limited. Expanding the database with artificial label-preserving modifications reduces overfitting most frequently. Therefore, we add to the database using various modifications to produce numerous modest changes in appearance and orientation before training the CNN model. The network is more resilient to various scenarios using straightforward data augmentation techniques. It has been observed that DL has attained cutting-edge performances in a range of applications [191]. This section is a brief introduction to some problems in the application of FER, also, this is followed by a brief history on the deep learning methods that have been applied in FER and then the deep learning techniques.

3.2.1. A Brief History

While the deep learning era is often associated with the McCulloch–Pitts (MCP) model [192] as the initial artificial neural model, it is essential to note that the term "deep learning" itself emerged much later to refer to the training of neural networks with multiple layers. The origins of deep learning can be traced back to the early development of artificial neural networks, including the perceptron in the 1950s and the backpropagation algorithm in the 1980s. Regarding their application in Facial Expression Recognition (FER), deep learning techniques began to garner attention in the late 2000s and early 2010s. While the first application of FER [14] with artificial neural networks (ANN) by Padgett and Cottrell [193] in 1996 laid some groundwork, it was not until subsequent years when deep learning methods displayed substantial enhancements in FER performance. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated impressive feature learning abilities in FER tasks. These models can automatically learn hierarchical representations from raw image or video data, enabling them to capture intricate facial features and patterns associated with different expressions. Despite the remarkable feature learning ability of deep learning, there are several challenges in applying it to Facial Expression Recognition (FER). Firstly, crafting FER systems based on deep neural networks (DNNs (deep neural networks are artificial neural networks with multiple hidden layers between the input and output layers. They learn hierarchical representations of the input data by progressively extracting higher-level features through interconnected artificial neurons. The weights between neurons are adjusted during training to minimize the difference between predicted and desired outputs)) encounters significant constraints due to their high data dependence. This is exacerbated by the fact that FER datasets are often small and tailored to specific tasks [152]. Given this, overfitting in results frequently occurs when applying training deep models on FER datasets. Before 2010, Facial Expression Recognition (FER) datasets were relatively small. Over time, these datasets gradually increased in scale

To put it simply, earlier research indicated that achieving end-to-end training in Facial Expression Recognition (FER) posed challenges when employing deep architectures on minimally preprocessed images. However, previous studies have showcased that incorporating task-specific data for pre-training or fine-tuning existing models can notably enhance the development of more sophisticated FER models. The second challenge was the requirement for facial expression datasets at that juncture. This posed difficulties in training neural networks with deep architectures, which are known for their prowess in achieving exceptional results in recognition tasks. A third explanation for substantial inter-subject variance is due to personal characteristics such as age, gender, ethnic origin, and level of expressiveness. Variations in position, lighting, and occlusions typical in unrestrained facial expressions may also create issues, so we trained models on corresponding datasets that encompassed these variations. Below, we will refer to the solutions to these problems.

(a) **Data augmentation** is an option for preventing overfitting in DL models.
(b) As time progressed, larger and more diverse datasets collected from real-world scenarios emerged. Notable examples include AffectNet [83], ExpW [95], EmotionNet [92,93], and RAF-DB [94,194,195]. These datasets enabled the development of Facial Expression Recognition (FER) methods with enhanced performances and promising outcomes in recognition tasks.
(c) To enhance facial expression recognition, an ideal dataset with diverse images of expressions, age, gender, and ethnicity can be utilized, along with multitasking deep networks and transfer learning techniques.
(d) **Preprocessing** [196] is pivotal in managing diverse data variations. Here are some reasons why preprocessing is advantageous: **(i) Standardization**: Preprocessing techniques like normalization or standardization ensure that features with different scales

or units are brought to a typical range, preventing certain features from dominating the learning process. **(ii) Noise reduction**: Filtering or denoising techniques applied during preprocessing minimize the impact of noisy data, improving the model's ability to extract meaningful features. **(iii) Feature extraction** (referred to above). **(iv) Dimensionality reduction** (referred to above). **(v) Handling missing values**: Preprocessing techniques, such as imputation or exclusion, address missing data, ensuring that the model is trained on complete and consistent data, resulting in reduced bias and improved accuracy and reliability of predictions.

As we mentioned, the era of deep learning era in Facial Expression Recognition (FER) began in 1996 with the introduction of the Artificial Neural Network (ANN) by Padgett and Cottrell [193]. In **2003, Matsugu** [107] employed the Convolutional Neural Network (CNN) in FER, while the architecture for LeNet-5 was proposed in 1998 by LeCun [197]. Furthermore, in 2011, Ranzato [112] introduced the Deep Belief Network (DBN) to the field of FER, although DBN was initially presented in 2006 by Hinton [198]. In 2014, a variant of DBN, the Boosted Deep Belief Network (Boosted DBN), was introduced by Liu [199].

In 2015, a **Recurrent Neural Network (RNN)**, originally created by Ruhelmart [200] in 1986, was utilized in FER by **Kahou** [117]. Moreover, in 2017, **Zhang** [120] in **2017** combined **Part-Based Hierarchical Bidirectional RNN (PHRNN)** and **Multi-Signal CNN (MSCNN)**. It is worth noting that many recent models in FER are hybrids, combining different models. It has been observed that more and more models are combinations or a hybrid network, i.e., **CNN-RNN and 3D convolutional networks (C3D),** which was introduced in **2016** by **Fan [119]**.

Additionally, the Deep Autoencoder (DAE), consisting of two DBNs, was introduced in 2006 by **Hinton et al. [201],** and in FER, the **Sparse Autoencoder Network (DSAE)** was introduced in **2018** by **Zeng [202]**. Lastly, **Generative Adversarial Networks (GAN)** were first introduced by Goodfellow et al. [203] in 2014 and have been widely used in the FER domain, including by Zhang in 2018 [95].

Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) [197,204–206] are specialized neural networks designed for processing data with a grid-like topology, such as time-series data arranged in a 1D grid, image data in a 2D grid of pixels, and volumetric data in a 3D grid [207].

CNNs have emerged as the mainstream approach in deep learning techniques for various tasks. CNNs are potent tools for image recognition tasks as they learn hierarchical representations of images. They have been successfully applied in a multitude of computer vision applications, including face detection, facial expression recognition, object detection, self-driving or autonomous cars, auto-translation, text prediction, handwritten character recognition, climate analysis, X-ray image analysis, cancer detection, visual question answering, image captioning, biometric authentication, document classification, and 3D medical image segmentation [204]. These applications demonstrate the versatility and effectiveness of CNNs in solving various visual recognition problems.

Their robustness to various changes like face location and scale variations makes them exceptionally effective [208]. They outperform multilayer perceptron (MLP) models in handling these variations due to their shift-invariance, convolutional operations, and pooling layers. This robustness allows CNNs to capture and extract features regardless of facial location or scale changes.

A typical CNN architecture consists of three main types of layers: (1.) **Convolutional Layer**, (2.) **Pooling Layer**, and (3.) **Fully Connected Layer**. The first layer of a CNN is typically a **convolutional layer**, which applies a set of learnable filters to the input image, generating activation maps that highlight certain features. As the data progress through the layers of the CNN, the network identifies increasingly complex patterns, such as facial features, shapes, and, ultimately, the full object.

CNNs can have multiple convolutional layers, allowing the network to build hierarchical representations of the input image. The depth hyperparameter determines the number of neurons in a layer connected to the same input area, with deeper networks being more complex. The stride hyperparameter specifies the number of pixels that shift across the input matrix in each convolution step. At the same time, zero-padding is used to add zeros around the edges of the input to control the size of the output volumes. In Figure 8, there is an example of a CNN architecture.

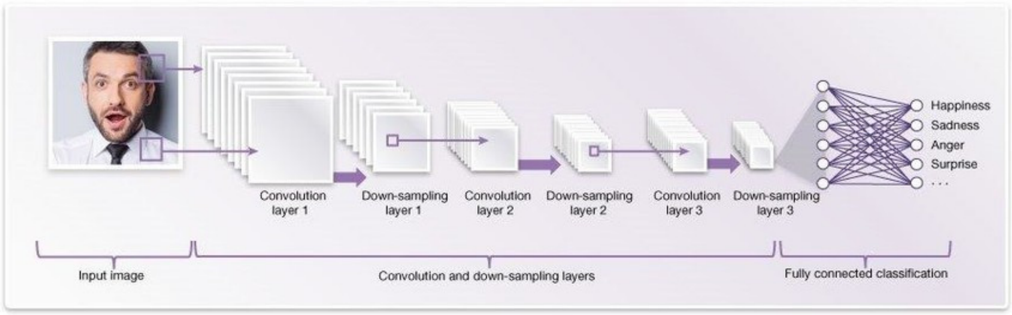


Figure 8. CNN architecture (https://github.com/somillko/Facial-Expression-Recognition accessed on 1 February 2023).

After the convolutional layers, also known as down-sampling, a **pooling layer** is often added to reduce the size of the convolved features, thereby decreasing the network's computational burden [209]. Pooling reduces complexity, improves efficiency, and limits the risk of overfitting.

The final layer of a CNN is typically a **fully connected layer**, which performs the classification task based on the features extracted by the preceding layers. The fully connected layer (FC) [210] comprises the weights and biases together with the neurons, and each node in the output layer connects directly to a node in the previous layer. This layer enables the 2D feature maps to perform the classification task, enabling the 2D feature maps to be based on the features extracted through the previous layers and their different filters.

Pretrained CNN models, such as AlexNet [211], VGG [212], VGG-face [213], GoogleNet [214], Inception [215], and ResNet [216], can be particularly useful for tasks like facial expression recognition. These models are trained on large-scale image datasets such as ImageNet [217], and have shown their effectiveness in extracting meaningful features from images.

Figure 9 illustrates that AlexNet and VGG have many parameters, primarily contributed by their fully connected layers. On the other hand, GoogLeNet and ResNet have fewer parameters while achieving an accuracy rate of approximately 70%.

The latest incarnations of Inception and ResNet flatten the steep straight line of other architectures, indicating an approaching inflection point where costs start to outweigh accuracy gains. Inception V4, a hybrid of ResNet and Inception, offers an impressive 80% accuracy and may be the best choice according to the plot.

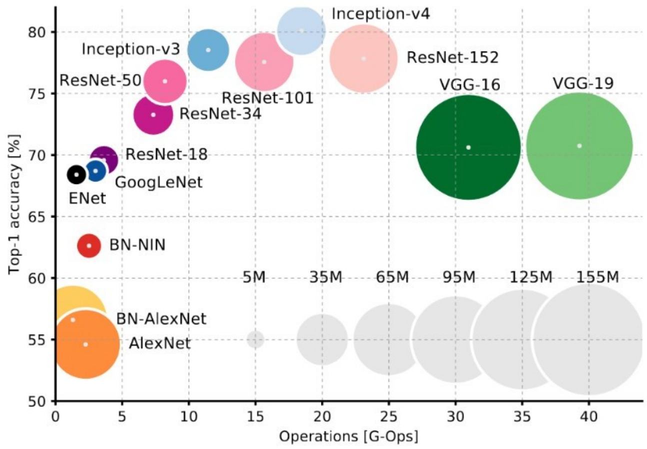


Figure 9. Accuracy vs. Operations, Size Parameters. This figure depicts the relationship between accuracy and the number of operations required for a single forward pass through the network. The size of each blob is proportional to the number of network parameters. A legend is provided in the bottom-right corner, indicating the range of parameters, which spans from 5 × 106 to 155 × 106. Both axes are shared across these figures, and grey dots are used to highlight the centers of the blobs [218].

Kahou et al. [115] proposed augmenting the dataset to build high-capacity models without overfitting, thereby enhancing facial expression recognition (FER) performance. Xception [219] is another promising model for FER that has shown slight improvements over Inception V3 on the ImageNet dataset. Xception uses depthwise separable convolutions to capture spatial features efficiently. DenseNet [220] is also recommended, as it enhances feature propagation, encourages feature reuse, reduces the number of parameters, and mitigates the vanishing gradient problem.

Furthermore, EfficientNet [221] has emerged as one of the top-performing models for image recognition. Its model variants range from EfficientNet-B0 to EfficientNet-B7. These models aim to achieve state-of-the-art accuracy while minimizing computational resources. However, it is important to note that training a Convolutional Neural Network, especially on large datasets, can be time-consuming and may require specialized hardware like GPUs.

While conventional CNNs excel in extracting spatial features from input images, they struggle to capture the temporal interactions in video sequences. This limitation highlights the need for specialized architectures, such as recurrent neural networks (RNNs) or spatiotemporal models, to model video temporal dynamics effectively.

In summary, strategies like data augmentation, leveraging advanced models like Xception, DenseNet, and EfficientNet, and exploring specialized architectures can significantly advance FER and other image recognition tasks. Tran et al. [222] proposed 3D-CNNs to capture the spatial and temporal properties in video clips. The main drawback is the increase in the number of training parameters.

As we mentioned before, preprocessing plays a crucial role in the performance of deep learning models. Data augmentation, cropping, down-sampling, and normalization are standard techniques for improving model robustness and accuracy. Data augmentation involves generating additional training samples by applying various transformations to the original data. Lopes et al. [203] demonstrated that combining these preprocessing methods can significantly improve the accuracy of CNNs.

Konda et al. [223] proposed a zero-bias model for the fully connected layer in CNNs, providing another avenue for optimization.

Researchers have suggested that more advanced and deeper CNN architectures can improve the accuracy and performance of computer vision algorithms. Ding et al. [224] proposed a novel architecture, FaceNet2ExpNet. The author initially suggested a probabilistic distribution function to characterize the high-level neuron response based on a previously modified face net. This leads to feature-level regularization, which uses the face net's rich face knowledge. Label supervision was also advocated in the second phase to increase the final discriminative capability.

Deep Belief Network (DBN)

Deep belief networks [198] are multi-layer generative models in machine learning that feature a deep, feedforward architecture composed of multiple hidden layers (see Figure 10). Unlike conventional neural networks, DBNs are built using unsupervised networks like Restricted Boltzmann Machines (RBMs) or autoencoders (AE). DBNs are versatile and can be employed for supervised and unsupervised learning tasks. One of their primary applications is in dimensionality reduction in the feature space. The learning process for DBNs is twofold: it involves layer-by-layer pre-training followed by fine-tuning [225]. Each RBM layer is trained sequentially from the bottom up in the pre-training phase. The fine-tuning phase utilizes back-propagation algorithms to adjust the network parameters. The introduction of layer-wise pre-training has been a significant factor in the success of DBNs.

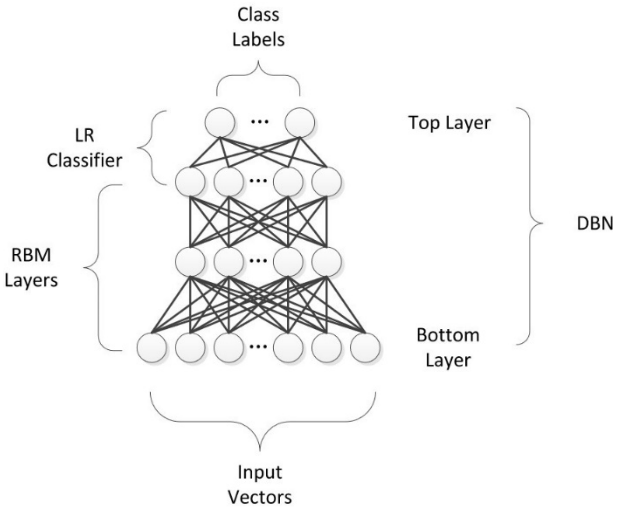


Figure 10. Structure of DBN [226].

Recurrent Neural Network (RNN)

Originally developed in the 1980s, Recurrent Neural Networks (RNNs) have recently gained renewed attention due to advances in computational power, the availability of vast datasets, and the introduction of Long Short-Term Memory (LSTM) in the 1990s [200]. Unique in their ability to maintain an internal memory, RNNs have emerged as one of the most promising neural network architectures currently in use. RNNs are particularly adept at processing sequential data, capturing and modeling temporal dependencies effectively. This makes them well-suited for various tasks, including natural language processing, speech recognition, and time series analysis. Their internal memory allows them to retain crucial information about the input sequences they process, enabling the accurate forecasting of future events. The architecture of a simple RNN is depicted in Figure 11.

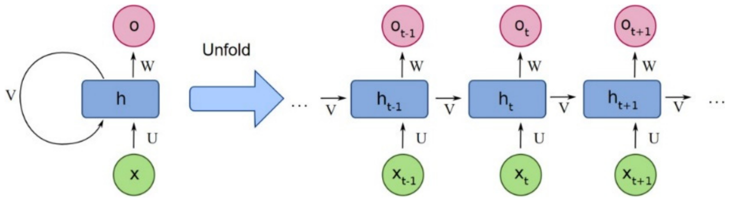


Figure 11. Simple RNN architecture (https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010 accessed on 1 February 2023).

Compared to other neural network architectures, RNNs can develop a deeper understanding of sequences and their contexts. Typically, RNNs are trained using the Backpropagation Through Time (BPTT) algorithm [227].

Recurrent Neural Networks have four main types:

- **One-to-one (or Vanilla NN)**: Handles one input to one output, typical in regular ML problems.
- **One too many**: Generates a sequence of outputs from one input, such as in image captioning.
- **Many to one**: Takes a sequence of inputs to produce one output, useful in sentiment analysis.
- **Many to many**: Takes a sequence of inputs and produces a sequence of outputs, as in machine translation.

To analyze sequential data, deep learning researchers have introduced several specialized tools, including Recurrent Neural Networks (RNNs) [200], Long Short-Term Memory networks (LSTMs) [228], and 3D Convolutional Neural Networks (3DCNNs) [222]. Among these, RNNs have successfully handled various types of sequential data. Their applications range from speech recognition and time series prediction—such as stock price forecasting—to natural language processing tasks like text mining and sentiment analysis. RNNs are also used in machine translation, where an input in one language is translated into an output in another language and action recognition tasks. RNNs have advantages over CNNs regarding learning the relationships in sequential or temporal data. They are adept at capturing dependencies and identifying temporal patterns. However, the choice between using RNNs or CNNs often depends on the specific characteristics of the data and the requirements of the task at hand. There are different RNN Architectures:

Bidirectional Recurrent Neural Networks (BRNN)

One issue of RNNs is that the network does not consider future inputs to make decisions. This problem is solved with BRNN [229], where inputs from future time steps are used to make decisions and improve the network's accuracy.

Gated Recurrent Units (GRU)

Introduced in 2014, Gated recurrent units (GRUs) aim to solve the vanishing gradient problem common in RNNs [230]. They utilize reset and update gates to control the flow of information and have fewer parameters than LSTMs.

Long Short-Term Memory (LSTM)

LSTMs [228] can learn long-term dependencies and are designed to address the vanishing gradient problem [228]. They use input, output, and forget gates to control the information flow. Similar to GRUs, these gates determine which information to retain.

The differences between GRUs and LSTM are illustrated in Figure 12.

Yu et al. [231] proposed an end-to-end architecture, Spatio-Temporal Convolutional LSTM (STC-NLSTM), that combined 3DCNN with Nested LSTM for robust facial expression recognition. Their model achieved impressive performances on multiple datasets. In particular, each convolutional layer's spatiotemporal characteristics were modeled using a T-LSTM, and the network's intermediate layers' multi-level features were encoded using a C-LSTM that integrated the outputs of all T-LSTMs. Yu et al. [231] tested their model on MMI [97,98], CK+ [76–78], Oulu-CASIA [88], and BP4D with a performance for each dataset of 84.53%, 99.8%, 93.45%, and 58%, respectively.

In [232], the authors suggested two methods: a double-channel Weighted Mixture Convolutional Neural Network (WMCNN) and a double-channel WMCNN–Long Short-Term Memory (WMCNN-LSTM). In brief, the WMCNN-LSTM network can swiftly identify facial emotions from the static picture characteristics provided by the WMDCNN network (WMDNN was first proposed by Yang et al. [233]). The WMCNN-LSTM network uses static picture features to acquire the temporal components of an image sequence, allowing for the exact identification of face expressions.

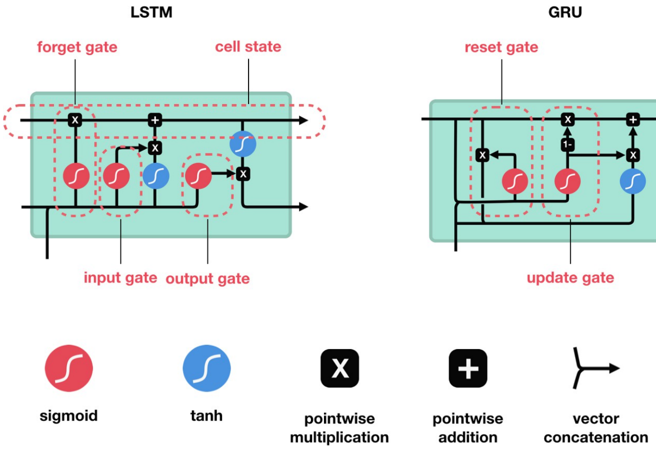


Figure 12. GRU and LSTM (source: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 accessed on 1 February 2023).

Dresvyanskiy et al. [234] introduced a multimodal system using CNN + LSTM architecture on non-lab-controlled data (AffWild2).

Deep-learning-based generative models have made significant strides in recent years, capitalizing on large datasets, innovative network designs, and advanced training techniques. These models can generate realistic content in various forms, including images, text, and audio. Two prominent generative models are Autoencoders (AEs) and Generative Adversarial Networks (GANs).

Autoencoder (AE)

An autoencoder is a specialized neural network designed to encode input data into a lower-dimensional form and then decode them back into their original state. The primary objective is to learn a compact representation of the input data and is useful for data compression, dimensionality reduction, and feature extraction. For instance, autoencoders often outperform traditional techniques like Principal Component Analysis (PCA) in dimensionality reduction.

The architecture of an autoencoder comprises two main components: (1) an encoder that compresses the input into a lower-dimensional representation and (2) a decoder that reconstructs the original input from this compressed form.

Autoencoders operate unsupervised, as they do not require explicit labels for training. They can compress various data types, offering an approximate but degraded (lossy) reconstruction of the original input.

GAN

GANs [235] employ the concept of adversarial learning, which involves two neural networks: the Generator and the Discriminator (see Figure 13). The Generator creates new instances, while the Discriminator evaluates whether the generated examples are authentic. The generator model is used for creating new instances, and the discriminator model determines whether produced examples are genuine examples from the domain or fraudulent ones created by the generator model.

As we can see, GANs have a range of applications in many areas, such as [236]:

Generating Examples for Image Datasets, Generating Photographs of Human Faces, Generating Realistic Photographs, Image-to-Image Translation, Text-to-Image Translation, Semantic-Image-to-Photo Translation, Generating New Human Poses, Photograph Editing, Face Aging, Super Resolution, Photo In painting, Clothing Translation, Video Prediction, 3D Object Generation, and many more.

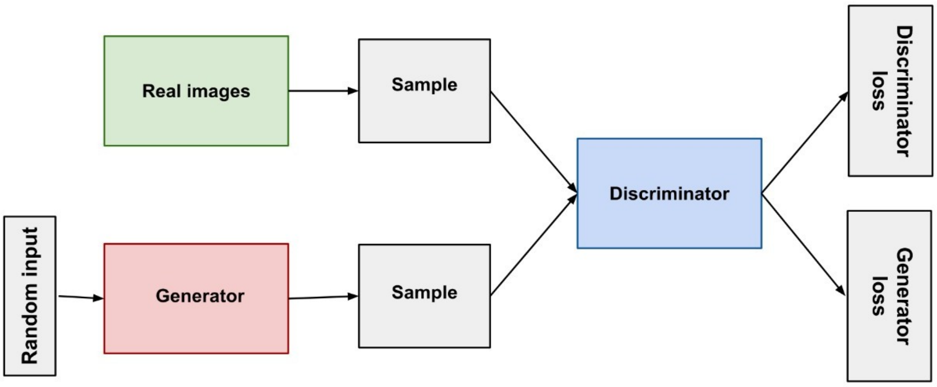


Figure 13. GAN (source: https://developers.google.com/machine-learning/gan/gan_structure accessed on 1 February 2023).

Regarding the types of GAN, there are Vanilla GAN, Conditional GAN (cGAN), Deep Convolutional GAN (DCGAN), CycleGAN, Generative Adversarial, Style GAN, and Super Resolution GAN (SRGAN).

In recent years, there has been a significant proliferation of innovative frameworks based on Generative Adversarial Networks (GANs) due to their rapid development. So, some good works that are worth mentioning about GAN are: Shen et al. [237], who introduced FaceFeat-GAN, a two-stage framework. The first stage involved competing in the feature domain to synthesize facial traits. In contrast, the second stage involved competing in the image domain to generate photo-realistic images that exhibited a wide range of variations while maintaining the subject's identity. Also, Shiri et al. [238] introduced Identity-Preserving Face Recovery from Portraits (IFRP) to restore a latent photorealistic face while maintaining the subject's identity. This approach involved utilizing a Style Removal network (SRN) and a Discriminative Network (DN).

Hybrid Model

Hybrid models are formed by integrating distinct, heterogeneous machine learning methodologies, such as features from deep learning and traditional machine learning (also known as handcrafted features). The most well-known hybrid models are CNN + DBN, CNN + RNN, and CNN + GAN, etc.

While deep learning has gained popularity due to the availability of large datasets and affordable GPU-based processing power, it often faces limitations in specific contexts like facial expression recognition, where datasets might be constrained. Traditional handcrafted algorithms, although generally less accurate than CNNs [239], do not require extensive datasets for generalization. However, these traditional methods are limited by their fixed transformations or filters that remain constant across different data sources.

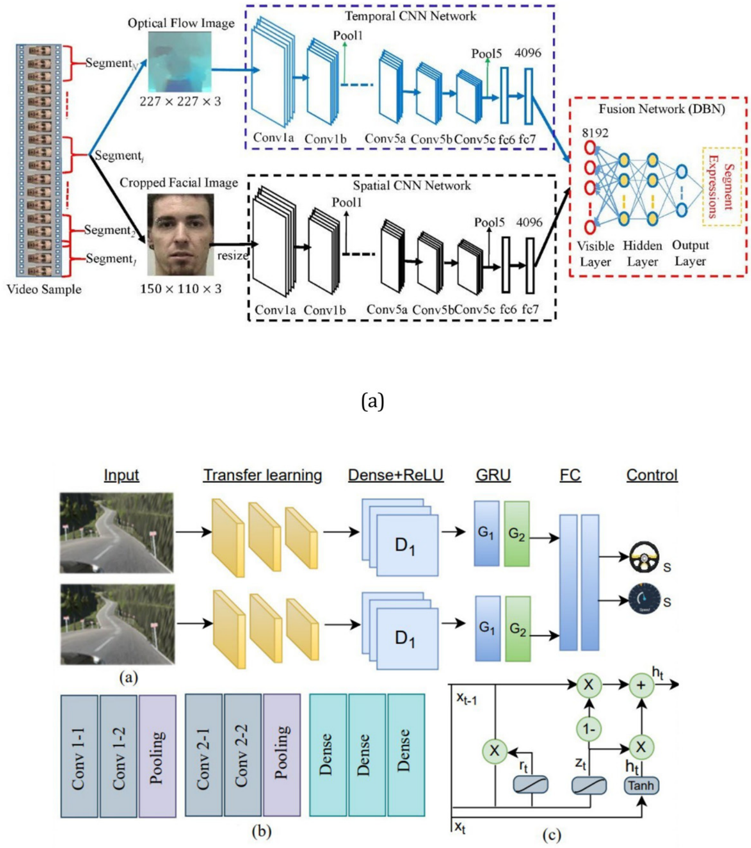
Another reason a hybrid model must be used is that handcrafted feature representations are low-level to discriminate dynamic facial expressions, so deep-neural-network-driven feature learning representations may achieve a better performance, i.e., a nice approach in [240] with spatial and temporal CNNs + DBN. Some applications of hybrid models might be autonomous driving to find the positions of other cars [241].

A hybrid model that combined traditional machine learning approaches and deep learning approaches was conducted by Pan et al. [242] focusing on the field of video-based facial expression recognition (VFER). Their study employed convolutional neural networks (CNN) and a histogram of an oriented gradient (HOG) to extract a wide range of information from video frames effectively.

Also, Jain et al. [58] developed another hybrid model that combined CNN and RNN architectures. This model demonstrated a superior performance, achieving an overall accuracy of 94%. Particularly, this model exhibited an excellent performance on the CK+ database, achieving an accuracy rate of 94.91% [76–78].

Similarly, Sun et al. [243] proposed a novel hybrid model that combined CNN and SIFT features with an SVM classifier. When compared to five state-of-the-art methods, their approach yielded the best results on the CK+ database, achieving an accuracy of 94.13%. Furthermore, a recent study by Abdulsattar et al. [244] explored the combination of CNN with Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP). It appears that the hybrid model involving CNN and HOG yielded the best accuracy.

In Figure 14, examples of two hybrid model frameworks for facial expression recognition are illustrated, as proposed in [119,240].



(a)

(b)

Figure 14. Examples of two hybrid model frameworks (from top to bottom [240,241]) (**a**) the outlined approach, which encompasses two distinct streams: the temporal stream and the spatial stream. The purpose of the temporal stream is to handle temporal signals, while the spatial stream focuses on processing spatial signals. These two streams are then combined to create a comprehensive video feature. Subsequently, a Support Vector Machine (SVM) is employed for the recognition of facial expressions. The method is structured into four key stages: (i) preprocessing of the video, (ii) extraction of features, (iii) fusion of features, and (iv) recognition of emotions [241]. (**b**) The suggested framework is divided into three main components: (a) the input section; (b) the recommended model, termed VGG-GRU; and (c) the mechanism for controlling both the steering angle and the vehicle's speed [240].

Graphs

Graphs are used as modeling tools to capture intricate relationships among facial features and expressions, providing a nuanced and detailed analysis of facial affect. By representing facial data as a graph, it becomes possible to incorporate contextual information and capture the dependencies between different facial components. Using graph-based approaches has significantly improved facial expression recognition models' performances and generalization capabilities. Unlike traditional methods that rely solely on appearance and geometry features, graph-based techniques offer a more comprehensive understanding of the underlying connections and interactions among facial attributes.

In recent advancements, deep learning models—specifically, Graph Neural Networks (GNNs)—have been at the forefront of graph-based facial affect analysis. These models leverage their capability for relational reasoning based on graph-based representations, allowing for a deeper exploration of the relationships within the facial graph. By integrating GNNs into their research, scientists can effectively capture the complex dependencies and patterns between facial features and expressions, leading to more accurate and robust facial affect analysis systems.

In summary, the adoption of graph-based methodologies and the integration of deep learning models are revolutionizing the field of facial expression recognition. These advanced techniques enable a more holistic understanding of facial affect, paving the way for improvements in emotion detection, human–computer interaction, and affective computing applications.

Transformers

A Transformer [245] is a deep learning architecture extensively used in various natural language processing (NLP) tasks, including language translation, text summarization, and language modelling. A Transformer is a type of deep learning model architecture that was introduced in the paper titled "Attention Is All You Need" by Vaswani et al. in 2017 [245]. This architecture utilizes a self-attention mechanism to weigh the importance of different words within an input sequence. This enables it to capture long-range dependencies effectively, thereby enhancing its performance on tasks requiring a comprehensive understanding of the input text. The Transformer model consists of an encoder and a decoder, each comprising multiple layers that feature self-attention mechanisms and feed-forward neural networks. The encoder processes the input sequence to generate a fixed-length vector representation, which is then used by the decoder to produce the output sequence, one word at a time.

3.2.2. State-of-the-Art

This section reviews cutting-edge deep neural networks designed for Facial Expression Recognition (FER).

3.2.3. Static Images Deep FER Networks

Loss Layer

In practical circumstances, FER often grapples with considerable intraclass variance and high interclass similarity. The standard softmax loss layer in CNNs mainly aims to separate features from different classes. To address these challenges, various researchers have proposed alternative loss layers.

Wen et al. first proposed the concept of center loss in 2016 for enhancing feature discriminability in face recognition tasks [246]. Center loss is utilized in face recognition tasks to enhance the discriminative power of learned features. This loss function addresses two fundamental yet challenging objectives: reducing intra-class dispersion and increasing inter-class differences. Specifically, center loss functions by computing a center vector for each class in the feature space. During the network's training phase, the algorithm minimizes the Euclidean distance between the features and their corresponding class center vectors. The dual effect of this minimization process is twofold: first, it contracts the feature vectors corresponding to the same class toward their shared center, thereby reducing intra-class variations. Second, it maximizes the distance between different class centers, thus increasing inter-class separability. Notably, when used with softmax loss, center loss empowers the network to learn highly discriminative features. This combined loss strategy has proven particularly effective in scenarios characterized by pronounced intra-class variability and subtle inter-class differences, enhancing face recognition models' accuracy and robustness.

To strengthen the discriminative capacity of feature representations in deep neural networks, Li et al. introduced a seminal architecture known as Deep Locality-Preserving Convolutional Neural Network (DLP-CNN) [94]. This architecture is meticulously designed to accentuate the importance of the local structure within each class by minimizing the distance to the K-nearest neighbors in the high-dimensional feature space. The focus on local structure enhances the model's efficacy in applications that require robust intra-class compactness.

Simultaneously, a novel loss function known as 'Island Loss' has emerged as a powerful mechanism for augmenting the discriminative quality of deep features. Unlike conventional loss functions, Island Loss adopts a dual-objective optimization strategy. It not only aims to minimize intra-class variances, but also to maximize inter-class separability. This is achieved by creating distinct clusters or 'islands' in the feature space, where each island corresponds to a specific class. Notably, Island Loss can be integrated seamlessly with other loss functions, such as softmax or center loss, offering a multi-faceted optimization landscape.

In a parallel research endeavor, a unique variant of the island loss layer was proposed in the IL-CNN architecture [247]. This architecture comprises three convolutional layers, each succeeded by batch normalization (BN) and PReLU (Parametric Rectified Linear Unit) activation layers. Preliminary results have suggested that IL-CNN demonstrates a comparable performance to existing state-of-the-art algorithms in facial expression recognition.

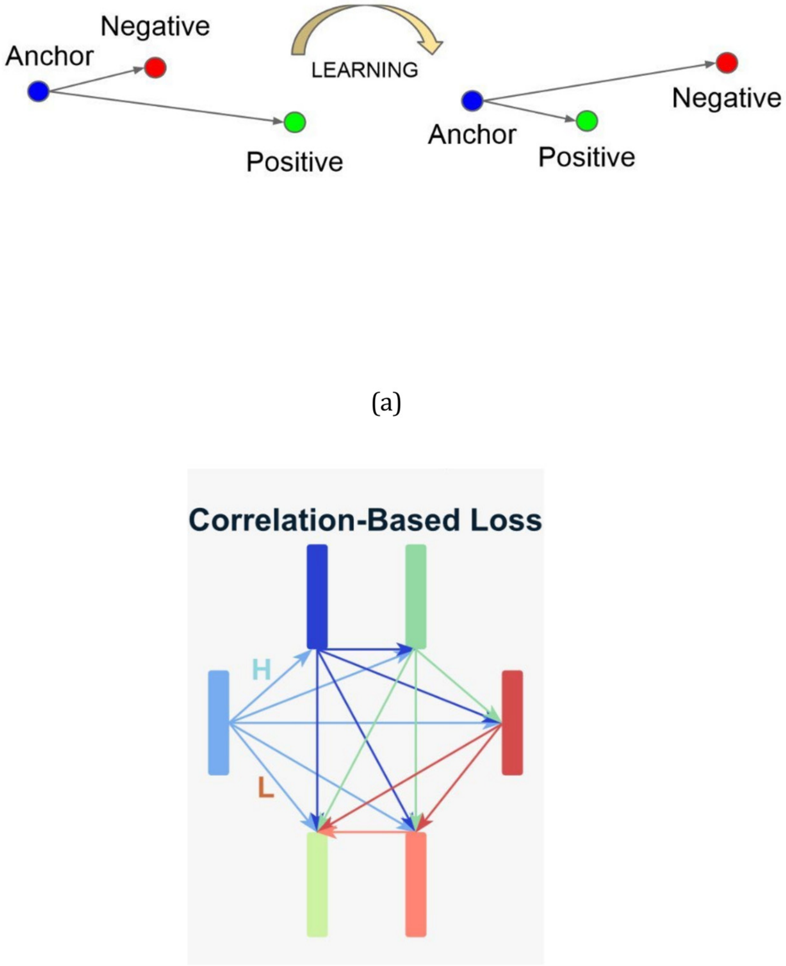
The salient distinction between these methods lies in their approach to feature optimization. While center loss concentrates on collapsing intra-class samples towards a central cluster, DLP-CNN emphasizes a distributed yet locally coherent representation by pushing feature vectors toward their K-nearest intra-class neighbors.

A seminal contribution to loss function is Triplet Loss, introduced by Schroff et al. [248]. This mechanism mandates that an anchor example be closer to a positive than a negative one in the embedded feature space. An extension to this paradigm is Exponential Triplet Loss [249], which adopts a sample-importance-driven strategy to refine the model, thereby yielding a more adaptable and resilient architecture. To address the intricacies associated with anchor selection and threshold tuning in identity-invariant FER, Liu et al. [250] proposed an (N + M)-tuplet clusters loss function. This loss function is designed to streamline the complexities inherent in handling a multitude of face identities and expressions.

Li et al., in 2019 [251], introduced a bifurcated loss mechanism, termed 'Separate Loss'. The architecture uses one term to amplify intra-class similarity while deploying a second term to attenuate inter-class similarity, thus achieving a balanced class separability. Also, Li et al. [252] proposed a novel method in FER, adaptive supervised AdaReg Loss, useful for observing CNNs as they learn expression representations with class imbalances. A novel loss function competitive with triplet loss is the Adaptive Correlation (Ad-Corre) Loss function by Fard et al. [60], and, as the backbone used Xception network, this architecture essentially consists of three parts (Feature Discriminator, Mean Discriminator, and Embedding Discriminator). A Future Discriminator directs the network to build the embedded feature. If two vectors are in the same class, they will be highly correlated; if they are in separate classes, they will be less correlated. The Mean Discriminator component causes the mean embedded feature vectors of distinct classes to be less similar to each other. The network is penalized by the Embedding Discriminator component for producing the different embedded feature vectors.

In Figure 15, we present a schematic representation of three state-of-the-art loss architectures that have been instrumental in shaping the current landscape of Facial Expression Recognition (FER), (a) island loss, (b) triplet loss (the distance between an anchor and a positive with the same identity minimized, while the distance between the anchor and a negative with a different identity is maximized), and (c) correlation-based loss, which considers all of the data in a mini-batch and directs the network to generate embedded feature vectors that are substantially connected for comparable classes and uncorrelated for different classes. H stands for strongly correlated, whereas L stands for less correlated.

Adding to this body of work, Farzaneh et al. [54] recently proposed a novel loss function termed Deep Attentive Center Loss (DACL). In this architecture, the attention mechanism is ingeniously incorporated by using the intermediate spatial feature maps extracted from the Convolutional Neural Network (CNN). This enables the network to calculate attention weights that are directly associated with feature significance, thereby enhancing the model's ability to focus on more discriminative aspects of the data.



(a)

(b)

Figure 15. Loss layer networks. (**a**) Triplet loss [248] and (**b**) correlation-based loss [60]. (**a**) This technique aims to reduce the gap between an anchor sample and a positive sample that share the same identity, while simultaneously increasing the gap between the anchor sample and a negative sample that has a different identity [248]. (**b**) This approach analyzes all samples within a mini-batch to guide the network in generating embedded feature vectors. The goal is for feature vectors of similar classes to exhibit strong correlation, whereas those of distinct classes should show weak correlation. Here, "H" signifies a strong correlation and "L" signifies a weak correlation [60].

Ensemble Network

While traditional Convolutional Neural Networks (CNNs) have made notable strides in the domain of Facial Expression Recognition (FER), they are not without limitations. The following challenges have been particularly salient:

1. Hyperparameter tuning in CNNs often requires significant domain expertise and computational effort.

2. The stochastic gradient descent algorithm has shown limitations in training deep architectures, particularly when handling large datasets. This is exacerbated by issues like vanishing gradients.
3. CNNs designed for FER are sensitive to a plethora of real-world variables, including age, gender, facial morphology, and ethnic background. Thus, face emotions have overlapping aspects, making them less robust and complicating implementation.

To mitigate these issues, ensemble methods have gained prominence. So, ensemble networks are primarily used for classification tasks rather than data augmentation. Contrary to common belief, ensemble networks are not merely a tool for data augmentation; rather, their primary function is to enhance classification accuracy by integrating the strengths of individual base classifiers. Two crucial factors must be considered when constructing an ensemble: (1) diversity among the base models to ensure complementary strengths, and (2) an effective strategy for aggregating the predictions of these models.

A seminal work by Jung et al. [253] introduced the Deep Temporal Appearance-Geometry Network (DTAGN), which unifies two specialized neural networks. The first, termed Deep Temporal Appearance Network (DTAN), uses a CNN to analyze information about the appearance of the face, while the second deep network, the Deep Temporal Geometry Network (DTGN), utilizes a fully connected Deep Neural Network (DNN) to capture facial landmark movements over time. These two networks are then combined to create the Deep Temporal Appearance-Geometry Network (DTAGN).

For preprocessing, a normalization process is used to make the input a fixed length, namely, detecting and cropping the faces of the image sequence, and then rescaling them to a size of 64 × 64 pixels. IntraFace is applied, which detects accurate facial landmark points such as two eyes, a nose, a mouth, and two eyebrows, which are all labeled as 49 facial landmark points.

The author tested the effectiveness of his network on three different datasets of facial images (CK+, MMI, and Oulu-CASIA) and found that it surpassed existing methods (97.25%, 81.46%, 70.24%, respectively).

Another commendable work integrated ensemble models, specifically VGG13, VGG16 [212], and ResNet, to achieve promising results in FER [254].

Mollahosseini et al. [255] proposed a deep CNN architecture, incorporating inception layers and adhering to the Network-in-Network paradigm [256]. This architecture (network-in-network) not only boosts a localized performance, but also mitigates overfitting issues. Their preprocessing pipeline involves Active Appearance Models (AAM) and Supervised Descent Models (SDM), followed by data augmentation. The architecture is made up of two convolutional layers (max pooling) and four inception layers. It seems that it is innovative work due to the inception layer for FER tasks across numerous datasets.

Fan et al. [257] proposed a Multi-Region Ensemble CNN (MRE-CNN) framework, combining three sub-networks (AlexNet, VGG-16, and VGG-Face) to enhance the learning power of CNN models. Also, in this network face detection, alignment, cropping, and resizing occurred. Then, data augmentation followed. Then, this method was tested with AFEW 7.0 80-82] and RAF-DB [94] with a good performance (47.43% and 76.73%, respectively).

Georgescu et al. [258] pioneered an approach that amalgamates handcrafted features derived from a bag-of-visual-words model with features automatically learned through Convolutional Neural Networks (CNNs). They employed a Support Vector Machine (SVM) classifier to integrate these features, offering a robust and versatile classification framework. In another groundbreaking work, a paper by Zhang et al. [240] introduced the STRNN model, a complex deep learning architecture adept at capturing both spatial and temporal information. The model utilizes multi-directional Recurrent Neural Network (RNN) layers for spatial variations and bi-directional temporal RNN layers for sequential data. This sophisticated architecture is further enhanced by sparse projection techniques, and empirical results have corroborated its superiority over existing methods.

Rajan et al. [259] proposed a novel deep learning framework that combines a convolutional neural network (CNN) with a long short-term memory (LSTM). The model incorporates two distinct preprocessing techniques aimed at compensating for illumination variations and retaining delicate edge features. The generated spatial feature maps are subsequently fused and processed through an LSTM layer, capturing the temporal dependencies between successive frames.

Khaireddin et al. [260] extended the well-known VGGNet architecture, incorporating additional refinements like Cosine Annealing. Their rigorous preprocessing steps, including data augmentation and image normalization, were instrumental in achieving a commendable 73.28% accuracy on the FER2013 dataset.

Moreover, Pham et al. [61] achieved a noteworthy milestone by proposing the Residual Masking Network (RMN), an ensemble of seven CNNs fortified with residual masking blocks. This architecture currently represents the state-of-the-art in FER, boasting the highest performance on the FER2013 dataset [89].

Wang et al. [261] took the innovative step of fusing ensemble networks with multitasking approaches, giving rise to the OAENet (Oriented Attention Ensemble) architecture. This model consists of two branches: one focused on capturing local highlights through oriented attention ensemble techniques and the other designed to extract high-level semantic features through multiple convolutional blocks. These branches are ultimately fused to produce the final classification outcome. In Figure 16, the NNs and BOVW network [258] are displayed.

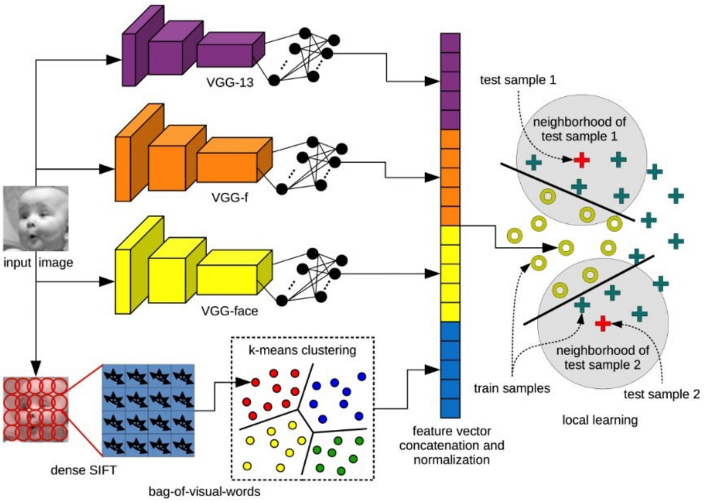


Figure 16. Ensemble network. NNs and BOVW network [258]. This network integrates automatically extracted features from convolutional neural networks (such as VGG-13, VGG-f, and VGG-face) with manually designed features derived using the bag-of-visual-words approach. Following the combination and L2 normalization of the feature vectors, a local learning strategy is applied [258].

Multitask Networks

Multitask learning has been increasingly recognized as a robust framework for addressing the complexities inherent in facial expression recognition (FER). In real-world applications, FER is influenced by many factors, including but not limited to subject identity, lighting conditions, and head orientation. Traditional FER models, which typically focus on single-task learning, cannot capture these intricate interactions among various latent components. However, multitask learning paradigms offer a more comprehensive approach by integrating additional relevant tasks, thereby enhancing feature sensitivity to expression-specific cues and mitigating the impact of confounding variables [262]. Moreover, multitask networks can serve dual purposes: classification and data augmentation.

While multitask learning has shown promise in various applications, one critical challenge is determining the optimal weighting of individual tasks. These weights significantly influence the efficacy of the multitask learning model [263].

Zhang et al. [120] presented a multisignal CNN (MSCNN) trained under dual supervision—recognition and verification tasks. Each task has its loss function designed to increase the discriminative ability between distinct facial expressions while minimizing the intra-class variance. This approach effectively directs the model's focus toward the nuances of facial expressions.

Noteworthy is the work of Ming et al. [263], who introduced a dynamic multitask learning framework capable of adaptively updating task weights based on their relevance throughout the training process. Impressively, this dynamic approach achieved remarkable performance metrics—a 99.5% accuracy on the CK+ dataset [76–78] and 89.6% on the Oulu-CASIA dataset [88]—demonstrating its practical utility in real-world scenarios.

Serengil et al. [264] proposed an innovative approach by integrating the famous LightFace face recognition library with a multifaceted facial feature analysis. This hybrid framework leverages state-of-the-art face recognition architectures, including VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib, and SFace, to achieve a top performance in face recognition. Notably, the framework adopts the same preprocessing techniques—namely, face detection and alignment—that are common to the aforementioned recognition models. To broaden its applicability, the system also performs analyses based on age, gender, sentiment, and race. Given robustness and simplicity, a basic VGG-Face model is created with pre-trained weights, which serve as the cornerstones for the complete solution.

Savchenko has made substantial contributions to multi-task learning networks, with some of his work achieving unparalleled accuracy metrics [39,40]. In one significant study, Savchenko introduced a multi-task learning model that concurrently addresses face identification, gender, ethnicity, and age classification on the UTKface dataset [265] and emotion recognition using the AffectNet dataset. The model leverages lightweight backbone architectures, such as MobileNet, EfficientNet, and ResNet, for computational efficiency. Preprocessing steps include using the MTCNN algorithm for face detection and subsequent cropping [40].

In subsequent work, Savchenko and his team presented Multi-task EfficientNet-B2 and its variant, Multi-task EfficientNet-B0 [39]. These architectures are designed to perform an array of tasks, including face detection, face identification, and facial expression recognition. Initially, face sequences for each subject are extracted using a combination of face detection, tracking, and clustering techniques. A single neural network, pre-trained on face recognition tasks, is then fine-tuned to identify emotional attributes in each frame. This is achieved through a robust optimization strategy tailored specifically for static images from the AffectNet dataset.

Notably, the extracted facial features can be used to swiftly infer collective emotional states, individual emotional expressions (e.g., happiness and sadness), and varying levels of student engagement, ranging from disengagement to high involvement. It is worth highlighting that this model has set the **state-of-the-art benchmark** for recognizing eight distinct emotions, with a **63.03%** accuracy rate on the large-scale AffectNet dataset. Moreover, it has secured the third position for the seven-emotion classification with an accuracy of 66.29%.

Zheng et al. [45] suggested a sophisticated multi-task learning framework, Discriminative Deep Multi-Tasking Learning (DDMTL), using a Siamese-based loss function. This innovative approach incorporates data distribution information and expression labels to enhance facial expression recognition capabilities.

Antoniadis et al. [37] advanced the field by proposing a multi-task learning system that excels at recognizing facial expressions "in the wild". Their approach leverages Graph Convolutional Networks (GCN) to exploit the intricate correlations between categorical and dimensional emotions.

In a landmark study published in 2021, Huang et al. [42] reported a model that achieved remarkable performance metrics—100% accuracy on the CK+ dataset and 90.04% on the FER+ dataset. The architecture comprises three critical components: a primary deep learning model, which can be instantiated as either AlexNet or Inception; dual attention mechanisms, specifically, Grid Wise attention for low-level feature extraction and Visual Transformer Attention for high-level feature discernment; and an FER module capable of differentiating between simple and complex facial expressions, utilizing the C-F labelling technique as a reference point. Furthermore, they introduced a large-scale model variant that amalgamates the first two components to deliver a superior performance. Additionally, they conceptualized a novel Emotional Education Mechanism (EEM) to enable the efficient deployment and optimization of lightweight FER models.

Foggia et al. [46], in 2023, recently unveiled a 12-task CNN model that offers a unique approach to multi-task learning by integrating three distinct paradigms of parameter sharing across tasks. This model employs diverse backbone architectures—MobileNet, ResNet, and SENet [266]—to perform a range of tasks, including gender, age, facial expression recognition, and ethnicity classification [46].

In Figure 17, different Mutli-tasking Networks, (a) EmotionGCN [37] and (b) Multi-task EfficientNet-B0 [40], are displayed.

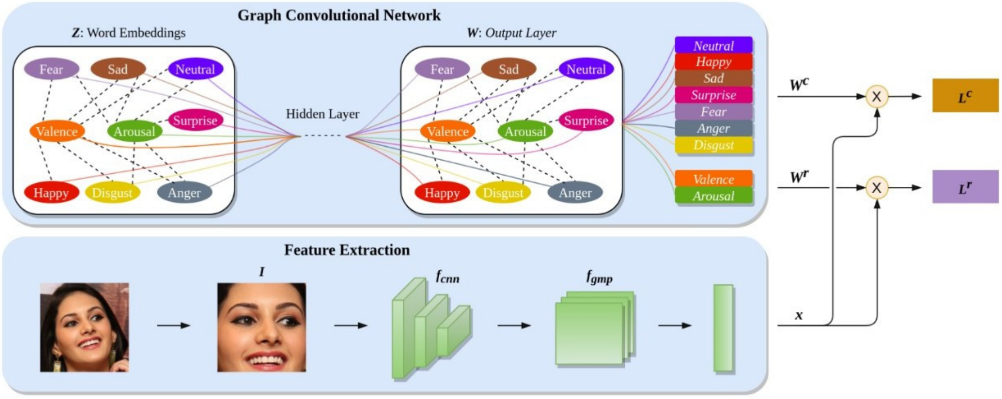


Figure 17. Multi-tasking Networks. Emotion GCN [37] The Emotion-GCN model is designed for Facial Expression Recognition (FER) in natural environments, incorporating a graph that links seven expression labels with two valence-arousal dimensions via an adjacency matrix. It utilizes Graph Convolutional Networks (GCNs) to process word embeddings into classifiers and regressors for mapping facial expressions. Image representations are extracted using a DenseNet-based CNN and refined by global max-pooling, enabling the model to perform both expression classification and valence-arousal regression through end-to-end training [37].

In [267], the Affective Behavior Analysis in-the-wild (ABAW) Competition has a Multi-Task Learning challenge. Participants are tasked with designing a unified algorithm capable of concurrently performing multiple affective analysis tasks, encompassing valence–arousal estimation, expression classification, and action unit detection. This initiative serves to stimulate advancements in affective computing, thereby facilitating the development of highly efficient algorithms for emotion analysis in ecologically valid settings.

Cascaded Networks

Cascade-forward neural networks resemble their feed-forward counterparts but feature connections from both the input and preceding layers. In a typical three-layer configuration, the output layer is directly linked to the hidden layer and the input layer. This cascaded architecture allows for the incremental integration of various modules to construct deeper networks that can perform tasks ranging from classification to data augmentation.

Yan et al. [48] proposed a real-time hybrid model for user profiling, encompassing gender, age, ethnicity, and emotion recognition from facial images. This model employs a

Sparse Autoencoder (SAE) and a Shallow Convolutional Neural Network (SCNN). The initial preprocessing steps involve face detection via the Dlib library, which utilizes the HOG-SVM algorithm, followed by image cropping, normalization, and data augmentation. The processed data are then subjected to SAE for image reconstruction and abstract feature extraction, which are subsequently integrated with a CNN for feature refinement. Classification is accomplished via a softmax layer. The model's experimental validation underscores its broad generalizability across various datasets and its suitability for real-time applications, owing to its shallow architecture.

In a recent contribution, a paper [47] introduced a cascaded attention-based network that synergistically combines attention mechanisms with pyramid features to exploit local spatial, multi-scale stereoscopic spatial context and temporal features.

In 2022, Li et al. [69] proposed an innovative approach to pre-train a general-purpose encoder for facial expression analysis through self-supervised contrastive learning. Utilizing the CRS-CONT method, the encoder attracts the features of positive pairs while repelling those of negative pairs. The resulting architecture demonstrates a remarkable adaptability to varied facial expressions, outperforming existing state-of-the-art methods. It also shows potential for reducing computational overhead in fully supervised feature learning scenarios.

GAN

Generative Adversarial Networks (GANs) have primarily been employed to synthesize realistic images, such as faces and numbers, thereby enriching data augmentation strategies for various machine learning tasks. While GANs are inherently generative models that create new data samples that mimic a given dataset, their utility extends beyond mere data generation. They consist of a generator and a discriminator network, which engage in a game-theoretic contest to improve each other's performances. Notably, the synthetic data produced by GANs can be harnessed as supplementary training data or for data augmentation to enhance classification tasks.

Hyang et al. [268] introduced a specialized GAN architecture designed to produce contextually valid facial expressions in dyadic human interactions [268]. Distinct from prior research that primarily focused on the facial attributes of generated identities, Hyang et al. leveraged the concept of dyads to model the reciprocal influence of facial expressions between interacting individuals.

Furthermore, the authors introduced a two-tiered GAN architecture tailored for interviewer–interviewee dynamics. The first tier generated dynamic sketches of the interviewer's facial expressions based on the expressions exhibited by the interviewee. Subsequently, the second tier synthesized realistic face images from these sketches. Through extensive quantitative evaluations, the authors established that their model was adept at generating visually convincing facial expressions that were contextually appropriate within dyadic interactions. Moreover, they demonstrated that the synthesized expressions on the interviewer's face were accurate emotional responses to the interviewee's behavior.

Zhang et al. [121] presented an end-to-end deep learning model based on a generative adversarial network (GAN) for simultaneously exploiting diverse postures and expressions for pose-invariant facial expression identification. The architecture employs an encoder–decoder scheme to craft both generative and discriminative identity representations for facial images. Remarkably, the identity representations are decoupled from variations in both expression and pose through the utilization of specialized expression and pose codes. This allows the model to automatically synthesize facial images across a spectrum of emotions and poses, thereby enriching the training dataset for facial expression recognition (FER). Both quantitative and qualitative evaluations across controlled and uncontrolled datasets substantiate the model's superiority over existing state-of-the-art methods.

Yang et al. [269] undertook a novel approach called De-expression Residue Learning (DeRL) for facial expression recognition. The method leverages a generative model, trained via a conditional GAN, to create a corresponding neutral face image for any given input facial image. While the generative model ostensibly filters out expressive information, these data are retained within the intermediate layers of the architecture. DeRL capitalizes on this by extracting the residual expressive information embedded within these layers, thereby generating highly informative features for FER. The efficacy of this approach was rigorously validated across seven public facial expression datasets. Pre-training was conducted on the BU-4DFE and BP4D-spontaneous datasets, while evaluations were performed on the CK+, Oulu-CASIA, MMI, BU-3DFE, and BP4D+ datasets. The comprehensive experimental results point to a marked improvement over previous methodologies. Wu et al. presented a groundbreaking architecture known as Cascade Expression Focal GAN (Cascade EF-GAN) to address the inherent limitations observed in current facial expression editing methodologies [270]. Existing techniques often suffer from the introduction of artifacts and blur, particularly around areas of the face with intense expressions. Moreover, they frequently generate undesirable overlapping artifacts when tasked with handling large-gap expression conversions, such as transforming an angry expression into a laughing one. The Cascade EF-GAN ingeniously uses localized expression and focuses on preserving identity-related features while minimizing the occurrence of artifacts and blur in the generated facial images. Furthermore, the architecture utilizes a cascade transformation strategy specifically designed to manage large-gap expression transformations, thereby alleviating overlapping artifacts and enhancing the realism of the edited facial expressions. Rigorous evaluations of multiple publicly available facial expression datasets substantiate the model's superiority over existing approaches in the realm of facial expression editing.

Sun et al. [70] introduced an innovative Discriminatively Deep Fusion (DDF) technique that addresses the ubiquitous challenge of limited labeled data in facial expression recognition [70]. The approach leverages an improved Conditional Generative Adversarial Network (im-cGAN) to generate a richer set of labeled expression samples. Additionally, DDF captures both global and local features from facial images and fuses them to generate a robust, composite feature representation. A specialized discriminative loss function (D-loss) is also introduced to further refine the discrimination capability of the fused features. Extensive experimental validation across multiple datasets revealed that the DDF approach outperforms existing state-of-the-art methods and excels in learning abstract representations of facial expressions, even when constrained by limited labeled data.

Graphs

Graph-based methodologies have emerged as a compelling framework for capturing the complex relationships among various facial features and expressions, thereby enabling a nuanced and comprehensive affective analysis [271,272]. These techniques have exhibited marked improvements in both the performance and generalizability of facial expression recognition models when compared to traditional methods that predominantly focus on appearance and geometric features. Recent advancements in this domain have particularly leveraged deep learning paradigms, such as graph neural networks, to conduct relational reasoning on graph-based representations of facial expressions.

Liao et al. [273] proposed a deep neural network called Facial Expression Recognition based on Graph Convolution Network (FERGCN). The proposed FERGCN framework comprises three pivotal modules: a feature extraction component to capture both global and local facial features, a graph convolutional network that refines expression information grounded on the topological graph of key facial points, and a graph-matching module aimed at enhancing the network's discriminative capability across varying expressions. The efficacy of FERGCN has been empirically validated on public datasets, achieving commendable accuracy rates of 88.23% on RAF-DB, 56.15% on SFEW, and 62.0% on AffectNet.

Wu et al. [274] proposed a new method for facial expression recognition using a Graph Convolutional Network (GCN) which is capable of processing non-Euclidean structured data. The methodology entails the construction of an undirected graph from facial images, which is achieved by amalgamating both fixed and random points. This constructed graph serves as the input to a specialized GCN, composed of six graph convolution layers, a fully connected layer, and a SoftMax layer. Each graph convolution layer is further equipped with a signal-filtering and a graph-coarsening layer. Experimental validation on two benchmark datasets, CK+ and JAFFE, reveals that the proposed model surpasses traditional methods that rely solely on fixed points, and even outperforms conventional convolutional neural network approaches in facial expression recognition tasks. This work substantiates the transformative potential of GCN-based methodologies in advancing the field of facial expression recognition.

Transformer

The Transformer architecture has had a profound impact on the field of natural language processing (NLP) and has been widely adopted in various applications, including machine translation, text generation, question–answering, and more.

In 2021, Aouayeb [43] proposed a vision Transformer jointly with a Squeeze and Excitation (SE) block for the FER task. Evaluated on the CK+ dataset, the architecture achieved an almost flawless performance with an accuracy rate of 99.8%.

Huang et al. [42] proposed a new framework for Facial Expression Recognition using two attention mechanisms in CNN-based models. The first attention mechanism employs a grid-wise strategy for low-level feature extraction, capturing the intricate dependencies between facial regions and facilitating effective parameter updates. The second mechanism incorporates a visual transformer approach for high-level semantic representation, enabling the capturing of long-range dependencies between facial regions. The framework was rigorously evaluated on three public datasets—CK+, FER+, and RAF-DB—and exhibited a superior performance, reaching a remarkable 100% accuracy on the CK+ dataset without necessitating additional training data. These findings underscore the efficacy of attention mechanisms in enhancing the FER performance in CNN-based architectures.

The article by Xue et al. [64] introduced the TransFER model, which proposes a novel approach to enhancing the performance of facial expression recognition (FER). The model comprises three fundamental components, namely Multi-Attention Dropping (MAD), ViT-FER, and Multi-head Self-Attention Dropping (MSAD). These components operate synergistically to enable the model to adaptively explore diverse local patches and learn rich, relation-aware local representations. Despite its effectiveness, the model was outperformed by one specific approach [62] across multiple benchmark datasets, although it surpassed other state-of-the-art methods.

In 2023, Li et al. [62] introduced FER-former, a novel multifarious supervision-steering Transformer approach for Facial Expression Recognition (FER) in the wild. The architecture incorporates multi-granularity embedding integration, a hybrid self-attention scheme, and heterogeneous domain-steering supervision. The model seamlessly fuses features from both CNNs and Transformers through a hybrid stem and employs an FER-specific transformer mechanism. Experimental results confirmed the model's superior performance over existing state-of-the-art methods across various benchmarks.

Moreover, Wasi et al. [275] proposed ARBEx as a framework and introduced it for facial expression learning (FEL) tasks. The framework leverages the Vision Transformer and incorporates a range of techniques to address challenges such as class imbalance, bias, and uncertainty in FEL tasks.

In Table 1, it is evident that FER-former outperforms other state-of-the-art methods, not just for one, but for three different datasets. Specifically, FER-former achieves the highest accuracy rates of 90.96%, 91.30%, and 62.1% for FER+, RAF-DB, and SFEW 2.0, respectively. These results demonstrate the effectiveness of FER-former in recognizing facial expressions across diverse datasets, thereby establishing its superiority over existing FER methods. In Figure 18, different Transformer methods in FER, a) FER-former [62] and (b) TransFER [64], are displayed.

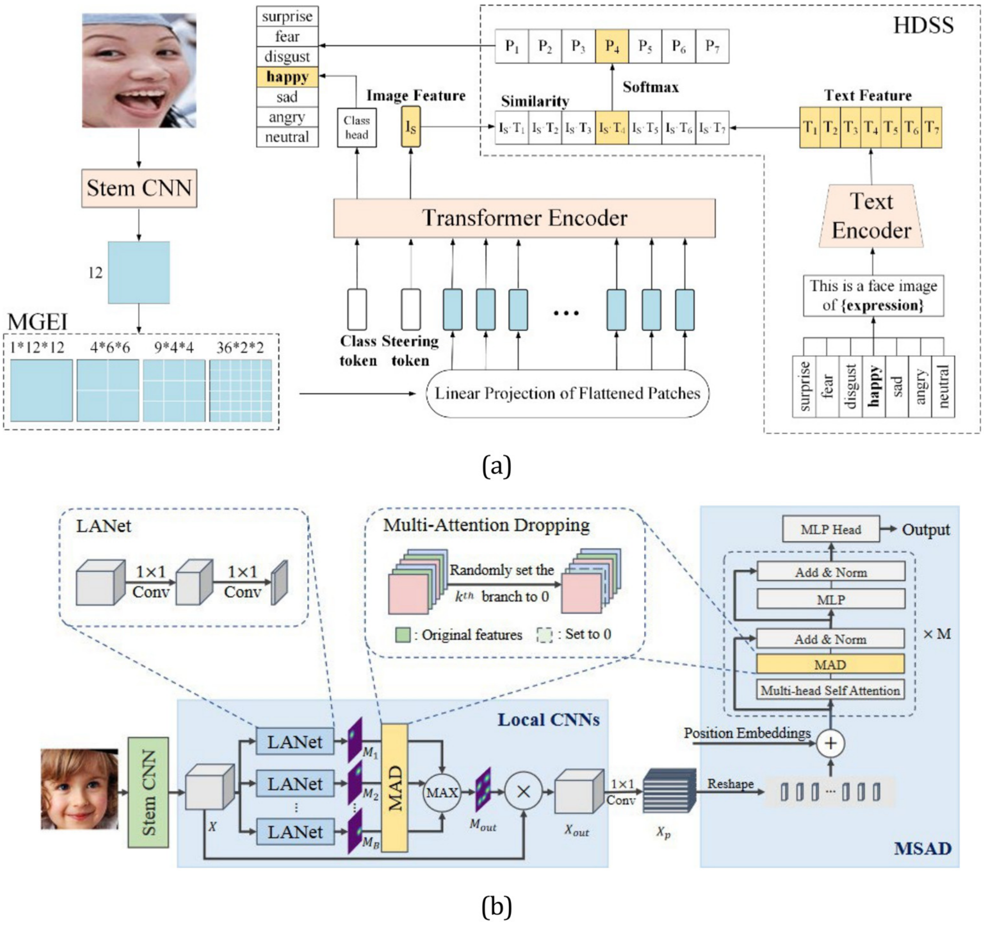


Figure 18. Transformers. (**a**) FER-former [62] and (**b**) TransFER [64]. (**a**) presents model's architecture FER-former, incorporating Multi-Granularity Embedding Integration (MGEI), hybrid self-attention, and Heterogeneous Domains-Steering Supervision (HDSS). It captures wide-ranging receptive fields and multi-modal semantics by embedding text into Multi-Head Self-Attention (MHSA), excluding convolutional position encodings for simplicity [62]. (**b**) TransFER model structure. Initially, facial images undergo processing by a foundational CNN to generate feature maps. These maps are further refined by local CNNs to identify various important feature regions. Subsequently, a 1 × 1 convolution and reshaping operations transform these maps into a sequence of feature vectors, suitable for input into the MSAD (a modified self-attention design within a Transformer encoder), which examines the interconnections among local patches. An MLP Head is then used to produce the final classification outcome. MAD directs the identification of diverse local patches, while MSAD leverages multi-head self-attention to uncover complex relationships between these patches [64].

3.2.4. Deep Sequence FER Networks

Frame Aggregation

Frame aggregation serves as a vital computational strategy in deep learning architectures, aimed at synthesizing information across multiple frames or images to enhance model performance. Traditional approaches for facial expression recognition have predominantly concentrated on static images; however, the incorporation of sequential video frames has shown potential for significant performance gains. Various methodologies have been developed to aggregate frame outputs in sequential data, with Decision-Level Frame Aggregation and Feature-Level Frame Aggregation emerging as two salient paradigms. In the interest of providing a comprehensive overview, a table summarizing the current state-of-the-art methodologies for facial expression recognition benchmarked on widely accepted datasets is presented in accordance with a person-independent evaluation protocol [152].

Decision-level frame aggregation is a technique for improving facial expression detection accuracy by combining the frames' output in a series [12,115]. In this approach, the k-class probability vectors from each frame in a sequence are amalgamated to render a consolidated output.

A noteworthy limitation of this methodology lies in the variability in the number of frames across different sequences, which precludes the direct concatenation of frame outputs. To address this challenge, two distinct aggregation procedures have been contemplated: Frame Averaging and Frame Expansion. These techniques aim to generate fixed-length feature vectors for each sequence, thereby ameliorating the issues stemming from sequence length variability.

Statistical coding, which does not require a fixed number of frames, can also be used. Each sequence's per-frame probability can be summed using vectors such as average, max, average of the square, average of maximum suppression, and others [276].

In contrast, feature-level frame aggregation involves aggregating the learned features of the frames inside the sequence.

Various statistical encoding techniques have been employed to this end, including the aggregation of the statistical moments—mean, variance, minimum, and maximum—of feature vectors across the temporal sequence. Moreover, matrix-based models such as eigenvectors, covariance matrices, and multi-dimensional Gaussian distributions offer alternative, yet sophisticated, avenues for feature aggregation. Recent advancements in the field have also explored the application of multi-instance learning to create video-level representations. In this context, cluster centers are calculated from auxiliary data, leading to a bag-of-words representation for each subset of video frames.

Kahou et al. made a notable contribution by concatenating nn-class probability vectors across ten predefined segments, thereby synthesizing a fixed-length video representation [115]. Although frame aggregation methods offer promising avenues for FER, they are not without limitations. For instance, the significance of individual frames in FER has often been overlooked. It is paramount to distinguish frames that contribute meaningful information for a particular expression from those that do not.

More specifically, they proposed a method called Frame Attention Networks (FAN) for video-based facial expression recognition. FAN consists of two modules: feature embedding and frame attention. The feature embedding module uses a deep CNN to convert face images into feature vectors. The frame attention module learns self-attention and relation-attention weights to aggregate the feature vectors into a discriminative video representation. Self-attention weights highlight important frames, while relation-attention weights model the relation between frame features and a global representation. FAN automatically identifies discriminative frames, improving performance in facial expression recognition. It is an end-to-end framework with a variable input and fixed-dimension output, enhancing video-based FER. FAN is an end-to-end framework that takes a facial video with a variable number of face images as an input and produces a fixed-dimension feature representation for FER. FAN consists of two modules: the feature embedding module and the frame attention module. Addressing this gap, Meng et al. [277] introduced the concept of Frame Attention Networks (FAN), a groundbreaking advancement in video-based FER. The architecture is bifurcated into two integral modules: the feature embedding module, which employs deep convolutional neural networks (CNNs) to transform facial images into robust feature vectors, and the frame attention module. The latter innovatively utilizes self-attention and relation-attention mechanisms to amalgamate these feature vectors into a singular, highly discriminative video-level representation. Self-attention mechanisms prioritize frames that are salient for expression recognition, while relation-attention models the interdependencies between individual frame features and a global representation. Remarkably, FAN operates as an end-to-end framework capable of processing a variable number of input frames to produce a fixed-dimension feature representation, thereby representing a significant leap forward in the adaptability and effectiveness of video-based FER solutions.

Spatiotemporal FER Network

The spatiotemporal FER network accepts a sequence of frames in a temporal window as an independent input without prior knowledge of the expression intensity and encodes more subtle expressions using both textural and temporal information.

First of all, Fan et al. [119] proposed a CNN-RNN network combined with a 3D convolutional network (C3D). This composite architecture demonstrated a superior performance compared to its individual components, thereby establishing a seminal contribution to the FER literature.

FER cannot identify facial expressions in general and can only distinguish the expressions of pre-trained human faces. Handling this difficulty necessitates the use of a dependable classifier or large datasets including many faces with varying natural dissimilarities, and several efforts have been made to offer individual-independent FER systems. To overcome this issue, Zhang et al. [120] took a different route and proposed a spatial–temporal network-based FER framework that combines a Multi-Signal Convolutional Neural Network (MSCNN) with a Part-based Hierarchical Recurrent Neural Network (PHRNN). While the MSCNN specializes in capturing the spatial features of the face, the PHRNN is designed to model the temporal variations in facial expressions. The model employs specialized loss functions to optimize the nuances of facial emotions, thereby achieving impressive accuracies of 98.5% on the CK+ dataset and 81.18% on the MMI dataset.

Hasani et al. [278] suggested a novel approach by integrating a 3D Inception-ResNet layer with an LSTM unit in a 3D Convolutional Neural Network. This architecture is particularly adept at capturing both the spatial and temporal relationships among facial images across frames, an achievement that has broad implications for the field.

A parallel innovation was brought forth by Liang et al. [41], who designed a deep convolutional BiLSTM architecture that bifurcates feature extraction into a Deep Spatial Network (DSN) and a Deep Temporal Network (DTN). These networks are specialized for spatial and temporal feature extraction, respectively, and converge into a Bidirectional LSTM (BiLSTM) for classification. Notably, their preprocessing pipeline employs a Multi-Task Cascaded Convolutional Network (MTCNN) for face detection and data augmentation, enhancing the system's robustness.

Ryumina et al. [59] introduced EmoAffectNet, a holistic framework for emotion identification that leverages a two-pronged architecture. The first component utilizes the backbone of the VGGFace2 ResNet50 model, trained in a balanced fashion to predict emotions from raw images. A temporal block is stacked atop the backbone to extend its capabilities and trained using dynamic Visual Emotion Datasets (VEDs) in a cross-corpus setting. This framework has proven its efficacy, securing a top-tier accuracy of 64% on the AffectNet dataset and ranking second in its category.

In [47], Zhu et al. proposed a cascade attention network fusion with multi-scale spatiotemporal features for facial expression recognition. The architecture leverages the ResNeXt network [279] to initially extract localized spatial features from a batch of facial images. These high-level spatial features are then meticulously stored and subsequently amalgamated with multi-scale stereoscopic features derived from an auxiliary process, thereby constructing a comprehensive geographical context for the facial expressions. This geographical context serves as the substrate for the cascade attention module, where attention aggregation features are generated. These features, in turn, are fed into a temporal sequential feature module, which is adept at isolating temporal patterns and ultimately classifying a range of facial emotions, including but not limited to anger and contempt.

In Table 1, a comprehensive performance evaluation is presented, encompassing methodologies from 60 papers across diverse datasets. The highest-performing results are denoted in boldface for ease of reference and to underscore their significance within the broader landscape of Facial Expression Recognition research. It seems that the datasets of the controlled category (especially lab group) achieved better performances than the in the wild datasets (for the controlled group, i.e., CK+ 97.25–100%, in contrast to in the wild group, i.e., AffectNet 61.07–66.46%). Also, managing controlled movie datasets presents a formidable challenge for researchers (i.e., SFW 2.0 55.15–62.18%).

Table 1. FER Methods: performance comparison.

A/A	Datasets	Method	Descriptor	Year	Performance (%)
1	CK+	[253]	DTAGN	2015	7 classes: 97.25
2		[224]	DCNN	2016	6 classes: 98.6
3		[58]	STRNN	2018	7 classes: 95.4
4		[277]	DCNN	2019	7 classes: 97
5		[41]	CNNs + BiLSTM	2020	7 classes: 99.6
6		[42]	FER-VT	2021	7 classes: **100!**
7		[43]	ViT and SE	2021	7 classes: 99.8
8		[44]	SL + SSL	2021	7 classes: 98.2
9		[63]	PAU-Net	2022	7 classes: 99.5
10	AffectNet	[280]	CAKE	2018	7 classes: 61.07
11		[281]	DSNN with a Supervised Loss function	2019	7 classes: 64
12		[37]	Emotion-GCN	2021	7 classes: **66.46**
13		[60]	Ad-Corre	2021	7 classes: 63.36
14		[54]	DACL	2021	7 classes: 65.20
15		[56]	DAN	2022	7 classes: 65.69
16		[39]	Multi-task EfficientNet-B2	2022	7 classes: 66.29
17		[59]	EmoAffectNet	2022	7 classes: 66.4
18	JAFFE	[58]	STRNN	2018	7 classes: 95.4
19		[282]	Hybrid CRNN	2018	7 classes: 94.9
20		[283]	NN	2018	7 classes: 93.8
21		[49]	Attentional Convolutional Network	2021	7 classes: 92.8
22		[275]	ARBEx	2023	7 classes: **96.67**
22	FER2013	[258]	CNN + Handcrafted features model BOVW	2019	7 classes: 75.42
23		[284]	Multi-Level Convolutional Neural Networks	2019	7 classes: 74.09
25		[285]	CNN hypeparemeters optim.	2021	7 classes: 72.16
26		[260]	VGGNet	2021	7 classes: 73.28
27		[286]	LHC	2021	7 classes: 74.42
28		[61]	Ensemble ResMaskingNet with 6 other CNNs	2021	7 classes: **76.82**
29	MMI	[287]	CNN	2015	7 classes: 93.33
30		[255]	2 CNN + 4 Inception layers	2016	7 classes: 77.9
31		[288]	CNN + CRF	2017	7 classes: 78.68
32		[259]	MBCNN-LSTM	2020	7 classes: 81.60
33		[50]	3D CNN	2021	7 classes: **96.69**
34	SFEW 2.0	[224]	DCNN	2017	7 classes: 55.15
35		[247]	IL-CNN	2018	7 classes: 52.52
36		[57]	RAN	2020	7 classes: 56.4
37		[66]	IPD-FER	2022	7 classes: 58.43
38		[69]	CRS-CONT	2022	7 classes: 60.09
39		[62]	FER-former	2023	7 classes: **62.18**

Table 1. *Cont.*

A/A	Datasets	Method	Descriptor	Year	Performance (%)
40	Oulu-CASIA	[289]	PPDN	2016	7 classes: 84.59
41		[290]	DCPN	2018	7 classes: 86.23
42		[291]	CNN	2018	8 classes: **91.67**
43		[263]	Multi-task learning CNN	2019	8 classes: 89.6
44		[48]	SCNN	2022	6 classes: 88.09
45		[47]	Cascade attention-based FER network	2022	8 classes: 89.29
46	RAF-DB	[292]	VGG-Face	2020	7 classes: 77.5
47		[293]	PSR network	2020	7 classes: 88.98
48		[294]	MixAugment	2020	7 classes: 87.54
49		[67]	EfficientFace	2021	7 classes: 88.36
50		[295]	RUL	2021	7 classes: 88.98
51		[68]	EAC	2022	7 classes: 89.99
52		[62	FER-former	2023	7 classes: **91.30**
53	AFEW 6.0	[296]	CNN–ResNet	2020	7 classes: **65.5**
54		[297]	Multi-level attention model	2020	7 classes: 55.17
55		[40]	Multi-task neural network	2021	7 classes: 59.27
56	FER+	[252]	EEM (KTN + STSN)	2020	7 classes: 90.49
57		[42]	CNN + grid-wise attention + visual transformer	2021	7 classes: 90.04
58		[64]	TransFER	2021	7 classes: 90.83
59		[65]	EASE	2022	7 classes: 90.26
60		[62]	FER-former	2023	7 classes: **90.96**

4. Facial Expression Datasets

To the best of our knowledge, this is the first attempt to study and catalogue the most well-known datasets and evaluate the performances of different state-of-the-art models on these datasets. Training neural networks with examples is a crucial factor in the success of deep learning. Researchers have access to various facial expression recognition (FER) datasets to facilitate this process. However, each dataset differs from others in terms of factors such as the number and size of images and videos and variations in illumination, population, and face pose [292]. These differences provide a diverse range of training examples for neural networks, enabling them to learn and generalize across various FER scenarios [298].

To meet the diverse needs of researchers in the field, an ideal benchmark database should include static and dynamic images of faces displaying prototypical expressions of emotion and various expressions involving the activation of single or multiple Action Units (AUs). Namely, datasets may consist of either static or dynamic images. Dynamic images or sequences contain more information than single images, including temporal information. The four categories of datasets are IC (image controlled), IU (image uncontrolled), SC (sequence controlled), and SU (sequence uncontrolled).

Controlled datasets (or lab): Controlled datasets or labs are created in environments such as research laboratories or studios. These datasets involve capturing facial expressions under well-controlled conditions, including lighting, camera angles, and backgrounds. Examples of controlled datasets include JAFFE [72,73], CK+ 76–78], and many more.

Uncontrolled datasets (or In-the-Wild): In-the-wild datasets consist of images or videos collected from real-world scenarios. These datasets present additional challenges due to lighting, occlusions, and pose variations, as well as challenges due to factors like variations in lighting, occlusions, diverse camera qualities, and pose variations. Examples of in-the-wild datasets include AffectNet [83] and Emotic [101,102], etc.

Researchers use a variety of datasets to evaluate the performance of facial expression recognition (FER) systems. These datasets cover various conditions, from controlled lab settings to challenging real-world scenarios. Using diverse datasets ensures the development of robust and accurate models that perform well across different environments and applications. Datasets are initially categorized based on the data type—images or videos. Subsequently, they are classified as controlled (lab) and uncontrolled (in the wild). The final sub-category is based on the year of the datasets creation. Datasets that belong in the image category:

1. **Image**
   (a) Controlled (Lab) JAFEE [72,73], MMI [97,98] (both in video and image), BU-3DFE [74,75], Multi-Pie [96], RAFD [89], TFD [87], Yale [79], KDEF [90,91], SFEW 2.0 [84,85], and CalD3r & MenD3s [129].
   (b) Uncontrolled (or in the wild) FER2013 [89], FER+ [71], Emotion Net [92,93], AffectNet [83], RAF-DB [94], AFEW 7.0 [80–82] ExpW [95], Aff-Wild2 [103], and Emotic [101,102].
2. **Video**
   (a) Controlled (Lab)
   (b) CK+ [76–78], MMI [97,98], Oulu-CASIA [88], MUG [99,100]. Uncontrolled (In the wild) Aff-wild2 [103].

Typically, FER experiments make use of multiple datasets, including but not limited to FER+, JAFFE, CK+, Yale, AffectNet, SFEW 2.0, FER2013, TFD, EmotioNet, Oulu-CASIA, RaFD, KDEF, RAF-DB, ExpW, MMI, and MUG. These datasets offer a rich set of facial images or videos, facilitating the training and evaluation of FER algorithms.

In Figure 19, there are examples of 12 different well-known datasets. In recent years, the use of 3D datasets, such as BU-3DFE, has become more prevalent. The creation of datasets featuring 3D faces and 4D videos represents an innovative and pioneering approach in today's research landscape [299].

This section focuses primarily on the methodologies used in database development. It outlines essential database features such as the number of samples, image resolution, subject demographics, types of emotions captured, categorization, and origin. Also, in Table 2 an overview of the facial expression datasets is illustrated.

Table 2. An overview of the facial expression datasets.

A/A	Dataset	Emotions	Year	Subjects	Samples	Origin	Resolution	Category
1	FER+ [71]	8	2016	N/A	35,887 images	N/A Microsoft	48 × 48	IU
2	JAFFE [72,73]	7	1998	10	213 static images	Japan	256 × 256	IC
3	BU-3DFE [74,75]	7	2006	100	2500 images	USA	512 × 512	IC
4	CK+ [76–78]	7	2000	123	593 image sequences	USA	640 × 490	SC
5	Yale [79]	6	2017	11	165 images	California, USA	168 × 192	IC
6	AFEW 7.0 [80–82]	7	2011	330	1809 videos	Canberra, Australia	N/A	SC
7	AffectNet [83]	7	2017	N/A	450,000 images	USA	256 × 256	IU
8	SFEW 2.0 [84,85]	7	2015	N/A	1766 images	Canberra, Australia	720 × 576	IC
9	FER2013 [86]	7	2013	N/A	35,887 images	Canada/USA	48 × 48	IU
10	TFD [87]	7	2010	N/A	112,234 images	Toronto, Canada	32 × 32	IC
11	Oulu-CASIA [88]	6	2009	80	2880 image sequences	Finland/China	320 × 240	SC
12	RAFD [89]	8	2010	67	1608 images	Netherlands	1024 × 681	IC
13	KDEF [90,91]	7	2018	70	4900 images	Sweden	562 × 762	IC

Table 2. *Cont.*

A/A	Dataset	Emotions	Year	Subjects	Samples	Origin	Resolution	Category
14	Emotion Net [92,93]	6	2016	N/A	950,000 images	Ohio, USA	Web images	IU
15	RAF-DB [94]	7	2017	N/A	29,672 images	Netherlands	N/A	IC
16	ExpW [95]	7	2018	N/A	91,793 images	Honk Kong	Web images	IU
17	Multi-Pie [96]	6	2009	337	755,370 images	USA	400 × 400	IC
18	MMI [97,98]	6	2002	75	740 images and 2900 videos	Netherlands	720 × 576	IC&SC
19	MUG [99,100]	7	2010	86	1462 sequences	Greece	896 × 896	SC
20	Emotic [101,102]	26	2019	34.320	23,751 images	N/A	N/A	IU
21	Aff-Wild2 [103]	7	2018	258	1,413,000 images and 260 videos	UK/Finland	1454 × 890	IU&SU
22	CalD3r & MenD3s [129]	7	2023	104 + 92	4678 images + 4038 images	Southern Europe/Brazil	1454 × 890	IC

For example, the Yale database includes a range of emotions such as happy, normal, sad, sleepy, surprised, and wink. Meanwhile, the Multi-Pie database features emotions like neutral, smile, surprise, squint, and disgust. The eight commonly captured emotions across datasets are neutral, contempt, anger, sadness, fear, disgust, surprise, and happiness. While 'in-lab' conditions are commonly used for FER research, 'in-the-wild' conditions pose unique challenges. These include limited sample sizes, ambiguous annotations, and unconstrained variations such as occlusions, pose differences, and variations in lighting [62]. Each of the explained datasets is briefly described below, and the categorization will be as follows:

image → Controlled, uncontrolled → Chronological order
video → Controlled, uncontrolled → Chronological order

4.1. Image

4.1.1. Controlled (or Lab)

JAFFE [72,73]

The JApanese Female Facial Expression (JAFFE) dataset (the first one which contains expressions in the lab) is one of the earliest static facial expression datasets, dating back to 1998. It comprises a collection of images depicting facial expressions posed by Japanese women. These images are accompanied by semantic ratings based on nouns that describe the various expressions. The JAFFE database is designed as a laboratory-controlled picture collection featuring 213 posed emotion examples. The dataset includes images from 10 Japanese female subjects, resulting in 213 sample images. Each of the ten subjects is presented with three or four photographs showcasing seven different facial emotions. These emotions are the six fundamental facial expressions: anger, disgust, fear, happiness, sorrow, and surprise. Additionally, each subject also has one neutral expression image. However, it is worth noting that this database can be challenging to use effectively due to the limited number of instances available per subject and expression.

MMI [97,98]

The Man Machine Interaction (MMI) database has 326 sequences from 32 subjects, with 213 sequences tagged with 6 expressions and 205 captured frontally. The participants' ages range from 19 to 62, with different ethnic backgrounds and sexes. This database has 740 images as samples and 75 subjects. MMI sequences are onset–apex–offset, starting from neutral, peaking in the middle, and returning to neutral. MMI presents more challenging situations with high interpersonal variability, non-uniform expressions, and accessories. The popular strategy for person-independent 10-fold cross-validation is to select the neutral and three peak frames in each frontal sequence.

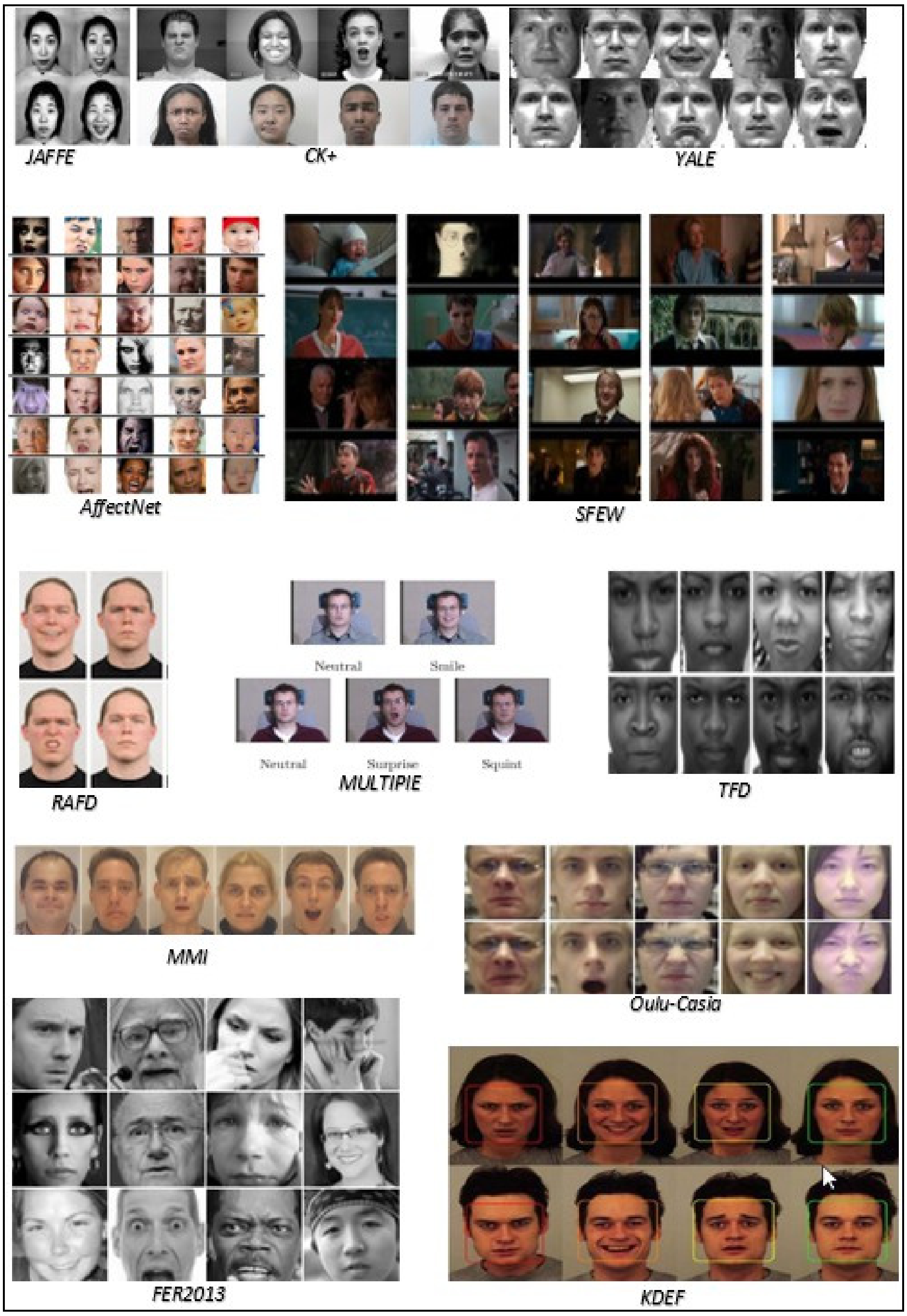


Figure 19. Examples of 12 different well-known datasets (from top to bottom and from the right to the left JAFFE [72,73], CK+ [76–78], Yale [79], AffectNet [83], SFEW 2.0 [84,85], RAFD [89], Multi-Pie [96], TFD [87], MMI [97,98], Oulu-CASIA [88], FER2013 [86], and KDEF [90,91]).

BU-3DFE [74,75]

The Binghamton University 3D Facial Expression (BU-3DFE) database is a comprehensive dataset containing 606 sequences of facial expressions from 100 participants of different genders and ethnicities. In total, it comprises 100 subjects and 2500 images. This dataset includes six universal facial expressions and seven expressions per participant, resulting in 2500 3D face expression models and 2500 two-view texture pictures. It is a valuable resource for various applications, including facial expression recognition, virtual reality, and animation, thanks to its provision of 25 immediate 3D expression models per person, offering detailed information. The dataset encompasses a diverse range of ethnic ancestries, including White, Black, East Asian, Middle Eastern Asian, Hispanic Latino, and others, with approximately 60% females and 40% males. Several approaches have been reported for 3D facial expression recognition [300].

Multi-PIE [96]

The Multi Pose, Illumination, Expressions (Multi-PIE) database contains a total of 755,370 images from 337 different individuals. The subjects in this dataset are predominantly men (235 or 69.7% vs. 102 or 30.3% females). Also, 60% of the people were Europeans and Americans, 35% were Asian, 3% were African Americans, and 2% were other. The mean age of the participants was 27.9 years. Also, this dataset offers demographic information such as gender, birth year, race, and whether the subject wears glasses.

RAFD [89]

The Radboud Faces Database (RaFD) is a high-quality "lab" database that comprises 67 models displaying eight emotional expressions. It includes Caucasian males and females, Caucasian children, and Moroccan Dutch males, totaling 67 subjects and 1608 images. This database was created by the Behavioural Science Institute of Radboud University Nijmegen and is intended for non-commercial scientific studies by accredited universities. Each model is trained to show eight emotional expressions, represented by three look directions and captured from five camera positions. The emotions include anger, disgust, fear, happiness, sadness, surprise, contempt, and neutrality.

Yale Face Database [79]

The Yale Face Database consists of 165 grayscale GIF images featuring 15 individuals. Within this database, there are 11 photos available for each individual, each corresponding to a specific facial expression. These expressions include happiness, neutrality, sadness, drowsiness, surprise, and winking. In total, the database comprises 11 subjects and 165 images.

TFD [87]

The Toronto Face Database (TFD) is a collection of facial expression datasets comprising a total of 112,234 photos. These facial images are normalized to 48 × 48 pixels, with the eyes equidistant and aligned at the same vertical coordinate. TFD introduces a partitioning scheme involving five folds, dividing the data into training, validation, and test sets at allocations of 70%, 10%, and 20% of the frames, respectively. Although the database contains a vast number of images (112,234 images), it unfortunately lacks specific details about the subjects. The dataset includes grayscale images of faces, each measuring 32 × 32 pixels, and is categorized into seven distinct emotional expressions. The dataset is further divided into labeled, public test, hidden test, and unlabeled subsets. Illustrations of faces displaying different expressions can be observed in Figure 19.

KDEF [90,91]

The Karolinska Directed Emotional Faces (KDEF) comprises a total of 4900 images showcasing various human facial emotions. These photographs portray 70 individuals, each expressing seven distinct emotional states. The database includes 4900 images and involves a total of 70 subjects. Each emotional expression is captured from five unique perspectives, providing a comprehensive view of facial expressions. Among the 70 individuals included, there are 35 females and 35 males, all within the age range from 20 to 30 years. Notably, the subjects do not have beards, mustaches, earrings, or eyeglasses, and efforts have been made to minimize visible makeup.

CalD3r & MenD3s [129]

The paper presents the CalD3r and MenD3s databases, featuring spontaneous 2D and 3D facial expressions from diverse young adults, elicited through validated affective images. These databases aim to enhance ecological validity in facial expression recognition, categorizing emotions into seven types and confirming them via participants' self-reports. This marks a significant advancement in creating more natural, real-world applicable emotion recognition systems.

4.1.2. Uncontrolled (or in the Wild)

FER2013 [86]

The FER2013 database, which was the first to include facial expressions "in the wild", compiles approximately 35,887 facial images originally utilized in the ICML 2013 Challenges in Representation Learning. These images were sourced from Google image search API, registered, shrunk to 48 × 48 pixels, and categorized into seven expressions: surprise, anger, fear, disgust, happiness, sadness, and neutrality. It should be noted that, although the database contains 35,887 images, there are no specific details provided regarding individual subjects. The dataset is divided into three categories: training, validation, and testing. The emotions most frequently represented in the dataset are happiness, anger, sadness, neutrality, fear, surprise, and disgust. This dataset has significantly contributed to the advancement of research in facial expression recognition and continues to be a valuable resource for researchers in this field.

FER+ [71]

The FER+ dataset is an extension of the original FER2013 dataset, initially compiled by Pierre Luc Carrier and Aaron Courville through the web crawling of facial images associated with emotion-related keywords. In this extended dataset, the images are meticulously labeled into one of eight distinct emotion categories: neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt. This database has 35,887 images, but similar to the original FER dataset, there are no specific details provided about individual subjects. Each image in FER+ was tagged by ten crowd-sourced taggers, resulting in a higher-quality ground truth for capturing the emotions depicted in still images compared to the original FER labels. Researchers working with the FER+ dataset can leverage these ten taggers to estimate an emotion probability distribution for each facial image, enabling the development of algorithms that can generate statistical distributions or SFEW 2.0 [84,85].

The Static Facial Expressions in the Wild (SFEW) was generated by choosing static frames from the AFEW database and calculating keyframes based on facial point clustering retrieved from movies. SFEW 2.0, the most widely used version, contained the benchmarking data for the SReco sub-challenge in EmotiW 2015. This database contains 1766 images, but there are no details about the subjects. SFEW 2.0 is divided into three sets: training (958 samples), validation (436 samples), and testing (372 samples). Each image is labeled with one of seven emotions: surprise, happiness, disgust, fear, neutral, anger and contempt. While the expression labels for the training and validation sets are available to the public, those for the testing set are available upon request.

Emotion Net [92,93]

The EmotioNet database is a vast collection comprising 950,000 annotated images focusing on Action Units (AU) (an Action Unit identifies facial muscle movements associated with expressions. The recognition of AUs helps to analyze facial expressions and infer emotions using computer vision and machine learning techniques)6. EmotioNet is a large-scale database containing one million photos of facial expressions gathered online. Of the 1 million images, 950,000 were tagged automatically using an AU detection algorithm, while the remaining 25,000 images underwent manual annotation, encompassing 11 AUs. It is worth noting that this database does not include specific details about the individuals represented in the images. Additionally, the EmotioNet Challenge includes a "second track" featuring six basic expressions, ten compound expressions, and a subset of 2478 photos with expression labels.

AffectNet [83]

AffectNet is a vast dataset comprising over 1 million facial images meticulously collected and annotated to create a database of facial emotions observed in real-world scenarios. The dataset is derived from images found on the internet using 1250 emotion-related keywords. These keywords were combined with terms related to gender, age, or ethnicity to obtain nearly 362 strings in the English language, such as "joyful girl", "blissful Spanish man", "furious young lady", or "astonished senior" in six languages and collected using three major search engines (Google, Bing, and Yahoo). About half of the retrieved images were manually annotated for the presence of seven discrete facial expressions and the intensity of valence and arousal. Approximately half of the retrieved images underwent manual annotation, which included the identification of seven discrete facial expressions and the quantification of intensity for valence and arousal. The remaining half of the images were annotated using a distribution model rather than the traditional category model. The database contains 450,000 images, and it is worth noting that specific details about the individuals represented in the images are not provided.

AffectNet empowers research on automated facial expression identification by offering two alternative emotion models. Researchers can utilize two baseline deep neural networks to categorize images and predict the strength of valence and arousal. These models have demonstrated superior performances when compared to traditional machine learning approaches and commercially available facial expression recognition systems.

B [94]

The Real-world Affective Faces DataBase (RAF-DB) is an extensive facial expression database consisting of around 30,000 (29,762 images) diverse facial images downloaded from the Internet. Using crowd-sourcing annotation, all the images were independently labeled by approximately 40 annotators. This dataset encompasses seven different emotion labels and includes approximately 30,000 facial images from thousands of individuals. Notably, the database does not provide specific details about the subjects featured in the images.

ExpW [95]

The Expression In-the-Wild (ExpW) dataset is designed for facial expression recognition and comprises 793 faces that are manually labeled with expressions. Each of the face photos is labeled with one of the seven fundamental expression categories: angry, disgust, fear, glad, sad, surprise, or neutral. This dataset includes 91,793 images, and specific details about the subjects featured in the images are not provided.

Emotic [101,102]

The EMOTion In Context (EMOTIC) is a database of images with people in real environments, annotated with their apparent emotions. The images are annotated with an extended list of 26 emotion categories combined with the three common continuous dimensions, Valence, Arousal, and Dominance. This database has 23,751 images, but there are no details about the subjects.

Aff-Wild2 [103]

Aff-Wild2, an extension of the Aff-Wild database, is a large-scale dataset for in-the-wild facial expressions. It contains 1,413,000 video frames and includes variations in pose, age, illumination conditions, ethnicity, and profession. The dataset is annotated for various facial expressions, encompassing both basic emotions (such as happiness, sadness, anger, and disgust) and complex emotions (like boredom, confusion, and surprise). While the database contains 23,751 images, it does not provide details about the subjects.

In [267], the focus of the ABAW Competition is on developing algorithms for the automatic analysis of affect, including facial expressions. This competition includes several challenges relevant to facial expression recognition: Valence–Arousal Estimation, Expression Classification, and Action Unit Detection. It aims to promote research and development in the field of affective computing. Additionally, the third Affective Behavior Analysis in-the-Wild (ABAW) Competition was held in conjunction with the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) in 2022. This competition builds upon previous events held at ICCV 2021, IEEE FG 2020, and IEEE CVPR 2017, and it also aims to advance the automatic analysis of affect. Overall, both the ABAW Competition and the Aff-Wild2 database are essential resources for researchers working on facial expression recognition. They highlight the growing interest in this specific area of affective computing.

4.2. Video

4.2.1. Uncontrolled (or Lab)

CK [76]

The Cohn-Kanade (CK) dataset was released in 2000 to promote research into the automatic detection of individual facial expressions. Since its release, the CK dataset has become the most widely used for evaluating and developing algorithms. It consists of 100 university students aged from 18 to 30 years. Of the total population, 65% are female, 15% are African American, and 3% are Asian or Latino, with the remaining 17% falling into other categories. The videos capture seven different facial expression. The database contains 1766 images, but there are no specific details about the subjects.

CK+ [76–78]

The Extended Cohn-Kanade (CK+) dataset is a laboratory-controlled database that consists of 593 video segments from 123 individuals of diverse genders and cultural backgrounds. Each video captures a face transitioning from neutral expression to a peak emotion, recorded at 30 FPS with a resolutions of either 640 × 480 or 640 × 490 pixels. The dataset includes 327 categorized sequences across seven expression classes: anger, contempt, disgust, fear, happiness, sorrow, and surprise. The absence of specified training, validation, and test sets complicates the comparison of algorithms evaluated using this dataset. The database contains 593 videos, but is based on 123 subjects.

Oulu-CASIA [88]

The Oulu-Chinese Academy of Science Institute of Automation (Oulu-CASIA) NIR & VIS facial expression database contains six expressions (surprise, happiness, sorrow, anger, fear, and disgust) from 80 participants, ranging in age from 23 to 58 years. Notably, 73.8% of the subjects are male. This database comprises a total of 2880 videos. Each participant was instructed to sit in front of the camera on a chair in the observation room. The camera-to-face distance was around 60 cm. They were instructed to mimic facial expressions based on example photo sequences provided to them. The imaging hardware used for this study operates at a frame rate of 25 frames per second and has a resolution of 320 × 240 pixels.

MUG [99,100]

The Multimedia Understanding Group (MUG) database comprises 86 sequences of images featuring 35 white women and 51 Caucasian men aged between 20 and 35 years. This database consists of 1462 videos and involves 86 subjects, each seated in front of a camera. They were instructed to perform six basic facial expressions (happiness, sadness, anger, fear, surprise, and disgust), along with experiencing laboratory-induced emotions. The database is divided into two parts. The first part consists of categorically labeled image sequences, each with varying lengths containing from 50 to 160 images. The second part involves recording participants while they watched a film designed to elicit emotions. The primary objective of this database is to provide ample material for recognizing genuine expressions in real-world scenarios.

AFEW 7.0 [80–82]

The Acted Facial Expressions In The Wild (AFEW) 7.0 dataset was constructed by various Australian universities and research organizations. Comprising 1809 videos and featuring 330 subjects, this dataset is an invaluable asset for research on dynamic facial expressions. Since 2013, it has served as the evaluation benchmark for the annual Emotion Recognition in The Wild Competition (EmotiW) (Emotion Recognition in The Wild Competition). AFEW includes video excerpts capturing spontaneous facial expressions under varied environmental conditions and is annotated with seven different expressions. The dataset's annotation has been continually updated to include data from reality TV shows. For the 2017 EmotiW challenge, AFEW was divided into three separate data partitions. Overall, the collection provides a rich array of real-world, or near real-world, facial expressions across diverse settings and scenarios.

4.2.2. Uncontrolled (or in the Wild)

The AffWild2 database [103] also includes video samples, totaling 260 videos.

5. Comparison Dataset and Methods

Below are some diagrams that approximate the number of open-access papers published in various fields, as well as the number of papers on Facial Expression Recognition (FER), that specifically mention certain datasets, for the years from 2018 to 2023. In Figure 20, the data reveal that CK+ was the most frequently used dataset in 2018. In 2019, AffectNet surpassed CK+ to take the lead. During 2020, both CK+ and AffectNet were commonly utilized. For the years 2021 and 2022, AffectNet regained its position as the most popular dataset. As of 2023, AffectNet continues to hold this status. In summary, over the six years from 2018 to 2023, AffectNet has been cited in 207 open-access papers. CK+ follows it with 144 citations, FER2013 with approximately 103, RAF-DB with 97, FER+ with 79, JAFFE with 66, Oulu-CASIA with 55, SFEW with 43, MMI with 40, and AFEW with 6, among others. These data were sourced from https://paperswithcode.com/ accessed on 1 February 2023.

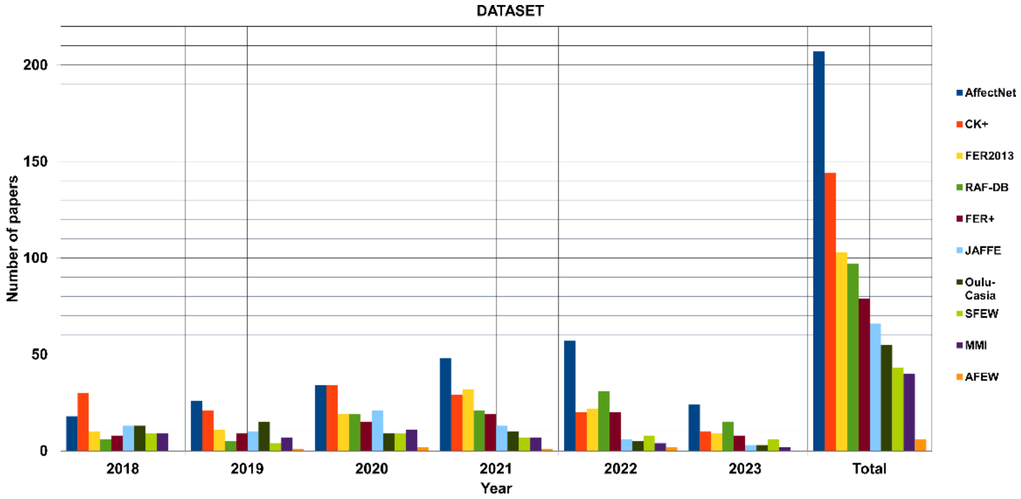


Figure 20. Number of papers per year (https://paperswithcode.com/ accessed on 1 February 2023).

Many models have been evaluated for performance, but some stand out notably. Initially, FER-former [62] emerges as one of the best models, leading as superior with three different datasets. Specifically, it achieved an accuracy of 90.96% on FER+, outperforming the second-best model with 90.25% [65]. On RAF-DB, it scored 91.3%, a 1.3% lead over the second-best model at 89.99% [68]. On SFEW 2.0, it reached 62.18%, which is 2.09% higher than the next best model [69]. This model stands out due to its incorporation of multi-modal fusion, a hybrid stem that combines CNNs and Transformers, FER-specific transformer mechanisms, heterogeneous domain-steering supervision, and multifarious token heads, and its demonstrated superior performance in facial expression recognition tasks.

Another state-of-the-art model is **FER-VT** [42], with the best accuracy on the CK+ dataset, an astonishing 100% accuracy without the need for extra training data. FER-VT introduces a novel framework for CNN-based Facial Expression Recognition (FER), incorporating two attention mechanisms at both the low-level feature learning and high-level semantic representation stages. By selectively focusing on important facial regions, FER-VT enhances the model's ability to capture discriminative features and extract meaningful representations. This framework addresses limitations in traditional CNN-based FER models and improves accuracy in recognizing facial expressions. This model distinguishes itself through its Grid-wise Attention Mechanism, which captures dependencies between facial regions and regulates parameter updates during low-level feature learning, thereby improving feature representation and discrimination. Furthermore, its Visual Transformer Attention Mechanism uses visual semantic tokens from high convolutional layer blocks to capture complex relationships and semantic cues in facial expressions. This contributes to the model's superior accuracy in recognizing facial expressions. The transformer mechanisms in FER-VT provide it with a distinct advantage over other models.

A noteworthy third model is TransFER [64], which comprises three key components: Multi-Attention Dropping (MAD), ViT-FER, and Multi-head Self-Attention Dropping (MSAD). Notably, the ViT (Vision Transformer) architecture plays a pivotal role in this model's exceptional performance.

One of the reasons why this model outperforms others is its emphasis on relation-aware representation learning. TransFER significantly focuses on learning relation-aware representations, a crucial aspect in Facial Expression Recognition (FER). This approach enables the model to capture dependencies and correlations among facial regions, enhancing its ability to understand and discriminate between facial expressions, ultimately improving performance. Additionally, TransFER leverages attention mechanisms, such as Multi-Attention Dropping (MAD) and Multi-head Self-Attention Dropping (MSAD). These mechanisms are instrumental in capturing fine-grained details and facilitating the exploration of diverse relationships among local facial patches. They prove highly effective in enhancing feature learning and discrimination within the FER context. Lastly, the model incorporates global–local Fusion, which combines the strengths of both global and local facial expression cues. This holistic approach enhances the model's capacity to capture comprehensive information and context, improving performance. Regarding its performance, TransFER achieves impressive accuracy scores of 90.83% on the FER+ dataset, 90.91% on RAF-DB, and 66.23% on Affectnet. These results position TransFER as the second-best performer in our survey across these datasets. For all the reasons mentioned, it is undoubtedly worth referencing this model.

A fourth model that merits mentioning is **Emotion-GCN [37]**. This state-of-the-art model boasts the highest performance, with an accuracy of 66.46% on the AffectNet dataset. The framework employs a Graph Convolutional Network (GCN) to capture the interdependencies between the categorical and dimensional models of affect. It learns a shared feature representation for both discrete and continuous recognition tasks. Additionally, the GCN is used to explicitly model the dependencies between facial expression classifiers and valence–arousal regressors.

Furthermore, STRNN [58] is a hybrid Convolutional-Recurrent Neural Network (CNN-RNN). There are several reasons why this model stands out as a state-of-the-art approach. First, the paper proposes a hybrid architecture that combines the strengths of Convolutional Neural Networks (CNNs) for spatial feature extraction with Recurrent Neural Networks (RNNs) for capturing temporal dependencies in facial expressions. This combination allows for a more comprehensive understanding of facial expressions, thereby improving accuracy. Second, the model takes into account the temporal dynamics present in facial expressions. Facial expressions are not static, they involve temporal variations. By modeling these temporal dependencies, the STRNN method can capture the evolution of expressions over time, leading to more accurate recognition. Collectively, its hybrid architecture and consideration of temporal dependencies position this paper as one of the state-of-the-art methods in Facial Expression Recognition.

Another noteworthy state-of-the-art model that has to be emphasized is the **AR-BEx** model [275]. This paper significantly enhances the performance of the cross-attention window-based Vision Transformer (ViT) in generating feature embeddings. The approach effectively tackles challenges related to inter-class similarity, intra-class disparity, and label ambiguity in facial expression recognition tasks. The model achieves robust representations through extensive augmentation and data refinement, enabling it to capture crucial variations in facial expressions. Furthermore, the window-based ViT's cross-attention mechanism improves the model's capacity to focus on relevant facial regions and attend to vital features. This integrated approach, overall, enhances the discriminative power and reliability of the feature embeddings, leading to more accurate and resilient facial expression recognition. It is undeniable that this model excels on the JAFFE dataset, achieving an impressive performance score of 96.67%. What makes this model truly special is its reliability balancing strategy—an approach designed to mitigate challenges stemming from partial and unbalanced data in facial expression learning (FEL). This strategy not only enhances the reliability of the model's predictions, but also boosts its overall performance. Additionally, throughout the paper, the transformer mechanism consistently contributes to an improved performance, making the ARBEx model a noteworthy advancement in the field of facial expression recognition.

A seventh model to take into consideration is the multi-rate **3DCNN** [50], which is based on a multi-rate signal processing scheme. This model employs 3D Convolutional Neural Networks (3DCNN) to harness feature maps from multiple frames as inputs, enabling the extraction of motion information. In this approach, the 3D CNN utilizes feature maps from multiple frames as inputs, allowing for the extraction of valuable motion information. The model consists of three different 3DCNN networks, each producing distinct feature vectors. Additionally, it leverages SENet, which incorporates self-attention mechanisms to recalibrate features by aggregating feature maps across spatial dimensions. The resulting emphasized features are then fed into a joint fusion classifier for emotion classification.

State-of-the-art models in facial expression recognition exhibit certain characteristics that contribute to their superior performance. These models often employ hybrid architectures, attention mechanisms, and fusion strategies. Hybrid architectures combine different networks, allowing for improved feature extraction, representation, and discrimination. Attention mechanisms, such as self-attention or grid-wise attention, selectively focus on important features or regions, capturing dependencies and semantic cues. Multi-modal fusion combines various sources of information, enhancing the model's understanding and recognition abilities by capturing complementary cues. Incorporating transformers enables state-of-the-art models to capture complex relationships, model dependencies, and extract meaningful representations, providing an advantage over other deep learning models and improving performance. Furthermore, models that consider temporal dependencies by incorporating recurrent neural networks or hybrid architectures capture the temporal dynamics of facial expressions, resulting in more accurate recognition. Techniques such as feature calibration and reinforcement, global–local fusion, and addressing specific challenges in facial expression recognition also contribute to the state-of-the-art performance of these models.

6. Discussion and Conclusions

Over the last decade, Facial Expression Recognition (FER) has garnered increasing attention from the scientific community. This paper serves as a comprehensive review of FER, starting with the methods developed in the pre-deep learning era. These initial methods primarily focused on facial feature extraction and emotion classification, the two most crucial steps in traditional FER systems. Moreover, this paper emphasizes the deep learning era, exploring its historical context, methodologies, and advancements. It extensively reviews state-of-the-art papers, offering a comprehensive overview of the latest research and developments in the field. Additionally, the paper presents a brief survey of the most commonly used FER datasets, detailing their respective characteristics, such as the types of emotions captured, the year of creation, the subjects involved, and their geographical origins. Despite the significant advancements and growing interest in FER, many unanswered questions and challenges remain. As we move forward, future research should consider addressing the following questions:

(1) How do human beings correctly identify facial expressions?

The mechanism behind the human identification of facial expressions remains not fully understood. It is a multi-faceted process involving the visual analysis of facial features, learned emotional associations, and innate biological responses. This intricate interaction of perception-acquired understanding and inherent biological mechanisms calls for further investigation.

(2) How can we make the models even faster than before?

While neural networks have significantly accelerated model performance and enhanced accuracy, the question arises: can we achieve even better? As technology evolves, we can expect to encounter models yielding increasingly efficient outcomes. One promising avenue is the development of hybrid models, which combine traditional methods with deep learning techniques. These hybrid approaches have shown promise in delivering improved results [301,302].

(3) How could we integrate facial expression analysis with other modalities?

The incorporation of multiple modalities, such as speech, facial recognition, and physiological and lexical data, has recently gained traction in the field, resulting in what is known as fusion models [303]. These models tackle the challenging task of harmonizing heterogeneous modalities for enhanced emotion expression recognition. For instance, in the Affective Behavior Analysis in-the-wild Competition, the Multi-Task-Learning challenge calls for the development of algorithms that can simultaneously perform multiple affective analysis tasks [267]. This concept mirrors multi-task learning in machine learning and could be considered as analogous to human cognitive multitasking. By advancing more efficient and practical algorithms for real-world emotional analysis, researchers may improve our understanding of human emotions and behaviors, thereby shedding light on the first question raised in this discussion.

(4) Will we see more models and research focused on thermal infrared imaging in the coming years?

Compared to facial expression recognition in the visible spectrum, the thermal infrared spectrum has received relatively less attention in research [304]. However, as technology advances and the need for more robust and versatile recognition systems grows, it is plausible that thermal infrared imaging will gain more prominence in future research endeavors.

(5) Is it possible to compress a deep model to reduce the large network parameters of deep models?

Deep learning models, particularly those with numerous layers, incur significant computational costs due to their extensive parameter count, leading to storage and training time challenges. Exploring advanced model compression techniques is essential to enhancing efficiency by reducing both the computational resources and time needed for model deployment and data processing.

(6) Is it possible to develop multimodal systems with a high performance on "in-the-wild" data?

While multimodal systems have demonstrated strong performances in controlled laboratory settings, they still face challenges in achieving ecological validity when applied to real-world "in-the-wild" data.

(7) It is observed that increasing the complexity of the models can lead to better results.

Deep learning models, such as deep neural networks, achieve this by learning hierarchical representations of data through the stacking of multiple layers of interconnected nodes or neurons. In the future, we can anticipate the emergence of increasingly complex classifiers that have the potential to yield even better results.

(8) Variation-free datasets limit deep learning models' effectiveness and adaptability.

Datasets that lack variations, such as occlusion, illumination, and resolution challenges, may seem ideal for training deep learning models. However, these variations enable deep learning models to generalize effectively, handle real-world scenarios, prevent overfitting, improve representation learning, and utilize their capacity optimally. Encompassing these challenges, training on diverse data empowers models to extract relevant features, patterns, and complexities, enhancing their performance, robustness, and ability to handle unseen instances. Removing variations curtails a model's capacity to adapt and generalize in different scenarios, ultimately impeding its real-world applicability. In summary, variations in data, even with inherent challenges, are indispensable for deep learning models to learn effectively and achieve superior results.

(9) How can we tackle ethical and social concerns associated with FER systems, including privacy, bias, and potential misuse?

Facial Expression Recognition (FER) systems raise ethical and social concerns regarding privacy, bias, and potential misuse. Privacy concerns arise when FER systems are employed for surveillance without consent, highlighting the need for transparent data collection practices and obtaining informed consent. Biased training datasets can lead to inaccurate results and perpetuate discrimination, making it crucial to ensure diverse and representative data collection. Additionally, the misuse of FER systems, such as emotional manipulation or discriminatory actions, must be prevented by establishing guidelines and regulations. It is essential to address these concerns to promote the ethical and responsible development and use of FER systems, safeguarding privacy, mitigating bias, and preventing potential misuse.

Author Contributions: Conceptualization, N.V.; methodology, N.V. and V.S.; writing—original draft preparation, T.K.; writing—review and editing, N.V. and V.S. and P.D.; supervision, P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the H2020 European Project affecTive basEd iNtegrateD carE for betteR Quality of Life TeNDER: TeNDER. Grant no. 875325.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Each dataset presented in this paper is referenced to the corresponding paper where the dataset was created.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

IC	Image Controlled
IU	Image Uncontrolled
SC	Sequence Controlled
SU	Sequence Uncontrolled
CVPR	Computer Vision and Pattern Recognition
ABAW	Affective Behavior Analysis in-the-Wild
AU	Action Unit
RNN	Recurrent Neural Network
ACN	Attentional Convolutional Network
DBN	Deep Belief Network
CNN	Convolutional Neural Network
ViT	Vision Transformer
SE	Squeeze and Excitation

SL	Supervised Learning
MSCNN	Multi-Signal CNN
PHRNN	Partial Based Hierarchical Bidirectional RNN
SSL	Self-Supervised Learning
DAN	Distract your Attention Network
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CAKE	Compact and Accurate K-dimensional representation of Emotion
DTAGN	Deep Temporal Appearance-Geometry Network
DTGN	Deep Temporal Geometry Network
DTAN	Deep Temporal Appearance Network
LBP	Local Binary Pattern
PReLU	Parametric Rectified Linear Unit
DCT	Discrete Cosine transform
PCA	Principal Component Analysis
PPDN	Peak-Piloted Deep Network
FER	Facial expression recognition
RAN	Region Attention Network
ARBEx	Attentive feature extraction with Reliability Balancing for robust fac1i7a7l0 Exp
cGAN	Conditional GAN
SAE	Sparse Autoencoder
VGG	Visual Geometry Group
GCN	Graph Convolutional Network
DTAGN	Deep Temporal Appearance-Geometry Network
DDMTL	Discriminative Deep Multi-Tasking Learning
3DCNN + DAP	3D CNN + Deformable Action Parts
DBN	Deep Belief Network
AU	Action Unit
DeRL	De-expression Residue Learning
SCNN	Sparse Autoencoder and CNN
DL	Deep Learning
RBM	Restricted Boltzmann Machines
PDM	Point Distribution Model
RUL	Relative Uncertain Learning
RF	Random Forest
AAM	Active Appearance Model
SDM	Supervised Descent Method
MoT	Mixtures of Trees
SCNN	Sparse Autoencoder and CNN
LDA	Linear Discriminant Analysis
GRU	Gated Recurrent Units
BPTT	Back Propagation Through Time
GAN	Generative Adversarial Network
STC-NLSTM	Spatio-Temporal Convolutional features with Nested LSTM
IFRP	Identity-Preserving Face Recovery from Portraits
SIFT	Scale-Invariant Feature Transform
FC	Fully Connected
MRE-CNN	Multi-Region Ensemble CNN
DLP-CNN	Deep Locality-Preserving CNN
AT-GCN	Attention Enhanced Temporal GCN
DeRL	De-expression Residue Learning
RMN	Residual Masking Network
BiLSTM	Bidirectional LSTM
WMCNN-LSTM	Weighted Mixture Convolutional Neural Network–LSTM

References

1. Pantic, M.; Rothkrantz, L. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]

2. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
3. Pantic, M.; Rothkrantz, L.J. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2004**, *34*, 1449–1461. [CrossRef] [PubMed]
4. Li, S.Z.; Jain, A.K.; Tian, Y.L.; Kanade, T.; Cohn, J.F. Facial Expression analysis. In *Handbook of Face Recognition*; Springer: New York, NY, USA, 2005; pp. 247–275.
5. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
6. Bettadapura, V. Face expression recognition and analysis: The state of the art. *arXiv* **2012**, arXiv:1203.6722.
7. Konar, A.; Chakraborty, A. *Emotion Recognition: A Pattern Analysis Approach*; John Wiley & Sons: Hoboken, USA, 2015.
8. Zhao, X.; Zhang, S. A review on facial expression recognition: Feature extraction and classification. *IETE Tech. Rev.* **2016**, *33*, 505–517. [CrossRef]
9. Martinez, B.; Valstar, M.F.; Jiang, B.; Pantic, M. Automatic analysis of facial actions: A survey. *IEEE Trans. Affect. Comput.* **2017**, *10*, 325–347. [CrossRef]
10. Azizan, I.; Khalid, F. Facial emotion recognition: A brief review. In Proceedings of the International Conference on Sustainable Engineering, Technology and Management (ICSETM-2018), Karnataka, India, 19–20 April 2018; Volume 20.
11. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* **2018**, *18*, 416. [CrossRef] [PubMed]
12. Wei, H.; Zhang, Z. A survey of facial expression recognition based on deep learning. In Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 9–13 November 2020; pp. 90–94. [CrossRef]
13. Mellouk, W.; Handouzi, W. Facial emotion recognition using deep learning: Review and insights. *Procedia Comput. Sci.* **2020**, *175*, 689–694. [CrossRef]
14. Patel, K.; Mehta, D.; Mistry, C.; Gupta, R.; Tanwar, S.; Kumar, N.; Alazab, M. Facial sentiment analysis using AI techniques: State-of-the-art, taxonomies, and challenges. *IEEE Access* **2020**, *8*, 90495–90519. [CrossRef]
15. Revina, I.M.; Emmanuel, W.S. A survey on human face expression recognition techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 619–628. [CrossRef]
16. Dang, V.T.; Do, H.Q.; Vu, V.V.; Yoon, B. Facial Expression Recognition: A Survey and its Applications. In Proceedings of the 2021 23rd International Conference on Advanced Communication Technology (ICACT), Pyeong Chang, Republic of Korea, 7–10 February 2021; pp. 359–367. [CrossRef]
17. Jampour, M.; Javidi, M. Multiview Facial Expression Recognition, A Survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2086–2105. [CrossRef]
18. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A.C. A Survey on Facial Emotion Recognition Techniques: A State-of-the-Art Literature Review. *Inf. Sci.* **2022**, *582*, 593–617. [CrossRef]
19. Khan, A.R. Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. *Information* **2022**, *13*, 268. [CrossRef]
20. Dujaili, M.J.A. Survey on facial expressions recognition: Databases, features and classification schemes. *Multimed. Tools Appl.* **2023**, *83*, 7457–7478. [CrossRef]
21. Cai, Y.; Li, X.; Li, J. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors* **2023**, *23*, 2455. [CrossRef]
22. Dulguerov, P.; Marchal, F.; Wang, D.; Gysin, C. Review of objective topographic facial nerve evaluation methods. *Am. J. Otol.* **1999**, *20*, 672–678.
23. Stanković, M.; Nešić, M.; Obrenović, J.; Stojanović, D.; Milošević, V. Recognition of facial expressions of emotions in criminal and non-criminal psychopaths: Valence-specific hypothesis. *Personal. Individ. Differ.* **2015**, *82*, 242–247. [CrossRef]
24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
25. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
26. Abdat, F.; Maaoui, C.; Pruski, A. Human-computer interaction using emotion recognition from facial expression. In Proceedings of the 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, Madrid, Spain, 16–18 November 2011; pp. 196–201.
27. Hickson, S.; Dufour, N.; Sud, A.; Kwatra, V.; Essa, I. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In Proceedings of the 2019 Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1626–1635.
28. Chen, C.H.; Lee, I.J.; Lin, L.Y. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Res. Dev. Disabil.* **2015**, *36*, 396–403. [CrossRef]
29. Assari, M.A.; Rahmati, M. Driver drowsiness detection using face expression recognition. In Proceedings of the 2011International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, 16–18 November 2011; pp. 337–341.
30. Zhan, C.; Li, W.; Ogunbona, P.; Safaei, F. A real-time facial expression recognition system for online games. *Int. J. Comput. Games Technol.* **2008**, *2008*, 542918. [CrossRef]
31. Allied-Market. Available online: https://www.alliedmarketresearch.com/emotion-detection-and-recognition-market/ (accessed on 1 February 2023).

32. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef] [PubMed]
33. Ekman, P.; Friesen, W.V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*; Ishk: Los Altos, CA, USA, 2003; Volume 10.
34. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
35. Matsumoto, D. More evidence for the universality of a contempt expression. *Motiv. Emot.* **1992**, *16*, 363–368. [CrossRef]
36. Ekman, P.; Cordaro, D. What is meant by calling emotions basic. *Emot. Rev.* **2011**, *3*, 364–370. [CrossRef]
37. Antoniadis, P.; Filntisis, P.P.; Maragos, P. Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition. In Proceedings of the 2021 16th International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021. [CrossRef]
38. Zangeneh Soroush, M.; Maghooli, K.; Setarehdan, S.; Motie Nasrabadi, A. Emotion classification through nonlinear EEG analysis using machine learning methods. *Int. Clin. Neurosci. J* **2018**, *5*, 135–149. [CrossRef]
39. Savchenko, A.V.; Savchenko, L.V.; Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2132–2143. [CrossRef]
40. Savchenko, A.V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In Proceedings of the 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 16–18 September 2021; pp. 119–124.
41. Liang, D.; Liang, H.; Yu, Z.; Zhang, Y. Deep convolutional BiLSTM fusion network for facial expression recognition. *Vis. Comput.* **2020**, *36*, 499–508. [CrossRef]
42. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [CrossRef]
43. Aouayeb, M.; Hamidouche, W.; Soladie, C.; Kpalma, K.; Seguier, R. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv* **2021**, arXiv:2107.03107.
44. Pourmirzaei, M.; Montazer, G.A.; Esmaili, F. Using Self-Supervised Auxiliary Tasks to Improve Fine-Grained Facial Representation. *arXiv* **2021**, arXiv:2105.06421.
45. Zheng, H.; Wang, R.; Ji, W.; Zong, M.; Wong, W.K.; Lai, Z.; Lv, H. Discriminative deep multi-task learning for facial expression recognition. *Inf. Sci.* **2020**, *533*, 60–71. [CrossRef]
46. Foggia, P.; Greco, A.; Saggese, A.; Vento, M. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Eng. Appl. Artif. Intell.* **2023**, *118*, 105651. [CrossRef]
47. Zhu, X.; He, Z.; Zhao, L.; Dai, Z.; Yang, Q. A Cascade Attention Based Facial Expression Recognition Network by Fusing Multi-Scale Spatio-Temporal Features. *Sensors* **2022**, *22*, 1350. [CrossRef] [PubMed]
48. Yan, L.; Sheng, M.; Wang, C.; Gao, R.; Yu, H. Hybrid neural networks based facial expression recognition for smart city. *Multimed. Tools Appl.* **2022**, *81*, 319–342. [CrossRef]
49. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef]
50. Park, S.J.; Kim, B.G.; Chilamkurti, N. A robust facial expression recognition algorithm based on multi-rate feature fusion scheme. *Sensors* **2021**, *21*, 6954. [CrossRef]
51. Akhand, M.; Roy, S.; Siddique, N.; Kamal, M.A.S.; Shimamura, T. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* **2021**, *10*, 1036. [CrossRef]
52. Mehendale, N. Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
53. Zhao, Z.; Liu, Q.; Wang, S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Trans. Image Process.* **2021**, *30*, 6544–6556. [CrossRef]
54. Farzaneh, A.H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 2402–2411.
55. Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; Cai, J. Facial motion prior networks for facial expression recognition. In Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP), Sydney, Australia, 1–4 December 2019; pp. 1–4.
56. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv* **2021**, arXiv:2109.07270.
57. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef] [PubMed]
58. Jain, N.; Kumar, S.; Kumar, A.; Shamsolmoali, P.; Zareapoor, M. Hybrid deep neural networks for face emotion recognition. *Pattern Recognit. Lett.* **2018**, *115*, 101–106. [CrossRef]
59. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [CrossRef]
60. Fard, A.P.; Mahoor, M.H. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access* **2022**, *10*, 26756–26768. [CrossRef]
61. Pham, L.; Vu, T.H.; Tran, T.A. Facial Expression Recognition Using Residual Masking Network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4513–4519. [CrossRef]

62. Li, Y.; Wang, M.; Gong, M.; Lu, Y.; Liu, L. FER-former: Multi-modal Transformer for Facial Expression Recognition. *arXiv* **2023**, arXiv:2303.12997.
63. Wang, X.; Zhang, T.; Chen, C.L.P. PAU-Net: Privileged Action Unit Network for Facial Expression Recognition. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *15*, 1252–1262. [CrossRef]
64. Xue, F.; Wang, Q.; Guo, G. TransFER: Learning Relation-Aware Facial Expression Representations with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 3601–3610.
65. Wang, L.; Jia, G.; Jiang, N.; Wu, H.; Yang, J. EASE: Robust Facial Expression Recognition via Emotion Ambiguity-SEnsitive Cooperative Networks. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 218–227.
66. Jiang, J.; Deng, W. Disentangling Identity and Pose for Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1868–1878. [CrossRef]
67. Zhao, Z.; Liu, Q.; Zhou, F. Robust lightweight facial expression recognition network with label distribution training. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 3510–3519.
68. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 418–434.
69. Li, H.; Wang, N.; Yang, X.; Gao, X. CRS-CONT: A Well-Trained General Encoder for Facial Expression Analysis. *IEEE Trans. Image Process.* **2022**, *31*, 4637–4650. [CrossRef]
70. Sun, Z.; Zhang, H.; Bai, J.; Liu, M.; Hu, Z. A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition. *Pattern Recognit.* **2023**, *135*, 109157. [CrossRef]
71. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283.
72. Lyons, M.J.; Akamatsu, S.; Kamachi, M.; Gyoba, J.; Budynek, J. The Japanese female facial expression (JAFFE) database. In Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 14–16.
73. The Japanese Female Facial Expression (JAFFE) Database. 2017. Available online: http://www.kasrl.org/jaffe.html (accessed on 1 February 2023).
74. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D facial expression database for facial behavior research. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 211–216.
75. BU-3DFE Dataset. 2016. Available online: https://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html (accessed on 1 February 2023).
76. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
77. Kanade, T.; Cohn, J.F.; Tian, Y. Comprehensive database for facial expression analysis. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 46–53.
78. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef] [PubMed]
79. Yale Face Database. 2017. Available online: http://vision.ucsd.edu/content/yale-face-database (accessed on 1 February 2023).
80. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. *Acted Facial Expressions in the Wild Database*; Australian Technical Report TR-CS-11; Australian National University: Canberra, Australia, 2011.
81. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
82. Dhall, A.; Goecke, R.; Ghosh, S.; Joshi, J.; Hoey, J.; Gedeon, T. From individual to group-level emotion recognition: Emotiw 5.0. In Proceedings of the 19th ACM international Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 524–528.
83. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
84. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Proceedings of the 2011 IEEE International Conference on Computer VISION Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 2106–2112.
85. Dhall, A.; Ramana Murthy, O.; Goecke, R.; Joshi, J.; Gedeon, T. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 423–426.

86. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013; Springer: Cham, Switzelrand, 2013; pp. 117–124.
87. Susskind, J.M.; Anderson, A.K.; Hinton, G.E. *The Toronto Face Database*; Technical Report 3; Department of Computer Scienice, University of Toronto: Toronto, ON, Canada, 2010; p. 29.
88. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
89. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
90. Lundqvist, D.; Flykt, A.; Öhman, A. Karolinska directed emotional faces. *Cogn. Emot.* **1998**, *91*, 630.
91. KDEF Face Database. 1998. Available online: https://www.kdef.se/home/aboutKDEF.html (accessed on 1 February 2023).
92. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.
93. Benitez-Quiroz, C.F.; Srinivasan, R.; Feng, Q.; Wang, Y.; Martinez, A.M. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv* **2017**, arXiv:1703.01210.
94. Li, S.; Deng, W.; Du, J. Reliable crowd sourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
95. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From facial expression recognition to interpersonal relation prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569. [CrossRef]
96. Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. Multi-PIE. In Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–8. [CrossRef]
97. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005.
98. MMI Facial Expression Database. 2017. Available online: https://mmifacedb.eu (accessed on 1 February 2023).
99. Aifanti, N.; Papachristou, C.; Delopoulos, A. The MUG facial expression database. In Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, Desenzano del Garda, Italy, 12–14 April 2010; pp. 1–4.
100. Multimedia Understanding Group (MUG) Database. 2017. Available online: https://mug.ee.auth.gr/fed (accessed on 1 February 2023).
101. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Context based emotion recognition using emotic dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2755–2766. [CrossRef]
102. Emotic Database. 2017. Available online: http://sunai.uoc.edu/emotic/ (accessed on 1 February 2023).
103. Kollias, D.; Zafeiriou, S. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *arXiv* **2018**, arXiv:1811.07770.
104. Bassili, J.N. Facial motion in the perception of faces and of emotional expression. *J. Exp. Psychol. Hum. Percept. Perform.* **1978**, *4*, 373. [CrossRef] [PubMed]
105. Padgett, C.; Cottrell, G. Representing face images for emotion classification. *Adv. Neural Inf. Process. Syst.* **1996**, *9*, 894–900.
106. Guo, G.; Li, S.Z.; Chan, K. Face recognition by support vector machines. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 28–30 March 2000; pp. 196–201.
107. Matsugu, M.; Mori, K.; Mitari, Y.; Kaneda, Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw.* **2003**, *16*, 555–559. [CrossRef] [PubMed]
108. Cohen, I.; Sebe, N.; Gozman, F.; Cirelo, M.C.; Huang, T.S. Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003 Proceedings, Madison, WI, USA, 18–20 June 2003; Volume 1, p. 1.
109. Wang, Y.; Ai, H.; Wu, B.; Huang, C. Real time facial expression recognition with AdaBoost. In Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, Cambridge, UK, 26 August 2004; Volume 3, pp. 926–929. [CrossRef]
110. Kotsia, I.; Pitas, I. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Process.* **2006**, *16*, 172–187. [CrossRef]
111. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef] [PubMed]
112. Ranzato, M.; Susskind, J.; Mnih, V.; Hinton, G. On deep generative models with applications to recognition. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2857–2864.
113. Zhong, L.; Liu, Q.; Yang, P.; Liu, B.; Huang, J.; Metaxas, D.N. Learning active facial patches for expression analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2562–2569.
114. Tang, Y. Deep learning using linear support vector machines. *arXiv* **2013**, arXiv:1306.0239.

115. Kahou, S.E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, Ç.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R.C.; et al. Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, Australia, 9–13 December 2013; pp. 543–550.
116. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model 2149 for dynamic expression analysis. In Proceedings of the Asian conference on computer vision. 2150 Springer; 2014; pp. 143–157.
117. Ebrahimi Kahou, S.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent neural networks for emotion recognition in video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 467–474.
118. Kim, B.K.; Lee, H.; Roh, J.; Lee, S.Y. Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, New York, NY, USA, 9–13 November 2015; pp. 427–434. [CrossRef]
119. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.
120. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [CrossRef]
121. Zhang, F.; Zhang, T.; Mao, Q.; Xu, C. Joint pose and expression modeling for facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3359–3368.
122. Liu, C.; Jiang, W.; Wang, M.; Tang, T. Group Level Audio-Video Emotion Recognition Using Hybrid Networks. In Proceedings of the 2020 International Conference on Multimodal Interaction, New York, NY, USA, 25–29 October 2020; pp. 807–812. [CrossRef]
123. Kumar, A.; Kaur, A.; Kumar, M. Face detection techniques: A review. *Artif. Intell. Rev.* **2019**, *52*, 927–948. [CrossRef]
124. Bhele, S.G.; Mankar, V.H. A review paper on face recognition techniques. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **2012**, *1*, 339–346.
125. Mutlag, W.K.; Ali, S.K.; Aydam, Z.M.; Taher, B.H. Feature extraction methods: A review. *J. Phys. Conf. Ser.* **2020**, *1591*, 012028. [CrossRef]
126. Jun, H.; Shuai, L.; Jinming, S.; Yue, L.; Jingwei, W.; Peng, J. Facial Expression Recognition Based on VGGNet Convolutional Neural Network. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 4146–4151. [CrossRef]
127. Abhishree, T.; Latha, J.; Manikantan, K.; Ramachandran, S. Face recognition using Gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique. *Procedia Comput. Sci.* **2015**, *45*, 312–321. [CrossRef]
128. Kotsiantis, S.B.; Kanellopoulos, D.; Pintelas, P.E. Data preprocessing for supervised leaning. *Int. J. Comput. Sci.* **2006**, *1*, 111–117.
129. Ulrich, L.; Marcolin, F.; Vezzetti, E.; Nonis, F.; Mograbi, D.C.; Scurati, G.W.; Dozio, N.; Ferrise, F. CalD3r and MenD3s: Spontaneous 3D facial expression databases. *J. Vis. Commun. Image Represent.* **2024**, *98*, 104033. [CrossRef]
130. Zeng, D.; Veldhuis, R.; Spreeuwers, L. A survey of face recognition techniques under Occlusioniet. *IET Biom.* **2021**, *10*, 581–606. [CrossRef]
131. Ekenel, H.K.; Stiefelhagen, R. Why is facial occlusion a challenging problem? In Proceedings of the International Conference on Biometrics, Alghero, Italy, 2–5 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 299–308.
132. Braje, W.L.; Kersten, D.; Tarr, M.J.; Troje, N.F. Illumination effects in face recognition. *Psychobiology* **1998**, *26*, 371–380. [CrossRef]
133. Wood, R.; Olszewska, J.I. Lighting-variable AdaBoost based-on system for robust face detection. In Proceedings of the 5th International Conference on Bio-Inspired Systems and Signal Processing, Algarve, Portugal, 1–4 February 2012; SciTePress Digital Library: Algarve, Portugal, 2012; pp. 494–497.
134. Zou, W.W.; Yuen, P.C. Very low resolution face recognition problem. *IEEE Trans. Image Process.* **2011**, *21*, 327–340. [CrossRef] [PubMed]
135. Abaza, A.; Harrison, M.A.; Bourlai, T.; Ross, A. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biom.* **2014**, *3*, 314–324. [CrossRef]
136. Prikler, F. Evaluation of emotional state of a person based on facial expression. In Proceedings of the 2016 XII International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Lviv, Ukraine, 20–24 April 2016; pp. 161–163.
137. Shreve, M.; Godavarthy, S.; Goldgof, D.; Sarkar, S. Macro-and micro-expression spotting in long videos using spatio-temporal strain. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–25 March 2011; pp. 51–56.
138. Hasan, M.K.; Ahsan, M.S.; Newaz, S.S.; Lee, G.M. Human face detection techniques: A comprehensive review and future research directions. *Electronics* **2021**, *10*, 2354. [CrossRef]
139. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. 1. [CrossRef]
140. Cootes, T.F.; Taylor, C.J. Active shape models—'smart snakes'. In Proceedings of the BMVC92: Proceedings of the British Machine Vision Conference, Leeds, UK, 22–24 September 1992; Springer: Berlin/Heidelberg, Germany, 1992; pp. 266–275.
141. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]

142. Liu, C.; Wechsler, H. Independent component analysis of Gabor features for face recognition. *IEEE Trans. Neural Netw.* **2003**, *14*, 919–928.
143. Zhang, Z.; Lyons, M.; Schuster, M.; Akamatsu, S. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 454–459.
144. Face-Apple. Available online: https://machinelearning.apple.com/research/face-detection#1 (accessed on 1 February 2023).
145. Mita, T.; Kaneko, T.; Hori, O. Joint Haar-like features for face detection. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 1, pp. 1619–1626. [CrossRef]
146. Johnston, B.; Chazal, P.d. A review of image-based automatic facial landmark identification techniques. *EURASIP J. Image Video Process.* **2018**, *2018*, 86. [CrossRef]
147. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [CrossRef]
148. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Incremental face alignment in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1859–1866.
149. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
150. Ren, S.; Cao, X.; Wei, Y.; Sun, J. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1685–1692.
151. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
152. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
153. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
154. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
155. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
156. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
157. Vijayakumar, S.; Schaal, S. Locally weighted projection regression: An o (n) algorithm for incremental real time learning in high dimensional space. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford, CA, USA, 29 June–2 July 2000; Morgan Kaufmann: Burlington, MA, USA, 2000; Volume 1, pp. 288–293.
158. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
159. Tomasi, C. Histograms of oriented gradients. *Comput. Vis. Sampl.* **2012**, 1–6.
160. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
161. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
162. Tharwat, A. Principal component analysis: An overview. *Pattern Recognit.* **2016**, *3*, 197–240.
163. Friesen, E.; Ekman, P. Facial action coding system: A technique for the measurement of facial movement. *Palo Alto* **1978**, *3*, 5.
164. Tharwat, A.; Gaber, T.; Ibrahim, A.; Hassanien, A.E. Linear discriminant analysis: A detailed tutorial. *AI Commun.* **2017**, *30*, 169–190. [CrossRef]
165. Comon, P. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
166. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
167. Hong, X.; Xu, Y.; Zhao, G. Lbp-top: A tensor unfolding revisit. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2017; pp. 513–527.
168. Bosch, A.; Zisserman, A.; Munoz, X. Image classification using random forests and ferns. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
169. Mase, K. Recognition of facial expression from optical flow. *IEICE Trans. Inf. Syst.* **1991**, *74*, 3474–3483.
170. Cohn, J.; Zlochower, A.; Lien, J.; Kanade, T. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 396–401. [CrossRef]
171. Barron, J.L.; Fleet, D.J.; Beauchemin, S.S. Performance of optical flow techniques. *Int. J. Comput. Vis.* **1994**, *12*, 43–77. [CrossRef]
172. Ekman, P.; Friesen, W.V. Facial action coding system. *Environ. Psychol. Nonverbal Behav.* **1978**, 22.
173. Canedo, D.; Neves, A.J. Facial expression recognition using computer vision: A systematic review. *Appl. Sci.* **2019**, *9*, 4678. [CrossRef]
174. Carreira-Perpinán, M.A. *A Review of Dimension Reduction Techniques*; Technical Report CS-96-09; Department of Computer Science, University of Sheffield: Sheffield, UK, 1997; Volume 9, pp. 1–69.
175. Calder, A.J.; Burton, A.M.; Miller, P.; Young, A.W.; Akamatsu, S. A principal component analysis of facial expressions. *Vis. Res.* **2001**, *41*, 1179–1208. [CrossRef]

176. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86. [CrossRef] [PubMed]
177. Vretos, N.; Tefas, A.; Pitas, I. Facial expression recognition with robust covariance estimation and Support Vector Machines. In Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing, Santander, Spain, 23–26 September 2012; pp. 1–5.
178. Deng, H.B.; Jin, L.W.; Zhen, L.X.; Huang, J.C. A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA. *Int. J. Inf. Technol.* **2005**, *11*, 86–96.
179. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [CrossRef]
180. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
181. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
182. Nefian, A.V.; Hayes, M.H. Hidden Markov models for face recognition. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), Seattle, WA, USA, 15 May 1998; Volume 5, pp. 2721–2724.
183. Michel, P.; El Kaliouby, R. Real time facial expression recognition in video using support vector machines. In Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, BC, Canada, 5–7 November 2003; pp. 258–264.
184. Dagher, I.; Dahdah, E.; Al Shakik, M. Facial expression recognition using three-stage support vector machines. *Vis. Comput. Ind. Biomed. Art* **2019**, *2*, 24. [CrossRef]
185. Sarker, I.H. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput. Sci.* **2021**, *2*, 420. [CrossRef]
186. Wu, S.T.; Roberts, K.; Datta, S.; Du, J.; Ji, Z.; Si, Y.; Soni, S.; Wang, Q.; Wei, Q.; Xiang, Y.; et al. Deep learning in clinical natural language processing: A methodical review. *J. Am. Med. Inform. Assoc. JAMIA* **2020**, *27*, 457–470. [CrossRef]
187. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* **2019**, *7*, 19143–19165. [CrossRef]
188. Sapoval, N.; Aghazadeh, A.; Nute, M.G.; Antunes, D.A.; Balaji, A.; Baraniuk, R.; Barberan, C.J.; Dannenfelser, R.; Dun, C.; Edrisi, M.; et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **2022**, *13*, 1728. [CrossRef]
189. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Columbus, OH, USA, 23–28 June 2014.
190. Wang, X.; Wang, K.; Lian, S. A survey on face data augmentation for the training of deep neural networks. *Neural Comput. Appl.* **2020**, *32*, 15503–15531. [CrossRef]
191. Deng, L.; Yu, D. Deep learning: Methods and applications. *Found. Trends®Signal Process.* **2014**, *7*, 197–387. [CrossRef]
192. McCelloch, W.; Pitts, W. A logical calculus of the idea immanent in neural nets. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
193. Padgett, C.W.; Cottrell, G. Representing Face Images for Emotion Classification. In Proceedings of the NIPS, Cambridge, MA, USA, 3–5 December 1996.
194. RAFD Face Database. 2010. Available online: http://rafd.socsci.ru.nl/RaFD2/RaFD?p=main (accessed on 1 February 2023).
195. Li, S.; Deng, W. Reliable Crowd sourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370. [CrossRef]
196. Olszewska, J.I. Automated face recognition: Challenges and solutions. In *Pattern Recognition Analysis and Applications*; IntechOpen: Rijeka, Croatia, 2016; pp. 59–79.
197. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
198. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
199. Liu, P.; Han, S.; Meng, Z.; Tong, Y. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1805–1812.
200. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. *Learning Internal Representations by Error Propagation*; Technical Report; California University San Diego, La Jolla Institute for Cognitive Science: La Jolla, CA, USA, 1985.
201. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
202. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649. [CrossRef]
203. Lopes, A.T.; De Aguiar, E.; De Souza, A.F.; Oliveira-Santos, T. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognit.* **2017**, *61*, 610–628. [CrossRef]
204. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
205. Ponti, M.A.; Ribeiro, L.S.F.; Nazare, T.S.; Bui, T.; Collomosse, J. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In Proceedings of the 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Niteroi, Brazil, 17–18 October 2017; pp. 17–41.

206. O'Shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
207. Le, T.; Duan, Y. Pointgrid: A deep network for 3d shape understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9204–9214.
208. Singh, S.; Nasoz, F. Facial expression recognition with convolutional neural networks. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 0324–0328.
209. Gholamalinezhad, H.; Khosravi, H. Pooling methods in deep neural networks, a review. *arXiv* **2020**, arXiv:2009.07485.
210. Zhang, C.L.; Luo, J.H.; Wei, X.S.; Wu, J. In Defense of Fully Connected Layers in Visual Representation Transfer. In *Proceedings of the Advances in Multimedia Information Processing—PCM 2017*; Zeng, B., Huang, Q., El Saddik, A., Li, H., Jiang, S., Fan, X., Eds.; Springer: Cham, Switzerland, 2018; pp. 807–817.
211. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
212. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
213. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the BMVC 2015—Proceedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015.
214. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
215. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
216. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
217. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
218. Canziani, A.; Paszke, A.; Culurciello, E. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv* **2016**, arXiv:1605.07678, 2016.
219. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
220. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
221. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
222. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features With 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
223. Konda, K.; Memisevic, R.; Krueger, D. Zero-bias autoencoders and the benefits of co-adapting features. *arXiv* **2014**, arXiv:1402.3337.
224. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 118–126.
225. Wang, H.; Raj, B. On the origin of deep learning. *arXiv* **2017**, arXiv:1702.07800.
226. Li, C.; Wang, Y.; Zhang, X.; Gao, H.; Yang, Y.; Wang, J. Deep belief network for spectral–spatial classification of hyperspectral remote sensor data. *Sensors* **2019**, *19*, 204. [CrossRef]
227. Werbos, P.J. Backpropagation through time: What it does and how to do it. *Proc. IEEE* **1990**, *78*, 1550–1560. [CrossRef]
228. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
229. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
230. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
231. Yu, Z.; Liu, G.; Liu, Q.; Deng, J. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing* **2018**, *317*, 50–57. [CrossRef]
232. Zhang, H.; Huang, B.; Tian, G. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognit. Lett.* **2020**, *131*, 128–134. [CrossRef]
233. Yang, B.; Cao, J.; Ni, R.; Zhang, Y. Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images. *IEEE Access* **2018**, *6*, 4630–4640. [CrossRef]
234. Dresvyanskiy, D.; Ryumina, E.; Kaya, H.; Markitantov, M.; Karpov, A.; Minker, W. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. *Multimodal Technol. Interact.* **2022**, *6*, 11. [CrossRef]
235. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
236. Applications of GAN. Available online: https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/ (accessed on 1 February 2023).
237. Shen, Y.; Zhou, B.; Luo, P.; Tang, X. Facefeat-gan: A two-stage approach for identity-preserving face synthesis. *arXiv* **2018**, arXiv:1812.01288.

238. Shiri, F.; Yu, X.; Porikli, F.; Hartley, R.; Koniusz, P. Identity-preserving face recovery from stylized portraits. *Int. J. Comput. Vis.* **2019**, *127*, 863–883. [CrossRef]
239. Li, J.; Lam, E.Y. Facial expression recognition using deep neural networks. In Proceedings of the 2015 IEEE International Conference on Imaging Systems and Techniques (IST), Macau, China, 16–18 September 2015; pp. 1–6.
240. Zhang, S.; Pan, X.; Cui, Y.; Zhao, X.; Liu, L. Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access* **2019**, *7*, 32297–32304. [CrossRef]
241. Khanum, A.; Lee, C.Y.; Yang, C.S. Deep-Learning-Based Network for Lane Following in Autonomous Vehicles. *Electronics* **2022**, *11*, 3084. [CrossRef]
242. Pan, X. Fusing HOG and convolutional neural network spatial–temporal features for video-based facial expression recognition. *IET Image Process.* **2020**, *14*, 176–182. [CrossRef]
243. Sun, X.; Lv, M. Facial expression recognition based on a hybrid model combining deep and shallow features. *Cogn. Comput.* **2019**, *11*, 587–597. [CrossRef]
244. Abdulsattar, N.S.; Hussain, M.N. Facial expression recognition using HOG and LBP features with convolutional neural network. *Bull. Electr. Eng. Inform.* **2022**, *11*, 1350–1357. [CrossRef]
245. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
246. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
247. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Tong, Y. Island Loss for Learning Discriminative Features in Facial Expression Recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 302–309. [CrossRef]
248. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
249. Guo, Y.; Tao, D.; Yu, J.; Xiong, H.; Li, Y.; Tao, D. Deep Neural Networks with Relativity Learning for facial expression recognition. In Proceedings of the 2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW), Seattle, WA, USA, 11–15 July 2016; pp. 1–6. [CrossRef]
250. Liu, X.; Vijaya Kumar, B.V.K.; You, J.; Jia, P. Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
251. Li, Y.; Lu, Y.; Li, J.; Lu, G. Separate loss for basic and compound facial expression recognition in the wild. In Proceedings of the Asian Conference on Machine Learning, PMLR, Nagoya, Japan, 17–19 November 2019; pp. 897–911.
252. Li, H.; Wang, N.; Ding, X.; Yang, X.; Gao, X. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Trans. Image Process.* **2021**, *30*, 2016–2028. [CrossRef]
253. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991. [CrossRef]
254. Bargal, S.A.; Barsoum, E.; Ferrer, C.C.; Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 433–436.
255. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
256. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
257. Fan, Y.; Lam, J.C.; Li, V.O. Multi-region ensemble convolutional neural network for facial expression recognition. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; Springer: Cham, Switzerland, 2018; pp. 84–94.
258. Georgescu, M.I.; Ionescu, R.T.; Popescu, M. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access* **2019**, *7*, 64827–64836. [CrossRef]
259. Rajan, S.; Chenniappan, P.; Devaraj, S.; Madian, N. Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM. *IET Image Process.* **2020**, *14*, 1373–1381. [CrossRef]
260. Khaireddin, Y.; Chen, Z. Facial emotion recognition: State of the art performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.
261. Wang, Z.; Zeng, F.; Liu, S.; Zeng, B. OAENet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognit.* **2021**, *112*, 107694. [CrossRef]
262. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
263. Ming, Z.; Xia, J.; Luqman, M.M.; Burie, J.C.; Zhao, K. Dynamic multi-task learning for face recognition with facial expression. *arXiv* **2019**, arXiv:1911.03281.
264. Serengil, S.I.; Ozpinar, A. HyperExtended LightFace: A Facial Attribute Analysis Framework. In Proceedings of the 2021 International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey, 27–28 October 2021; pp. 1–4. [CrossRef]

265. Zhang, Z.; Song, Y.; Qi, H. Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5810–5818.
266. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
267. Kollias, D. ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Multi-Task Learning Challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 18–24 June 2022; pp. 2328–2336.
268. Huang, Y.; Khan, S.M. Dyadgan: Generating facial expressions in dyadic interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 June 2017; pp. 11–18.
269. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2168–2177.
270. Wu, R.; Zhang, G.; Lu, S.; Chen, T. Cascade EF-GAN: Progressive Facial Expression Editing with Local Focuses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
271. Liu, Y.; Zhang, X.; Li, Y.; Zhou, J.; Li, X.; Zhao, G. Graph-based facial affect analysis: A review. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2657–2677. [CrossRef]
272. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
273. Liao, L.; Zhu, Y.; Zheng, B.; Jiang, X.; Lin, J. FERGCN: Facial expression recognition based on graph convolution network. *Mach. Vis. Appl.* **2022**, *33*, 40. [CrossRef]
274. Wu, C.; Chai, L.; Yang, J.; Sheng, Y. Facial expression recognition using convolutional neural network on graphs. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 7572–7576.
275. Wasi, A.T.; Šerbetar, K.; Islam, R.; Rafi, T.H.; Chae, D.K. ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning. *arXiv* **2023**, arXiv:2305.01486.
276. Perveen, N.; Gupta, S.; Verma, K. Facial expression recognition system using statistical feature and neural network. *Int. J. Comput. Appl.* **2012**, *48*, 17–23. [CrossRef]
277. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame Attention Networks for Facial Expression Recognition in Videos. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870. [CrossRef]
278. Hasani, B.; Mahoor, M.H. Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
279. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
280. Kervadec, C.; Vielzeuf, V.; Pateux, S.; Lechervy, A.; Jurie, F. Cake: Compact and accurate k-dimensional representation of emotion. *arXiv* **2018**, arXiv:1807.11215.
281. Hayale, W.; Negi, P.; Mahoor, M. Facial expression recognition using deep siamese neural networks with a supervised loss function. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–7.
282. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* **2018**, *49*, 839–847. [CrossRef] [PubMed]
283. Salmam, F.Z.; Madani, A.; Kissi, M. Emotion recognition from facial expression based on fiducial points detection and using neural network. *Int. J. Electr. Comput. Eng.* **2018**, *8*, 52. [CrossRef]
284. Nguyen, H.D.; Yeom, S.; Lee, G.S.; Yang, H.J.; Na, I.S.; Kim, S.H. Facial emotion recognitionusing an ensemble of multi-level convolutional neural networks. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 1940015. [CrossRef]
285. Vulpe-Grigoras¸i, A.; Grigore, O. Convolutional neural network hyperparameters optimization for facial emotion recognition. In Proceedings of the 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 25–27 March 2021; pp. 1–5.
286. Pecoraro, R.; Basile, V.; Bono, V. Local multi-head channel self-attention for facial expression recognition. *Information* **2022**, *13*, 419. [CrossRef]
287. Burkert, P.; Trier, F.; Afzal, M.Z.; Dengel, A.; Liwicki, M. Dexpression: Deep convolutional neural network for expression recognition. *arXiv* **2015**, arXiv:1509. 05371.
288. Hasani, B.; Mahoor, M.H. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 790–795.
289. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-piloted deep network for facial expression recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 425–442.
290. Yu, Z.; Liu, Q.; Liu, G. Deeper cascaded peak-piloted network for weak expression recognition. *Vis. Comput.* **2018**, *34*, 1691–1699. [CrossRef]

291. Kuo, C.M.; Lai, S.H.; Sarkis, M. A Compact Deep Learning Model for Robust Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
292. Kollias, D.; Cheng, S.; Ververas, E.; Kotsia, I.; Zafeiriou, S. Deep Neural Network Augmentation: Generating Faces for Affect Analysis. *Int. J. Comput. Vis.* **2020**, *128*, 1455–1484. [CrossRef]
293. Vo, T.H.; Lee, G.S.; Yang, H.J.; Kim, S.H. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [CrossRef]
294. Psaroudakis, A.; Kollias, D. MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2367–2375.
295. Zhang, Y.; Wang, C.; Deng, W. Relative Uncertainty Learning for Facial Expression Recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17616–17627.
296. Zhou, H.; Meng, D.; Zhang, Y.; Peng, X.; Du, J.; Wang, K.; Qiao, Y. Exploring emotion features and fusion strategies for audio-video emotion recognition. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 562–566.
297. Kumar, V.; Rao, S.; Yu, L. Noisy student training using body language dataset improves facial expression recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 756–773.
298. Adrian, R. Deep Learning for Computer Vision with Python Volume 1; Pyimage-Search. 2017. Available online: https://bayanbox.ir/view/5130918188419813120/Adrian-Rosebrock-Deep-Learning-for.pdf (accessed on 1 February 2023).
299. Cheng, S.; Kotsia, I.; Pantic, M.; Zafeiriou, S. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5117–5126.
300. Sun, N.; Tao, J.; Liu, J.; Sun, H.; Han, G. 3-D Facial Feature Reconstruction and Learning Network for Facial Expression Recognition in the Wild. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 298–309. [CrossRef]
301. Wu, Z.; Wang, X.; Jiang, Y.G.; Ye, H.; Xue, X. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In Proceedings of the 23rd ACM International Conference on Multimedia, New York, NY, USA, 26–30 October 2015; pp. 461–470. [CrossRef]
302. Dang, C.N.; Moreno-García, M.N.; De la Prieta, F. Hybrid deep learning models for sentiment analysis. *Complexity* **2021**, *2021*, 9986920. [CrossRef]
303. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
304. Khan, M.M.; Ward, R.D.; Ingleby, M. Automated classification and recognition of facial expressions using infrared thermal imaging. In Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems, Singapore, 1–3 December; 2004; Volume 1, pp. 202–206.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.