SN Computer Science (2024) 5:96
https://doi.org/10.1007/s42979-023-02423-7

REVIEW ARTICLE

Facial Emotion Recognition in-the-Wild Using Deep Neural Networks: A Comprehensive Review

Hadjer Boughanem[1] · **Haythem Ghazouani**[1,2] · **Walid Barhoumi**[1,2]

Received: 25 June 2023 / Accepted: 15 October 2023
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2023

Abstract

Facial expressions are a crucial aspect of human communication that provide information about emotions, intentions, interactions, and social relationships. They are a universal signal used daily to convey inner behaviors in natural situations. With the increasing interest in automatic facial emotion recognition, deep neural networks have become a popular tool for recognizing emotions in challenging in-the-wild conditions that are closer to reality. However, these systems must contend with external factors that degrade the quality of facial features, making it challenging to determine the correct emotion classes. In this paper, we first provide a summary of the various fields that use facial recognition systems under in-the-wild context. Then, we extensively examine the major datasets utilized for in-the-wild facial expression recognition, taking into account their appropriateness for this context, the challenges related to their application, the coverage of various emotions, and the potential domains of application. The analysis is conducted rigorously, emphasizing the merits and demerits of each dataset and advocating for their pertinence and effectiveness in real-life situations. We also present an expanded taxonomy of facial emotion recognition in-the-wild, while focusing mainly on deep learning methods and covering the manufacturing steps of a facial emotion recognition system and the different possible techniques for each step. Finally, we provide a discussion, insights, and conclusion, making this survey a reference point for researchers interested in the in-the-wild context, while providing a better understanding of the different datasets' compositions and specificities. This survey can help advance research on deep facial emotion recognition in-the-wild and serve as a resource for methods, applications, and datasets in the field.

Keywords FER datasets · Emotion recognition · Deep learning · State of the art · In-the-wild · Applications of facial emotion analysis

Introduction

Faces, which are ubiquitous in day-to-day life, offer a rich source of information for human communication. Indeed, facial expressions are the most common means for conveying emotions, intentions, and social relationships. In particular, recognizing emotions, which are a crucial part of interpersonal interactions and discussions, is essential for effective communication and adaptability to the environment [53]. Emotions can be transmitted through different channels, including facial expressions, vocal tones, gestures, gaze directions, and postures. However, it has been observed that facial expressions play a dominant role in expressing emotions and conveying intentions. In fact, facial expressions are facial changes that occur in response to the internal emotional states of individuals, their intentions or their social communications. Thus, facial expression analysis has been an active research topic for behavioral scientists since

✉ Walid Barhoumi
walid.barhoumi@enicarthage.rnu.tn

Hadjer Boughanem
hadjer.boughanem@fst.utm.tn

Haythem Ghazouani
haythem.ghazouani@enicar.u-carthage.tn

[1] Institut Supérieur d'Informatique d'El Manar, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), Université de Tunis El Manar, 2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisia

[2] Ecole Nationale d'Ingénieurs de Carthage, Université de Carthage, 45 Rue des Entrepreneurs, 2035 Tunis-Carthage, Tunisia

Published online: 13 December 2023

Darwin's work in 1872. Facial expressions are generated by contractions of facial muscles resulting in temporally deformed facial features, such as eyelids, eyebrows, nose, lips, and skin texture, often revealed by wrinkles and bulges [64]. These facial expressions can be used to recognize various emotions, such as happiness, sadness, surprise, fear, anger, and disgust. Moreover, emotion recognition can help to identify the mood of individuals and their reactions to different situations, and can be thus used in various fields ranging from psychology to virtual reality gaming. Consequently, automatic facial emotion recognition has been an active area of research in recent years due to its numerous applications and potential benefits. Indeed, Facial Expression Recognition (FER) has become increasingly popular in recent years owing to its wide range of applications, including healthcare, security, entertainment, and marketing. While controlled environments, such as laboratories, can provide ideal conditions for FER research, they do not accurately represent the complexity of the real world. This has led to a growing interest in developing FER systems that can perform reliably in challenging environments, including in-the-wild scenarios. In-the-wild conditions are characterized by naturalistic settings where facial expressions are captured in real-world situations with varying lighting conditions, poses, occlusions, and expressions. Unlike controlled environments, in-the-wild conditions are not designed to optimize the quality of facial features, and this imposes significant challenges for FER systems. In this framework, Convolutional Neural Network (CNNs) have become the tool of choice for recognizing emotions under difficult conditions in-the-wild. They are capable of recognizing patterns within large amounts of data, making them particularly suited for FER tasks. However, even with the power of today's modern CNNs, in-the-wild FER systems must contend with external factors that degrade the quality of facial features. For instance, varying lighting conditions can cause shadows on the face, which can thereby obscure important features, such as the eyes and the mouth. Occlusions, such as glasses or masks, can also obscure facial features and make it challenging to determine the correct emotion classes. Despite these issues, the development of FER systems that can perform reliably in the wild has the potential to revolutionize a wide range of industries. For instance, FER systems could be adopted in healthcare for detecting early signs of depression or anxiety disorder. Within the framework of security and privacy, FER tools could contribute for detecting suspicious behavior in public spaces. In entertainment, FER systems could be used to personalize content based on the emotional state of the user.

In this study, we provide a comprehensive overview of the latest research efforts on facial emotion recognition under challenging in-the-wild conditions. Specifically, we aim to provide the research community with an understanding of the various datasets used for emotion recognition in-the-wild, especially their challenges and their suitability for specific target domains. We begin by summarizing the use of emotion recognition systems in various fields while highlighting the importance of developing automated emotion recognition tools that can perform reliably under in-the-wild environments. We then review the state-of-the-art of in-the-wild FER datasets while analyzing their variability and challenges. Then, to provide a deeper understanding of FER in the wild, we present an expanded taxonomy of FER, while focusing on deep learning methods and covering the manufacturing steps of a facial emotion recognition system. Specifically, we discuss the different possible techniques for each step, including face detection, alignment, feature extraction, and classification. Finally, we provide a deep discussion and conclusion, highlighting the importance of this review as a reference point for researchers interested in the area of emotion recognition in-the-wild. In fact, we aim to advance research on deep emotion recognition by providing a comprehensive overview of methods, applications, and datasets in the field, and by increasing awareness of the specificities of the in-the-wild contexts. Ultimately, this study can serve as a valuable resource for researchers and practitioners seeking to develop reliable FER systems that can operate effectively in real-world situations.

Overall, the main contributions of this paper are threefold:

1. This is the first survey, as best as we know, that specifically focuses on deep learning works within the framework of in-the-wild facial expression recognition.
2. The study provides a comprehensive overview of well-known and extensively used FER datasets, while systematically evaluating their suitability within the in-the-wild scenario. The assessment includes a thorough analysis of the strengths and limitations of each dataset. Moreover, practical recommendations are offered considering their relevance across different application domains.
3. We introduce a new taxonomy for in-the-wild deep learning models, while categorizing them based on inputs, utilized modalities, and used networks, incorporating different levels of refinement. The taxonomy allows for a precise classification of techniques and architectures, facilitating deeper insights into their nuances and contributions. A comprehensive analysis is provided, further elucidating the significance of the studied techniques and architectures.

The remaining part of this paper is organized as follows. In Section "Applications of Emotion Recognition in-the-Wild", we present the most challenging factors of FER in-the-wild, while discussing different FER applications. In Section "In-the-Wild Facial Expression Datasets", we thoroughly investigate the leading datasets used for recognizing facial expressions in real-world settings, considering their

suitability, challenges, emotional coverage, and potential applications in different domains. The analysis is conducted with scientific rigor, while highlighting the strengths and weaknesses of each dataset and arguing for their relevance and usefulness in practical scenarios. In Section "Taxonomy of Existing Deep Learning-Based FER in-the-Wild", we provide a detailed review of FER in-the-wild based on deep learning techniques, what is the primary focus of this study. We first organize the review according to the input starting with the pre-processing techniques and moving on to the input modalities. A taxonomy is then presented according to the network architecture, training procedure, and classification strategy, providing a systematic framework for understanding the various approaches in the literature. Finally, in Section "Discussion and Insights", we provide a discussion before summarizing the key findings and contributions of this study.

Applications of Emotion Recognition in-the-Wild

Despite showing high recognition accuracy on lab-controlled datasets; which are collected in the laboratory settings such as frontal faces, high-resolution images, and good illumination conditions (e.g., JAFFE, CK+ [13, 32, 56], KDEF [78]); emotion recognition systems achieve much lower accuracy when dealing with datasets collected in real-world uncontrolled environments, also referred to as "in-the-wild" datasets. These datasets exhibit occlusions, large variation in head-pose and face size, low image resolution, lighting change, face orientation, and movement blur. Other major difficulties associated with emotion recognition in-the-wild include non-frontal head-pose which can be an obstacle to detecting the faces and interpreting the facial expressions. Moreover, it may be laborious to discriminate some emotions that do not substantially exhibit their original emotional classes [11], in addition to the existence of some uncommon expressions [49], especially when the number of samples is low for a fair few of these classes. Nevertheless, emotion recognition in the wild has attracted more attention in the fields of computer vision and artificial intelligence, and a variety of real-world applications involving in-the-wild emotion recognition have been promoted for encoding emotions based on facial expression representations. These applications have utility in various domains, such as security, healthcare, education, gaming, access control, and video surveillance. Noteworthy applications include drowsiness detection for driver safety [39, 67, 68, 71], pain analysis in healthcare [3, 31, 36], client satisfaction [14, 19], video conferencing [10, 26, 27], and facial expression recognition for cognitive sciences [16]. Many of these applications can be leveraged in daily life and uncontrolled environments. Consequently, there is a pressing need to enhance the performance of automatic facial emotion recognition under in-the-wild conditions. In the subsequent sections, we focus on domains where emotion recognition systems from facial expressions are commonly employed to facilitate daily tasks.

Healthcare

To be effectively deployed in healthcare, an emotion recognition system should be tested and validated under natural conditions to ensure its reliability, given the nature of healthcare environments and conditions. In fact, in these conditions, facial expressions are not staged or controlled, and the environment can be highly variable, which may impact the quality and the reliability of facial expression recognition. Additionally, patients' facial expressions may vary widely depending on various factors, such as pain, anxiety, depression, and medication effects, which can be difficult to control or predict. Moreover, many healthcare providers are often limited in their evaluation time, and they generally rely on nonverbal cues, such as facial expressions, to assess the patient's emotional state, making it more challenging to achieve accurate recognition results. Therefore, it is essential for a FER system that should be employed in healthcare to handle these external factors to perform well in real-world situations, highlighting the importance of defining it within an in-the-wild context. In addition, healthcare providers may use FER systems to monitor patients' emotional states during their care, such as post-operative pain or during counseling sessions, to provide timely interventions for addressing negative emotional states. These emotional states may be unpredictable and diverse, further emphasizing the need for effective emotion recognition systems that can accurately recognize emotions in an in-the-wild context. Several research works on emotion recognition have focused on this application domain. For instance, Anneketh et al. [79] have proposed an automated psychometric analyzer that can perform sentiment analysis in short span of time leading to more accurate results using the voice analysis feature. They have based their research work on the survey of current technological advancements and previous research of sentiment analysis and emotion recognition. Their model is intended for the psychology field (Mental Health Care) by combining various methods and algorithms to complete an automated psychometric analysis or to develop a self-serving psychometric kiosk. Being very cheap and time-saving, the designed tool would solve many problems faced by doctors, especially psychologists, and patients. Moreover, it could help not only doctors to monitor the civil patients' health but also army health care, corporate sector recruitment and employee health. This could be applicable when using speech-based emotional analysis from recorded voice in addition to typed and handwritten text-based sentiment

analysis. The developed concepts have allowed to make a self-serving medical kiosk or a psychometric analyzer that is capable of performing fast computational linguistics, thus producing a short-crisp summary of emotional health of the patient based on previous records, medications, and treatments. This concept led to save the time of both doctors and patients.

Overall, looking into the complexity in the e-health sector, the recourse to a system capable of patients' pain recognition, especially for those who cannot express orally their "sufferance physique", is inevitable. It would be a great help for the medical staff and a substantial improvement of quality of life. In this context, Francisco et al. [62] have implemented a tool, whose principal function is performing remote patients' monitoring based on computer vision and supervised learning, to automatically detect their emotions. They have been inspired by the fact that facial expressions are valid indicators of the degree of pain of persons. According to the authors, in addition to recognizing images confusing human being, this pain recognition tool would be a great advance in smart cities; through improving the e-health sector in the babies care sector, the reduction of the continuous tracking, saving costs, and a more objective diagnosis of pain suffered by babies. Another research work that forms part of the emotion detection within the e-health field has been carried out by Lucey et al. [52]. In fact, authors have investigated the replacement of patient self-report, which is the most widely used technique to measure pain in clinical setting, through coding pain as a series of facial Action Units (AUs) on a frame by-frame basis. For patients with shoulder injuries, authors have processed video data, before describing an Active Appearance Model (AAM)-based computer vision system carrying out automatic detection of frames in videos illustrating patients suffering from pain. The implementation of a system detecting patients' pain automatically is of great importance thanks to many reasons. First, in hospital setting, it would improve greatly the efficiency and overheads associated with monitoring patient healing progress. Second, it could allow to overcome the issues associated with detecting spontaneous data, such as pain, due to the major facial expressions and considerable head motion. In addition, such system could potentially provide significant advantage in patient care and cost reduction.

Automotive

In the automotive industry, an FER system would typically face an in-the-wild context due to the uncontrolled nature of driving situations. Unlike in controlled lab environments, the lighting, the noise, and the environmental conditions can vary widely, which can negatively impact the reliability and the accuracy of automated facial expression recognition. Additionally, drivers' facial expressions are often affected by factors, such as fatigue, distraction, or frustration, which can be challenging to predict or control. Therefore, an FER system that is robust and able to handle these external factors is necessary for successful real-world implementation. Furthermore, in the automotive sector, an emotion recognition system can be employed in advanced driver assistance systems to detect drowsiness and/or distraction while alerting the driver to take corrective actions, further highlighting the importance of accurate and reliable facial expression recognition in a such promising in-the-wild area. In fact, nowadays, automobiles should ensure the driver's pleasure in addition to his transportation and security. This emotional interaction is related directly with the feeling of the driver in a driver–vehicle environment [17]. To deduce the emotional state of drivers, three technologies are mainly used: emotion recognition from speech, emotion recognition from facial expressions, and emotion recognition from physiological data. Martin et al. [75] have combined the three mentioned techniques into a test car and they have demonstrated the applicability of the proposed solution for assessing the driver's pleasure. They have implemented a wireless sensor system for emotion recognition consisting of gloves hosting sensing elements measuring skin temperature and resistance, a wrist pocket containing sensing electronics, in addition to a base unit that enhances data and stores it on an SD card. In this study, two cars have been equipped with sensors and eight invited nonprofessional participants aging from 33 to 53 years. The obtained results have showed that emotional speech most significant parameters are pitch, intensity, and energy changes over frequency bands. For the facial expression recognition, it was hard to distinguish because of the presence of non-emotional-related head motion while observing the traffic. Physiology training data for classifiers were insufficient to achieve a desirable confidence, what makes the authors assume that these data should not be included into the final result report. As a conclusion, they have stated that positive emotion measures in real-world settings can be realized using affect sensors and that useful results are obtained only with an analysis of the combination of all modalities' results.

Furthermore, Motor Vehicle Accidents (MVA) constitute a leading cause of injury and death, especially in USA, what has led researchers to study the implementation of methods into human automotive interfaces that are capable of analyzing driver's inattention and stress. Among them, Cruz et al. [20] have designed a system performing early warning that alerts the driver when he/she is in a poor psychological state, especially stressed or inattentive. They have tested their developed system based on an unconstrained and unique brand-new dataset provided by "Motor Trend Magazine" from their best driver car of the consecutive years 2014 and 2015. Authors have suggested a face detector unifying state-of-the-art approaches and controlling the quality of face

detection outputs. It is called reference-based face detection. On the other hand, they presented a new method for facial feature extraction encoding the facial spatio-temporal behavior and removing texture background. This achievement constitutes a promising approach for the automatic observation of driver inattention and stress in real-world applications. The drivers' performances are directly related to their emotional state while driving and this is highly affected by both positive and negative emotions [74]. When sensing anger, the driver underestimates the risk's level, thus increasing accident probability, whereas fear may lead to anxious mood and higher concentration by perceiving a situation as a possible risk. On the other side, positive emotional state implies to follow a risky driving behavior and tempts drivers to show reckless manners [51, 61]. For the reasons mentioned above, monitoring the emotions is important to prevent the bad effect of negative feelings on the driving experience, especially on risk estimation. In this context, [50] have aimed to develop a driving monitoring system detecting emotionally affected drivers and taking over control mainly in critical situations. For data collection, all experiments were done with a research test vehicle FASCAR previously implicated in research by Fischer et al. [29] and aiming the assessment of driver assistance systems. For driving experiments, the scenarios have been carried out on a designated test ground at the DLR Institute of transportation systems compound in Braunschweig in Germany. The presented study has focused on only four frequently occurring emotional states: neutral, positive, frustrated, and anxious. The collected data from all the experiments have been used to train machine learning algorithms. Authors have concluded that their implemented recognition and monitoring system is able to recognize behavioral factors of drivers, while driving cars, to mitigate negative safety impact of negative emotional states of drivers. This has enabled the monitoring of the driver's emotional state during driving.

Education and Learning

Within the framework of education and learning, emotion recognition systems need to operate reliably in an in-the-wild context to accurately recognize emotions in various unstaged and uncontrolled settings. In fact, several external factors, such as social interactions, learning materials, and teachers' attitudes, can significantly and unpredictably influence students' emotions. Thus, detecting and recognizing a wide range of emotional states accurately require a robust and adaptable FER system that can work well in real-world scenarios. Such a system can play an important role in providing personalized learning environments that effectively consider the students' emotional states and cognitive requirements, emphasizing the need for high accuracy and reliability in an in-the-wild context.

In fact, emotions have an important role on students learning and achievement, since they control the student's attention while affecting their motivation to learn and influence their self-regulation of learning. As stated by [54], self-regulated learning and motivation mediate the effects of emotions on academic achievement. Especially, positive emotions positively affect academic achievement, especially when they are mediated by self-regulated learning and motivation. In [25], the authors have developed a facial expression recognition system, based on CNNs, with the aim of being incorporated into e-learning systems. In the light of the prominent results obtained, the authors have integrated the designed solution into an educational game in Morocco, with four students aged between 8 and 12 years (two participants with learning difficulties), to prove the important role of emotion recognition in e-learning. It provides a learning environment which fosters learning and helps students with their learning difficulties by improving some of their elementary skills, such as reading and writing. The participants were invited to play in the game, while the system analyses their emotion in real time, and at the end of the game, the system saves the results. The outcomes show that emotions were detected, and that the system reached state-of-the-art results [25]. Furthermore, since in an e-learning environment, the learners' emotions have a crucial role to detect their concentration level, Krithika.L.B et al. [43] used excite, disturb, and moving pattern of eyes and head to infer a set of information such as detection of the emotional state of learners, and detection of their eyes and head movement, what could prove how much that information are important to evince the interest of learners to the courses topics and their concentration on. The first step in this work consists to detect the face and thereafter the eyes area, to recognize the motion and the status of the learners' eyes. The output of this first step is used as a criterion to decide whether the learner is focused and interested to the topics or not. However, it is usual that the learner is bored, and turns his face or rotates his head, where the eyes cannot be neither visible nor detected. This measure of head rotation from the frames captured strengthens the judgment if the learner is interested or not by the topic, in addition to the measurement of the concentration level. The proposed system has been tested with five students taking a course lecture (video), and the results have been pertinent, showing that the quality of the e-learning could be remarkably enhanced by considering the concentration level. This could be performed by interpreting eye and head movements, and also by detecting the negative emotions during the courses.

Customer Interest and Satisfaction

Consumer interest quantification and satisfaction is an interesting and promising way to conduct marketing research and

business. In this context, it may be necessary for a salesperson to track clients' interests using an emotion recognition system to correctly interpret and understand their behaviors. This could effectively help in observing the consumers' behaviors during the shopping experience, which represents another complex real-world situation. Such FER systems could help service providers by effectively comprehending the needs of the clients while enhancing the services provided. For instance, in [15], an effective method for facial emotion detection based on facial expression has been proposed to recognize customer's satisfaction using machine learning techniques. The method has been designed to assess customer satisfaction using five different classifiers (SVM, random forests, KNN, decision tree, and AdaBoost). In [6], the authors previously conducted a study that revealed an erratic relationship between individuals' capacity for emotion recognition and the outcomes of their work-related activities. They have focused on client satisfaction as a core work-related outcome across professions based on the differential effects of emotion recognition combined to empathy. They have proved that client satisfaction is jointly predicted by service providers' ability to recognize emotions and demonstrate empathy. Likewise, Indira et al. [37] have proposed a methodology for client satisfaction estimation as an alternative option of the ordinary method of gathering clients' reaction. The suggested method has characterized face features using deep convolutional neural networks and Haar cascade classifier. Then, the consumer fulfillment has been classified into one of three classes to show if the client is satisfied, not satisfied, or neutral. In another work proposed by Nethravathi and Aithal [59], for real-time customer satisfaction analysis, the authors have investigated a deep learning-based system composed of three cascaded CNNs for observing customer actions, focusing on interest identification. The proposed method combines facial expressions and head-pose estimation to determine client attention while estimating head posture.

Military Sector

The military sector is an inherently sensitive domain as it is closely tied to security, necessitating heightened scrutiny and increased vigilance. Moreover, the stressful nature of military missions leads generally to several psychiatric problems in military healthcare mainly Post-Traumatic Stress Disorder (PTSD) and depression. To reduce their effects, it is primordial to detect and treat these issues earlier. Nonetheless, the used classical methods, such as psychological examination and interviews, are not as efficient as required by doctors. Therefore, the recourse to automatic detection would be a great help for military personnel facing those psychological problems. In this line, Tokuno et al. [76] have made use of the existing link between emotion change and mental diseases, that is widely applied by doctors to understand mental condition of their patients. They have investigated the detection of emotion change occurring when undergoing mental stress through natural speaking voice. Their investigation was applied in military healthcare after collecting data from volunteering members of military medical staffs who are facing high stress level during their missions. For this purpose, they have used Sensibility Technology (ST) emotion as an emotion analyzing system in speech recognition of Japanese mother tongue. This algorithm of ST emotion is composed of two steps: first, calculating parameters and then conducting decision tree analysis. The obtained results from two study groups have evidenced the change in subjects' emotions. This fact proves the efficiency of such systems for screening of mental status in military situation and eventually in civilian one. In addition, the decision-making is also a crucial step for warfighters as it is often made during moments of extremely distressing situations. This fact could lead warfighters to make decisive judgements that they might regret later because of their tragic consequences. In this framework, Oden et al. [60] have suggested infusing emotional intelligence training into current military training practices and programs that equip them with necessary skills allowing the recognition and then the regulation of their emotions. Authors have exposed the Infantry Immersion Trainer (IIT), which is a training media developed and applied by the United States marine corps while making personnel learn mastering these skills. The benchmark scenarios are supposed to accurately portray emotions and evoke their responses. The IIT replicates a small village in which the stresses and challenges of a small unit operation have been recreated. This training media is combining Hollywood-like sets, scent generators, pyrotechnics, role players, animatronics, projected characters, and immersive environment. Recently, it was upgraded to provide more flexible support to the experimentation including a dedicated network. The IIT includes, currently, a mixed-reality character called "Angry Grandmother" portraying an angry Afghan grandmother of an insurgent giving warfighters experience dealing with uncooperative characters.

In-the-Wild Facial Expression Datasets

Numerous datasets have been produced in real-world environments in order to be as close as possible to the reality conditions [34]. The main objective is to provide suitable data for designing and evaluating reliable systems capable of recognizing spontaneous emotions in real time. In this study, we have summarized the publicly available datasets that illustrate the basic emotion classes while being widely used within the framework of in-the-wild emotion recognition. Table 1 provides an overview of the mentioned datasets,

including the number of images or video samples, the emotion distribution, the image size, the default color space, and additional information (age, type of images…). Since the emotions from the in-the-wild datasets are spontaneous, most of the acquired images are affected by challenging factors, such as different head positions, complex backgrounds, variable distances from the camera, multi-face scenes, subject movements, low lightness conditions, and poor resolutions. In addition to the uncontrolled conditions of the acquisition, the investigated datasets are also characterized by the presence of babies' and children's facial images, occluded faces, and different skin colors and features. In what follows, we present an in-depth analysis of the most prominent facial expression datasets used for in-the-wild emotion recognition, considering their suitability for this context, the challenges associated with their use, the range of emotions they cover, and their potential applications in various domains. This analysis is provided in a scientifically rigorous manner, highlighting the strengths and weaknesses of the studied datasets and arguing for their relevance and usefulness in real-world scenarios (Fig. 1).

The AFEW Dataset

The Acted Facial Expressions in-the-Wild (AFEW) dataset [22] is a dynamic temporal facial expression dataset. It contains video clips collected from 37 movies covering spontaneous expressions, different head poses, wide age, varied face resolutions and focuses, what makes it close to real-world illumination. AFEW is a suitable dataset for the in-the-wild context, since the expressions are performed by actors in a more natural way compared to staged expressions. However, there are still some limitations to its use in real-world scenarios. In fact, the actors in the videos are not necessarily representative of the wider population, and the lighting and other environmental factors may not fully reflect the real-world variability. In terms of challenges, AFEW has a relatively small number of samples compared

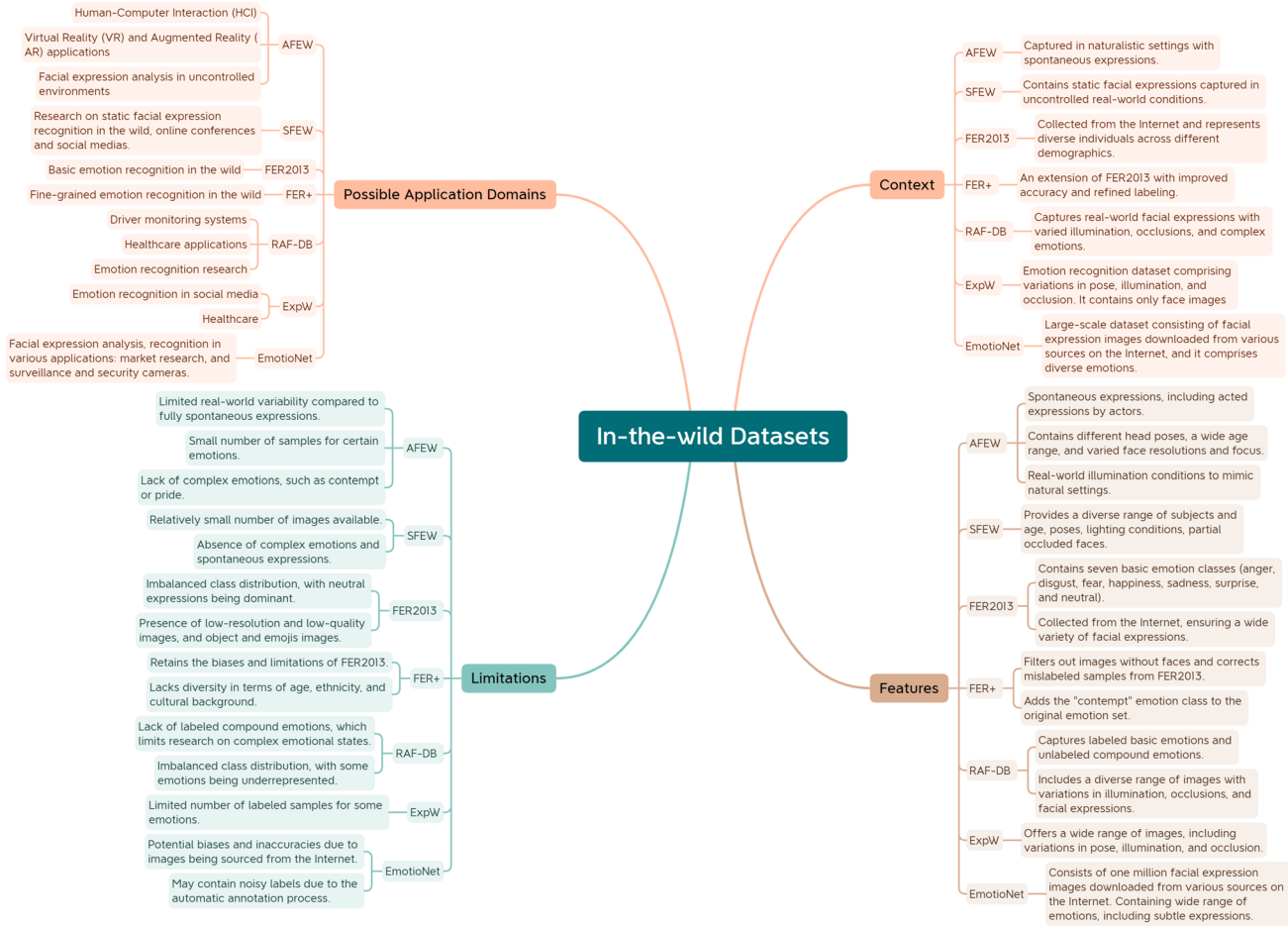

Fig. 1 Overview of the publicly available facial expression datasets for in-the-wild emotion recognition. For each dataset, we have outlined the most crucial detail to be aware of, in terms of "Context" (sources and acquisition settings), "Features" (contents), "Possible Application Domains", and "Limitations"

to some other datasets, what may limit the generalizability of models trained on it. Additionally, the dataset does not include some more complex emotions, such as contempt, which may be important in some applications. Thus, the application domain suitability of AFEW is mainly focused on research in the field of FER in the wild, particularly for developing and evaluating deep learning models. The availability of labeled facial expressions videos and their corresponding emotions makes AFEW a useful resource for the development of more advanced FER systems. Furthermore, the AFEW dataset is not perfect for the in-the-wild context. Its limitations, especially the small number of samples and the lack of more complex emotions, should be considered before adopting it for research or development of FER systems. AFEW dataset can be more suitable for investigations on basic emotions under the in-the-wild context, such as detecting emotions in movies and TV shows.

The SFEW Dataset

The Static Facial Expressions in-the-Wild (SFEW) dataset [21] is a static version of the AFEW dataset. It has been built by extracting static frames from the movies of AFEW. The SFEW dataset is composed of three sets: the training set, the validation set, and the test set, containing 958, 436, and 372 images, respectively. All the images have been captured in the wild with different lighting and environmental conditions. The SFEW dataset is suitable for in-the-wild applications due to its diverse range of subjects, poses, and lighting conditions. However, it has some limitations, such as the fairly low number of instances and the absence of more complex emotions. The SFEW dataset can be used for various applications, such as emotion recognition and facial expression analysis. However, its small size may limit its applicability in certain contexts, and researchers may need to supplement it with other datasets for better performance. Overall, the SFEW dataset is a useful resource for studying basic facial expressions in an in-the-wild context, but it has some limitations and may not be sufficient for more complex applications. It can be more suitable for research on static facial expression recognition under the in-the-wild context, such as identifying emotions in social media profiles or security cameras, as well as in e-learning and online conferences.

It is worth noting that the published descriptions of the SFEW and the AFEW datasets do not explicitly provide detailed statistics or specific information about the ethnicities present in theses datasets. However, it can be inferred that they contain samples from various ethnic backgrounds. Nevertheless, the majority of the samples in these datasets are likely to illustrate ethnicities prevalent in the United States, given that they primarily consist of facial expressions from movies that are predominantly American. However, without a specific demographic information or an explicit breakdown of ethnic representation, it is challenging to provide a comprehensive analysis of the exact ethnicities present in these datasets.

The FER2013 Dataset

The Facial Expression Recognition 2013 (FER2013) dataset [33] is an unconstrained collection gathered by the Google Image Search API, from images available on the Internet. The FER2013 dataset is composed of the seven basic emotion classes containing 4953 images for “Anger”, 547 “Disgust” images, 5121 “Fear” images, 8989 “Happiness” images, 6077 “Sadness” images, 4002 “Surprise” images, and 6198 images for “Neutral”. The dataset has been labeled using crowdsourcing. While FER2013 is a widely used dataset, it has some limitations when it comes to fitting for the in-the-wild context. First, the dataset has a relatively limited diversity in terms of facial expressions and emotions. Additionally, the images in the dataset are not always reflective of real-world scenarios, since they were collected from the Internet which may comprise a significant portion of images taken in controlled environments. Nevertheless, given the fact that it has been collected from the Internet, the FER2013 dataset encompasses subjects with a broader range of age and ethnicity. However, the dataset itself does not provide precise information about the specific age groups or ethnicities represented within it. Therefore, while it can be assumed that the FER2013 dataset includes a diverse range of age and ethnic backgrounds, the lack of exact information makes it difficult to provide detailed insights into the specific demographics of the dataset. Furthermore, the FER2013 dataset has been criticized for having a large portion of mislabeled or ambiguous images, which can significantly affect the accuracy of FER models trained on this dataset. To address these limitations, several modifications of the original FER2013 dataset have been proposed, such as the FER+ dataset, which provides more accurate labels for ambiguous images. Overall, the FER2013 dataset is useful for benchmarking FER models but may not fully reflect the complexity of the in-the-wild context, particularly in terms of the diversity of emotions and environmental factors that can influence facial expressions. It can be more suitable for research on basic emotion recognition under the in-the-wild context, such as analyzing emotions in social media posts or user-generated contents.

The FER+ Dataset

The FER+ dataset [5] is a curated version of FER 2013. The images that do not contain faces have been removed, the original images have been relabeled, and a "contempt" emotion class has been added. The FER+ dataset is splited into three sets containing 25,045 images for training, 3191

images for validation, and 3137 images for testing [30]. One challenge of the FER+ dataset is that the images were primarily collected from Western cultures, which may limit the generalizability of the dataset to other cultures. Overall, the FER+ dataset is well suited for evaluating FER systems in the wild due to its large number of images, more accurate annotations, and the additional emotion of contempt. However, it is important to consider the cultural and emotional limitations of the dataset when applying it to different domains. The additional annotations make it suitable for research on fine-grained emotion recognition within the context of in-the-wild FER, such as detecting subtle emotional expressions in social media or online conversations.

The RAF-DB Dataset

The Real-world Affective Face DataBase (RAF-DB) [44] is a real-world dataset collected from the Internet. It contains 29,672 diverse ranges of facial images, of which 15,339 images are labeled with the seven basic emotion sets (six emotions and neutral) divided into two groups: training set including 12,271 samples and testing set including 3,068 samples. The remaining of the dataset includes 12 other classes of compound emotions, which are not annotated. Images in this dataset are of great variability in terms of subjects' ages, genders, and ethnicities. It covers a diverse range of images, including those captured in natural settings with varying head poses, lighting conditions, occlusions (e.g., glasses, facial hair, or self-occlusion), post-processing operations (e.g., various filters and special effects), and expressions that are not staged or controlled. The RAF-DB dataset is consequently well suited for investigations under the in-the-wild context as it contains a diverse range of images, including those captured in natural settings with varying illumination, occlusions, and expressions that are not staged or controlled. The dataset poses several challenges for FER systems, including dealing with complex emotions, such as contempt and disgust, and recognizing subtle expressions in the presence of noise and occlusion. Furthermore, this dataset contains expressions that are often subtle or mixed with other emotions. The RAF-DB dataset is suitable for several application domains, including driver monitoring systems, where it can be used to detect drowsiness, distraction, or frustration in real-world driving situations. Additionally, it can be used for healthcare applications, such as monitoring pain or anxiety levels in patients during medical procedures. The RAF-DB dataset's suitability for these applications is due to its diversity, which allows for training of robust FER systems that can work well in real-world scenarios. It can also be appropriate for research on emotion recognition under crowded in-the-wild context, such as analyzing emotions in response to music or speeches.

The ExpW Dataset

The Expression in-the-Wild (ExpW) dataset [87] contains 91,793 facial images downloaded from the Web, which illustrate different ethnicities, including the six basic emotions in addition to the neutral one. However, like the previously mentioned datasets, ExpW does not explicitly provide detailed information or statistics regarding the ethnicities of the individuals included in the dataset. The ExpW dataset is particularly well adapted for the in-the-wild context as it comprises a wide range of images including variations in pose, illumination, and occlusion. The dataset was annotated by expert human coders, what ensures high-quality annotations of the emotions depicted in the images. However, one of the major challenges of the ExpW dataset is the class imbalance, since some emotions are under-represented in the dataset compared to others. This can lead to biased or inaccurate results in some cases. Additionally, the dataset does not include information about the temporal dynamics of the facial expressions. This could limit its applicability for tasks that require understanding the temporal aspects of emotional expressions. Nevertheless, the ExpW dataset is suitable for a wide range of applications, including emotion recognition in social media, market research, and healthcare. The dataset can also be employed for training and evaluating machine learning models for affective computing, including deep learning models that can handle the complexities of in-the-wild scenarios.

The EmotioNet Dataset

It is a very large annotated dataset with 1 million facial expression images downloaded from the Internet using WordNet. A total of 950,000 images were annotated by the proposed detection model of [9] and AU intensities. The remaining 25,000 images were manually annotated with 11 AUs. The EmotioNet dataset, which provides 23 classes (six basic emotions and 17 compound emotions), includes a diverse range of samples representing both genders and a majority of ethnicities and races. The EmotionNet dataset has been designed to address the limitations of existing FER datasets, which often lack diversity and do not reflect real-world scenarios. The dataset images are more challenging than those in other datasets due to variations in pose, illumination, and occlusion, making it suitable for training and evaluating in-the-wild emotion recognition models. However, the sheer sizes of the datasets and the diversity of emotions present pose several challenges for researchers. Furthermore, collecting and annotating such a large dataset are a significant task, and ensuring the quality and accuracy of annotations can be difficult. Additionally, training models on such a large dataset can be computationally intensive and may require specialized hardware. EmotionNet is suitable

for a variety of FER applications, including emotion recognition in social media and online platforms, video conferencing, and smart homes. However, due to its large size and complexity, it may be more challenging to use for certain applications compared to smaller datasets. It could be rather appropriate for research on fine-grained emotion recognition under the in-the-wild context, such as detecting subtle.

The AffectNet Dataset

The contraction of Affect from the InterNet (AffectNet) [58] is also a very large dataset covering the cropped facial images, the facial landmark points, and the affect labels. It comprises about 1 million facial images, divided into two groups. The first group contains 450,000 images manually annotated into both discrete categorical and continuous dimensional (valence and arousal) models, while being grouped into eight emotion categories. The second group contains about 550,000 images automatically annotated. The AffectNet dataset includes a wide range of ethnicities and nationalities. However, the dataset does not provide specific information or statistics regarding the exact ethnicities and nationalities represented. The dataset is suitable for training deep learning models for real-world applications such as healthcare, automotive, education, and social media. Nevertheless, the AffectNet dataset has several challenges, such as label noise, class imbalance, and inter-annotator agreement, what requires careful pre-processing and evaluation to ensure the quality and reliability of the dataset. One of the challenges of AffectNet is the presence of label noise, which can arise due to the subjective nature of emotion recognition and the use of crowd-sourced annotations. To address this challenge, the dataset creators have used a validation scheme to identify images with ambiguous labels before excluding them from the training set. Additionally, they have performed a detailed analysis of the label distribution, what allows them to propose a weighting scheme to balance the class distribution. In fact, another challenge of AffectNet is the class imbalance, which can lead to biased models that perform well on dominant emotions but poorly on under-represented ones. To overcome this challenge, the dataset creators have proposed a weighted loss function that assigns higher weights to under-represented emotions during training. They also introduced a fine-grained emotion hierarchy that enables the transfer of knowledge between related emotions.

The Aff-Wild Dataset

Affect in-the-Wild (Aff-Wild) dataset [84] is one of the largest publicly available datasets, with a total of more than 500 videos, which were collected mainly from YouTube and annotated with regards to valence and arousal, and more than 10,000 facial images in-the-wild annotated with regards to 16 FAUs. It illustrates people that react to various situations for measuring continuous affect in the valence-arousal space in-the-wild. The Aff-Wild dataset is challenging, because it contains spontaneous expressions that are not staged or controlled, and the videos have varying lighting conditions, camera angles, and audio quality. Additionally, the dataset includes occlusions, head movements, and non-frontal poses, what can make facial expression recognition more difficult. The Aff-Wild dataset is suitable for various challenging applications, such as emotion recognition, affective computing, and human–computer interaction in real-world scenarios. In fact, the Aff-Wild dataset can be used to develop and evaluate algorithms for facial expression recognition under in-the-wild conditions. The dataset can also be used to study the relationships between facial expressions, emotions, and social contexts, providing insights into human behavior and communication. Overall, the Aff-Wild dataset is a valuable resource for researchers and practitioners in the field of spontaneous emotion recognition, as it provides a challenging and diverse set of real-world facial expressions that can be used to develop and evaluate algorithms for in-the-wild applications. In particular, the Aff-Wild dataset could be utilized for developing FER systems for video gaming and virtual reality applications that can adapt to users' emotional states, providing a more immersive and engaging experience.

The Aff-Wild2 Dataset

Affect in-the-Wild dataset in its second version (Aff-Wild2) is the extended version of the Aff-Wild dataset with 260 more videos and 1,413,000 new video frames. All the videos have been downloaded from Youtube while containing large variations in age, head pose, lighting conditions, ethnicity (Caucasian, Hispanic or Latino, Asian, Black, or African American) and profession (actors, athletes, politicians, and journalists). The additional data have been concatenated to the original Aff-Wild dataset, and the size of the new dataset is equal to 558 videos, with 2,786,201 frames displaying reactions of 458 subjects (279 male and 179 female) from babies and young children to elderly people. The dataset is partitioned into three sets: training, validation, and test sets consisting of 350, 70, 138 videos, respectively, with 1,601,000, 405,000, and 780,201 frames, respectively [40]. One person could appear only in one of the three sets. The Aff-Wild2 dataset is well appropriate for training deep learning models that can handle the complexities of real-world scenarios. The dataset is particularly suitable for emotion recognition and behavior analysis within in-the-wild videos. In fact, it can be used for various in-the-wild applications, such as automotive, pain assessment, and videos surveillance. It is worth noting that the Aff-Wild and the Aff-Wild2

datasets display a big diversity in terms of subjects' ages, ethnicities, and nationalities, and they are also characterized by great variations and diversities of environments (Table 1).

Taxonomy of Existing Deep Learning-Based FER in-the-Wild

One major drawback of conventional FER methods based on hand-crafted features, particularly under in-the-wild environments, can be attributed to the limitations of low-level features in extracting pertinent local facial information. Moreover, these methods often fall short in capturing high-level salient information that is necessary for accurate FER under challenging real-world conditions [12]. This deficiency significantly impacts the overall rate of emotion recognition, notably under uncontrolled in-the-wild environments. Deep Learning (DL) methods have emerged as a potential solution to address the challenges associated with feature extraction and other aforementioned issues. However, even though DL methods possess robust feature learning and extraction capabilities, they still encounter difficulties when applied to facial emotion recognition Nonetheless, deep learning methods necessitate a substantial amount of training data instances to mitigate the risk of overfitting. Furthermore, in the context of facial emotion recognition in the wild, it is crucial to address the issue of high inter-subject variation caused by diverse factors, such as ethnic origins, ages, genders, and expression intensities. Previous research conducted by Zhang et al. [86] has delved into this issue, specifically investigating the intricate relationship between identity and facial expressions. However, distinguishing uncommon expressions remains challenging due to subtle differences. To overcome this challenge, various methods, including convolutional neural networks as well as other deep learning-based architectures [28], have been proposed to reduce intra-class variation. Additionally, the quality of the data significantly impacts the accuracy of facial emotion recognition, since it affects particularly the features extracted. Thus, employing specific data pre-processing techniques becomes essential for enhancing data quality. In this section, we propose a taxonomy for existing relevant in-the-wild FER systems (Table 2). In fact, existing DL-based solutions have been classified according to two primary components: the input and the

Table 1 Summary and characteristics of the reviewed in-the-wild facial expression datasets

Dataset	Nb. of samples	Size (px)	Emotions	Color space	Other information
AFEW	957 video clips	720 × 576	6 basic emotions + neutral	RGB	330 subjects aged 1–70 years, ethnicities prevalent in the United States
SFEW	1,766 images	143 × 181 (aligned) 720 × 576 (non-aligned)	6 basic emotions + neutral	RGB	Subjects aged 1–70 years, ethnicities prevalent in the United States
FER2013	35,887 images	48 × 48	6 basic emotions + neutral	gray scale	many images of cartoons and emojis, broad range of ethnicities
FER+	31,373 images	48 × 48	6 basic emotions + neutral + contempt	gray scale	Non-face images were removed, large range of ethnicities
RAF-DB	29,672 images	100 × 100 (aligned) various sizes (original)	6 basic emotions + neutral + 12 compound	RGB gray scale	Variability in subjects (age, gender, head pose…), lighting conditions, occlusions, and great variability in ethnicity
ExpW	91,793 images	880 × 658	6 basic emotions + neutral	RGB gray scale	Non-face images were removed in the annotation process, diverse ethnicities
EmotioNet	1 million images	various sizes	6 basic emotions + 17 compound	RGB	12 Action Units (AUs) annotated, different ethnicities
AffectNet	1 million images	349 × 349	8 categories	RGB gray scale	Valence and arousal (continuous), about 450,000 subjects, wide range of ethnicities
Aff-Wild	500+ videos with 10,000+ frames	Mean of 1454 × 890	6 basic emotions + neutral	RGB	200 subjects (130 men and 70 women), high diversity in ethnicities
Aff-Wild2	558 videos with 2,786,201 frames	Mean of 1454 × 890	6 basic emotions + neutral	RGB	258 new subjects (149 men and 109 women), 12 AUs and valence and arousal, big diversity in ethnicities

Table 2 Summary of the proposed taxonomy for facial emotion recognition in-the-wild using deep learning techniques

			Techniques	Examples of studies
Input	Pre-processing	Face detection	Google Vision API	[73]
			Mixture of trees method	[1]
			Deepfake face detector	[38]
		Normalization		[85]
		Regions of interest		[63]
				[46]
		Data augmentation	Geometric DA	[58]
			Geometric DA using data generator API	[1]
			Blur, shear, color and contrast variation + geometric DA	[65]
			Color jitter, Color jitter with crop, random crop	[38]
	Modalities	Static	2d images	[47]
				[57]
		Dynamic	3d expressions	[42]
				[48]
		Multimodal		[23]
				[77]
				[82]
Network	Architecture	Single stream		[5]
				[11]
		Multiple stream	Same block	[57]
				[12]
			Different/multiple blocks	[38]
				[41]
				[73]
			Handcrafted+CNN	[65]
		Cascade network		[69]
				[88]
				[83]
	Multitask networks			[69]
				[2]
				[18]
				[35]
	Training strategy		Fine-tune with additional FER datasets	[65]
			Pretrained models with CNN trained from scratch	[1]
	Classification strategy		Fully connected NN	[1]
			kernel SVM, logistic regression	[23]
			partial least squares	[4]
			(LSTM RNN for audio classification)	
			SVM classifier	[13]

network. The input component encompasses diverse FER modalities and pre-processing methods, serving as crucial initial steps in the FER process. On the other hand, the network component focuses on feature extraction and classification. In fact, the network component is divided into three sub-parts that are dedicated to the analysis of most relevant DL architectures, the adopted training strategies, and the employed classification techniques. The proposed taxonomy provides a comprehensive framework for understanding and categorizing in-the-wild FER systems based on their pre-processing techniques, modalities, architecture, training strategy, and classification strategy (Fig. 2). Moreover, Fig. 3 depicts the generic flowchart of a typical emotion recognition system, comprising the standard steps as well as various possible techniques for each step.

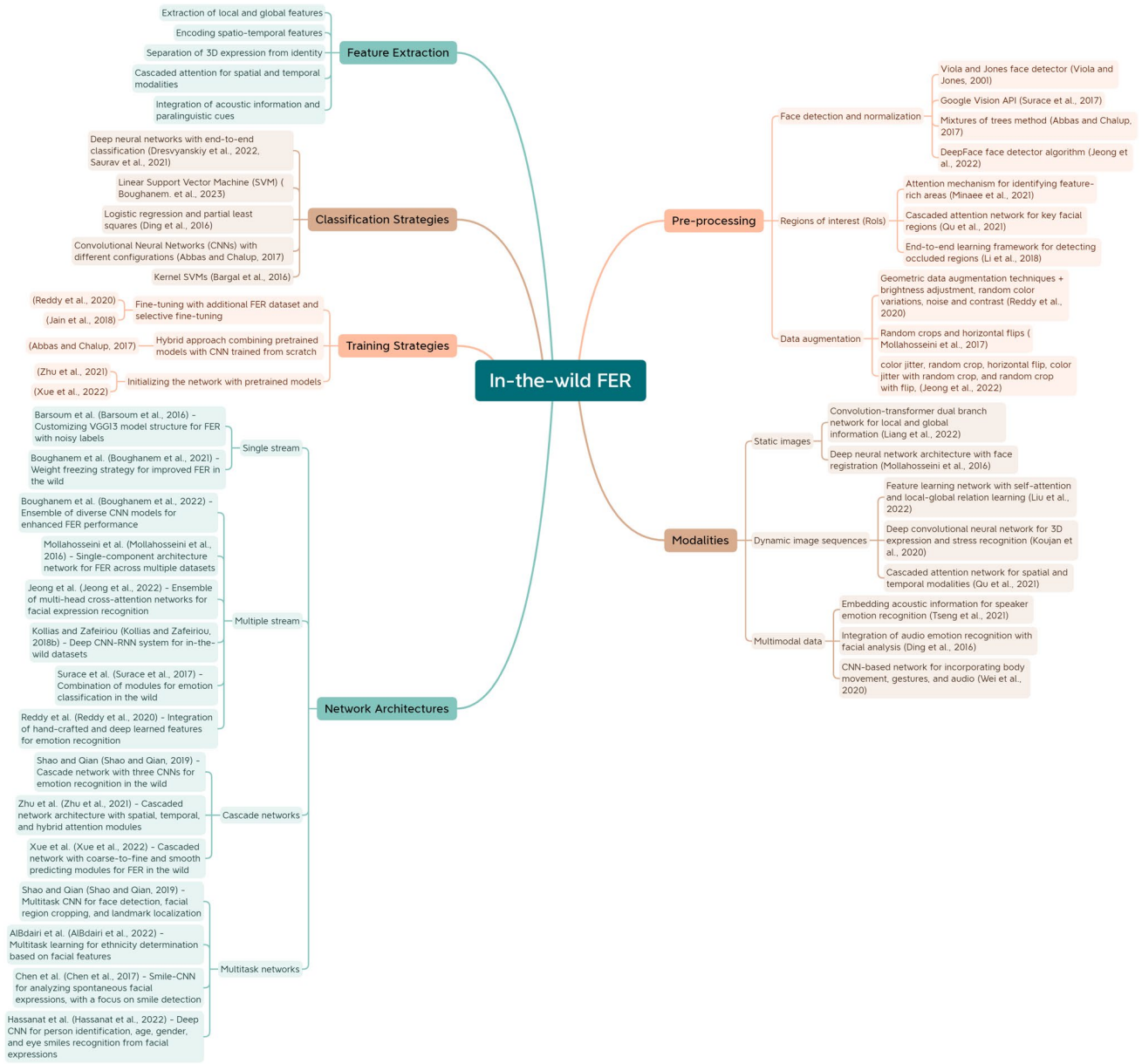

Fig. 2 The proposed taxonomy for existing in-the-wild FER systems through a scheme figuratively representing various hypothetical methods used for in-the-wild emotion recognition. These systems have been associated according to the inputs' modalities, the adopted DL network architectures, the used pre-processing methods, and the adopted strategies for feature extraction, training, and classification

Inputs

The performance of facial emotion recognition methods is profoundly influenced by the quality of the input data. Consequently, prior to initiating the training stage, which aims to extract meaningful features, pre-processing steps are predominantly required to enhance the data quality. Various techniques have been employed for this purpose, including normalization, region of interest selection, and Data Augmentation (DA), among others. These techniques play a crucial role in refining the input data and ensuring thereafter its suitability for subsequent analysis and training processes.

Pre-processing

The pre-processing step operates on the input data to optimize the recognition overall performance of FER in-the-wild. It involves the extraction of the faces from the entire images, by excluding the backgrounds and the non-face parts while keeping only facial parts or regions of interest. The pre-processing step is also required to reduce noise from

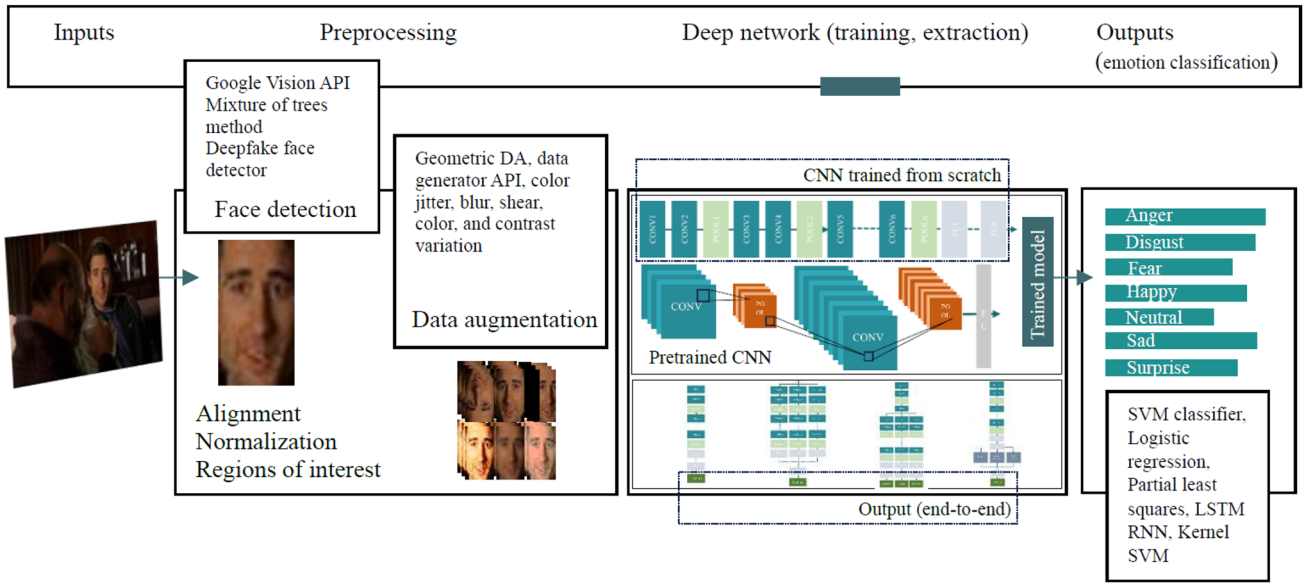

Fig. 3 Flowchart of a typical DL-based feature extraction system for facial emotion recognition

images and to improve the quality before analyzing the facial features.

- **Face detection and normalization**: The selection of an efficient face detection method is essential in the context of FER systems. This step involves separating facial parts from non-facial regions, as well as eliminating background and non-face areas from facial images. Face detection serves as a crucial prerequisite for enabling the learning of meaningful features. The Viola and Jones face detector [80] is a classic and widely used algorithm for face detection, which is known for its computational simplicity, ease of use, and robustness, particularly for frontal faces. However, under the challenging in-the-wild conditions; where factors like large pose variations, multiple faces in a single image, and occluded facial parts are prevalent; the Viola and Jones algorithm may not always be the optimal choice. Thus, many researchers have explored alternative approaches for face detection under such uncontrolled conditions. For instance, Surace et al. [73] have utilized the Google Vision API, which is a commercial library, to perform face detection by identifying a list of frames containing isolated faces. Differently, Abbas and Chalup [1] have focused on classifying the emotion of a group, incorporating face features extracted from the detected faces along with the contextual information of the scene, as the images were captured in unconstrained environments. They have investigated the technique of mixture of trees for detecting faces automatically within an image. Similarly, under uncontrolled conditions, Jeong et al. [38] have used the DeepFace face detector algorithm to accurately select the face region from facial images. Moreover, in addition to face detection, other challenges in FER systems arise from variations in illumination across uncontrolled environments, which can deteriorate the image quality. Furthermore, issues like poor acquisition and digitization conditions could embed noise to the images. To ensure robustness against changes in lighting conditions, the input data are often normalized [85] to achieve better inference generalization. This normalization step plays a crucial role in enhancing the quality and consistency of the input data for subsequent processing and analysis for in-the-wild FER systems.
- **Regions of interest**: In general, conventional practices involve using the entire face as input for the feature learning step in facial emotion recognition systems. However, relying solely on raw data may overlook important information, and certain facial regions can have a detrimental impact on the performance of FER systems. Indeed, factors such as occluded regions in uncontrolled environments can introduce confounding elements to the system. Therefore, it becomes crucial to consider the selective utilization of facial regions to enhance the accuracy and robustness of in-the-wild FER systems. In fact, by focusing only on informative facial regions while excluding potentially misleading or occluded areas, we can mitigate the negative effects of confounding factors and improve thereafter the overall performance of the system. Thus, numerous studies have emphasized the significance of directing in-the-wild FER systems to focus on specific facial regions known as Regions of Interest (RoIs) that contain the most expressive information. For instance, Minaee et al. [55] have highlighted the need to pay more

attention to particular regions to gain a deeper understanding of the underlying emotions in facial images. To address this issue, they have incorporated an attention mechanism that identify feature-rich areas of the face. Besides, they have introduced a visualization technique to identify important facial parts and highlight the most salient regions, enabling the classifier to accurately detect emotions. Similarly, Qu et al. [63] have identified the eyebrows, eyes, nose, and mouth as key regions on the face, while recognizing their significance in reducing the influence of person-specific attributes. They have proposed a cascaded attention network which combines spatial and temporal attention mechanisms designed for video analysis. The temporal attention mechanism selects automatically the peak expressions from image sequences without prior knowledge, while the spatial attention mechanism has focused on crucial facial areas, drawing attention to the key regions. Recently, given the challenges posed by the in-the-wild environments and the negative impact of occluded facial parts on FER systems, Li et al. [46] have developed an end-to-end learning framework. Their approach has employed a convolutional neural network with an attention mechanism capable of detecting occluded regions of faces. Moreover, the mechanism has prioritized the most discriminative non-occluded regions. The authors have reported that their proposed framework has significantly improved the recognition accuracy for both occluded and non-occluded faces. Notably, the method effectively shifted attention from the occluded zones to the related visible regions, further enhancing FER performance in challenging conditions.

- **Data augmentation**: In the context of facial emotion recognition in the wild, the availability of an adequate amount of labeled training data is crucial for deep learning-based methods. However, existing datasets often fall short in providing sufficient samples to ensure promising results for the emotion recognition task. Consequently, data augmentation becomes a necessary and vital part of deep learning-based emotion recognition systems. In fact, most of researchers employ data augmentation techniques to expand the dataset size, thus improving the performance of FER systems, particularly in challenging in-the-wild scenarios. In this context, geometric data augmentation techniques are commonly employed due to their simplicity, efficiency, and computational feasibility. Apart from increasing the number of data instances while reducing the risk of overfitting, data augmentation techniques aim also to enhance the model's robustness against variations in scale, lighting conditions, and minor shifts. For instance, Reddy et al. [65] have incorporated several techniques, such as blur, shear, zoom, brightness adjustment, width and height shift, rotation, random noise, translation, random color variations, random contrast, random distortion, and horizontal flipping. These augmentations were randomly applied to the available images during the training stage. Similarly, Mollahosseini et al. [58] have generated new image instances during training by utilizing five crops of size 224×224 and their corresponding horizontal flips, randomly applied. Likewise, Abbas and Chalup [1] have employed rotation, shifting, zooming, and horizontal flipping, using the image data generator API from Keras, to augment the training set. In their work, Jeong et al. [38] have performed various data augmentation techniques, including color jitter, random crop, horizontal flip, color jitter with random crop, and random crop with flip, to prevent overfitting while improving the generalization capability of the model. Overall, by employing diverse data augmentation techniques in FER systems, researchers aim to address the challenges posed by limited training data, enhance the model's ability to handle variations in the wild, and improve the overall performance and robustness of in-the-wild FER systems.

Modalities

Based on the feature representation, the input modality in facial emotion recognition can be classified into two main approaches. The first approach focuses on static images, and it solely utilizes spatial information from individual images without considering the temporal aspect. In this approach, each image is treated as an independent input, and the relationship between consecutive frames is not considered. On the other hand, the second approach involves dynamic inputs, specifically videos, which take into account the temporal and spatial relationship between consecutive frames in the sequence. This approach recognizes the importance of capturing temporal dynamics in facial expressions by analyzing the changes and patterns across multiple frames. By considering the sequential nature of video data, in-the-wild FER models can extract meaningful information from the temporal domain, complementing the spatial information obtained from individual frames.

- **Static images**: Facial expression images are widely utilized in the field of facial emotion recognition due to the availability of numerous annotated datasets and the rich set of features contained within facial images. Although datasets captured under uncontrolled (in-the-wild) conditions are relatively scarce compared to those obtained in laboratory settings, significant advancements have been made in achieving promising accuracies in emotion recognition using static facial images. For instance, Liang et al. [47] have addressed the challenges posed by occlusions and head-pose variations in FER in-the-wild sce-

narios by leveraging the benefits of both local and global facial information extracted from static images. They have proposed a convolution-transformer dual-branch network consisting of two convolutional neural networks. The first CNN focuses on extracting local region features, such as curves, edges, and lines, while the second CNN captures global features from the original image. These features are then fused to provide a comprehensive representation for accurate classification. Similarly, Mollahosseini et al. [57] have tackled the FER issue, across multiple well-known standard face datasets, by proposing a deep neural network architecture. Their approach has involved using registered static facial images as input and subsequently classifying these images into their respective emotion categories. To achieve face registration, they have adopted a descent method along with linear regression to extract 49 facial landmarks. Overall, facial expression images have been extensively employed in FER research, with notable advancements made in addressing challenges such as occlusions and head-pose variations in uncontrolled environments. Researchers have developed deep neural network architectures that leverage both local and global facial information, along with registration techniques to ensure accurate emotion classification across diverse datasets.

- **Dynamic image sequences** In-the-wild FER systems can benefit from considering the temporal relation among successive frames in facial expression sequences, leading to improved performance. Within the framework of video-based FER, significant advancements have been made in designing efficient and robust systems by incorporating the temporal factor. For instance, Liu et al. [48] have proposed a feature learning network that effectively captures the emotional intensity expressed in video sequences. Their approach utilizes self-attention learning and local–global relation learning modules to encode spatio-temporal features from video clips. By considering the temporal dynamics, their network demonstrates notable improvements in FER performance. Koujan et al. [42] have focused on leveraging the rich dynamic information associated with videos to enhance FER performance. They have addressed the separation of 3D expression from identity by constructing a large-scale dataset comprising facial videos with variations in dynamics, expressions, identities, appearances, and 3D poses. A deep convolutional neural network has been trained on this dataset, showcasing effectiveness in assessing 3D expression parameters and achieving satisfactory results in emotion recognition from facial images, as well as stress recognition from facial videos. Nevertheless, the dynamic changes in facial actions and expression appearances pose challenges for video-based FER, particularly in in-the-wild scenes. In this framework, Qu et al. [63] have proposed a cascaded attention network for facial expression recognition, leveraging both spatial and temporal modalities. The spatial attention module focuses on crucial facial regions, while the introduction of a temporal attention block helps eliminate redundant information across time frames. This approach leads to improved representation by assigning appropriate weights to each step in the temporal dimension, resulting in enhanced performance and effectiveness. Overall, considering the temporal relation among successive frames in video sequences has shown significant improvements for in-the-wild FER performance. Researchers have developed feature learning networks, separation techniques, and cascaded attention networks to effectively capture spatio-temporal features, leading to enhanced accuracy and effectiveness in recognizing facial expressions, particularly under the challenging in-the-wild conditions.
- **Multimodal data**: In addition to static and dynamic methods, various other modalities, such as audio, text, and physiological aspects, can be utilized in hybrid systems to assist in emotion recognition in the wild. In fact, multimodal systems aim to combine multiple input modalities to extract informative features. For instance, Tseng et al. [77] have demonstrated the effectiveness of embedding acoustic information into contextualized lexical representations and incorporating a parallel stream to integrate paralinguistic cues with word meanings, thereby providing crucial affective information for speaker emotion recognition. This multimodal approach has achieved notable performance in capturing emotions. Likewise, in the context of in-the-wild facial emotion recognition from videos, Ding et al. [23] have addressed the challenge by incorporating audio emotion recognition alongside facial analysis using deep neural networks. The integration of both subsystems has proved to be efficient, resulting in state-of-the-art performance. To tackle the specific challenges posed by unconstrained environments, such as head movement and face deformation, Wei et al. [82] have extended the focus beyond facial expressions. They have introduced additional modalities, including body movement, gestures recognition, and audio (speech), to enhance the recognition of emotions in the wild. Through a CNN-based network, multiple features extracted from videos and audios were combined, demonstrating robustness in emotion recognition tasks, particularly in in-the-wild environments. Overall, multimodal approaches integrating various modalities, such as audio, text, and physiological aspects, have shown promise in enhancing emotion recognition systems. These approaches have successfully utilized techniques, such as embedding acoustic information, incorporating parallel streams, and combining different modalities to capture affective cues and improve the performance of

emotion recognition, especially in challenging in-the-wild scenarios.

Deep Learning Networks

In the realm of facial emotion recognition in-the-wild, the effectiveness of the adopted network architecture plays a crucial role in achieving accurate and robust emotion classification. To comprehensively understand and categorize the most relevant architectures in the literature, we present herein a novel taxonomy specifically focusing on the network aspect of FER systems. The taxonomy encompasses diverse architectural paradigms, including single stream, multiple stream, cascade networks, and multi-task networks. Moreover, the investigated architectures have been analyzed based on their design, training strategy, and classification strategy. Unlike traditional taxonomies that often provide limited insights into the network-based approaches, the proposed taxonomy aims to offer a fresh perspective by organizing FER architectures according to their underlying network designs. Indeed, by classifying architectures based on their structural characteristics and integration of various streams, this taxonomy provides a more detailed and nuanced understanding of existing DL-based models for FER in-the-wild. The different architectures are represented in Fig. 4.

Architecture

FER in-the-wild presents several challenges, involving limited availability of datasets and variation in environments, such as illumination changes and head poses. These factors can contribute to some issues like overfitting and reduced performance. To mitigate these issues while enhancing FER performance, researchers have developed various network

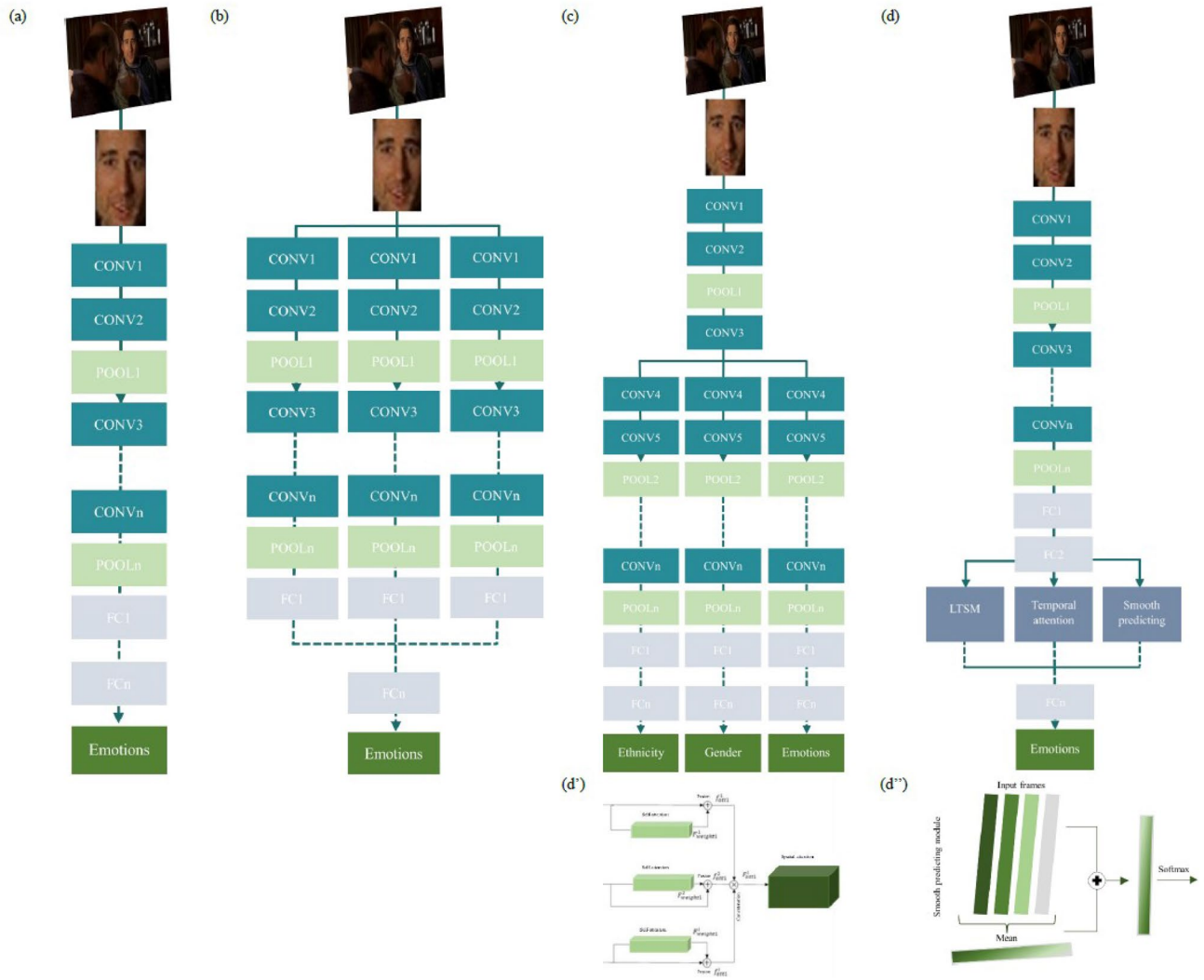

Fig. 4 Deep network architectures: **a** single stream network; **b** multi-stream network; **c** multi-task network; **d** cascaded network with LSTM, temporal attention, or smooth predicting, **d'**: temporal attention module, and **d''**: smooth predicting module

architectures with tailored blocks. In what follows, we will discuss different network architectures that incorporate these blocks to address the challenges of FER within in-the-wild scenarios.

- **Single stream** The single stream architecture serves as a fundamental framework for facial emotion recognition networks. It involves the utilization of a single convolutional neural network to extract meaningful features from facial images, before performing the emotion classification. To address the challenge of noisy labels, Barsoum et al. [5] have conducted a comprehensive study where they have employed facial expression recognition as a representative task. They have demonstrated the effectiveness of learning a deep CNN from noisy labels by customizing the VGG13 model's structure. Through their approach, they have successfully mitigated the impact of label noise while achieving improved performance in FER tasks. Furthermore, Boughanem et al. [11] have explored the potential of their method within the established structure of FER systems. They have focused on leveraging deep features extracted from a CNN model with the aim of introducing a weight freezing strategy in specific shallow layers. This freezing technique has allowed to enhance the performance of FER networks by stabilizing the learning process and reducing overfitting. The suggested method has surpassed several existing state-of-the-art methods, demonstrating significant improvement in spontaneous emotion recognition under in-the-wild conditions.
- **Multiple stream**: Extracting a sufficient number of discriminative features from individual networks may not be adequate to achieve optimal facial emotion recognition performance, especially when dealing within the context of "in-the-wild". To address this limitation, many researchers have investigated the integration of multiple networks within the same architecture to increase the recognition accuracy. Recent studies have focused on this concept, aiming to leverage the resourcefulness derived from multiple intelligent resources rather than relying solely on a single network. For instance, in [12], authors have carefully ensured the complementarity between features by selecting diverse network architectures. The ensemble of networks was formed at the feature level, leveraging three convolutional neural network models with a remarkable number of parameters, leading to efficient learning capability and fast convergence speed. Differently, Mollahosseini et al. [57] have proposed a single-component architecture network to address the FER problem across multiple standard face datasets. Their architecture consisted of two convolutional layers followed by max pooling and four inception layers for each. This design has aimed to reduce the computational complexity required for training the network, resulting in increased classification accuracy across subject-independent and cross-dataset evaluation scenarios. Other researchers have explored the combination of modules with different blocks. For instance, Jeong et al. [38] have designed a uni-task approach for facial expression recognition in the wild using an ensemble of multi-head cross-attention networks. The proposed architecture has comprised two modules: a feature clustering network and attention phases consisting of a multi-head cross-attention network and an attention fusion network. This method has demonstrated promising results with high average performance on validation and test sets. Furthermore, Kollias and Zafeiriou [41] have presented a method that combines deep convolutional and recurrent neural networks. The method incorporates knowledge from related networks trained on different datasets through an extended loss function. This improves performance on diverse datasets without discarding original learning. The obtained results have demonstrated the effectiveness of the proposed method in facial emotion recognition in-the-wild. Additionally, Surace et al. [73] have proposed an effective classification solution for emotions in the wild through the combination of networks. Their architecture has consisted of two modules: a bottom–up module that identifies emotions expressed by isolated faces in images while returning an average emotion estimation, and a top–down module that estimates group emotions using scene descriptors integrated into a Bayesian classifier. The obtained results have shown the superiority of the network system formed by combining different modules compared to using isolated modules. Multiple stream networks can also incorporate a combination of deep learning and hand-crafted features. For instance, Reddy et al. [65] have addressed the challenge of limited cues for identifying distinguishable feature patterns in the wild using hand-crafted facial landmark points as low-level representations, combined with deep learned features. This integration of hand-crafted and deep neural network features has significantly improved the performance of emotion recognition in images captured in unconstrained environments.
- **Cascade networks**: Cascaded networks employ a sequential combination of multiple modules, each assigned with distinct tasks, to construct a deeper network that is both more effective and more efficient. These networks have emerged as a promising architecture for facial emotion recognition in the wild, offering increased effectiveness and efficiency. In fact, many researchers have recognized the superiority of cascaded networks over other architectures and they have demonstrated that combining multi-task CNN modules within this framework can yield significantly improved results. This is exemplified

in the study conducted by Shao and Qian [69], where they have devised a network comprising three convolutional neural networks with distinct architectures for emotion recognition in the wild. Each network module has been dedicated to solving a specific problem. The first CNN has employed six modules to address complex topology and overfitting issues. The second module, a dual-branch CNN, works in tandem with the first module while focusing on extracting both Local Binary Pattern (LBP) features and deep learning features, enabling the estimation of global and local texture features. The third module consists to extract more specific features to overcome limitations associated with the availability of training samples. Similarly, in the study by Zhu et al. [88], a cascade network has been proposed to deal with facial expression recognition from video sequences captured in natural environments. The cascaded network architecture has consisted of three modules: a spatial feature extraction module, a temporal feature extraction module, and a hybrid attention module. The hybrid attention module has been integrated into the cascaded network to recognize facial expressions in videos, leading to effective performance improvements. This integration of the hybrid attention module within the CNN architecture has proved particularly advantageous under many challenging in-the-wild conditions. More recently, Xue et al. [83] have proposed a cascaded network to enhance FER performance in the wild. Their cascaded network has incorporated a coarse-to-fine module, which allows to facilitate the transition from coarse to fine classification, and a smooth predicting module, which further improved performance by capturing both universal and unique expression features.

- **Multitask networks**: In the context of facial emotion recognition in-the-wild, many studies have primarily focused on the singular task of emotion recognition, often overlooking the potential advantages of incorporating other related tasks. However, it has been acknowledged that solely addressing the FER task may not be adequate in certain scenarios. To enhance the performance of in-the-wild FER systems, many recent works have explored the integration of additional tasks within the framework of FER. Notably, facial landmark localization and facial action unit detection have emerged as common examples. For instance, Shao and Qian [69] have employed a multi-task convolutional neural network in the preprocessing stage to detect faces, crop facial regions, and accurately locate facial landmarks, such as the eyes, nose, and mouth corners. Similarly, other studies have embraced multi-task learning approaches to determine the ethnicities of individuals based on facial features [2]. Furthermore, the Smile-CNN [18], a deep convolutional network, has been proposed to analyze spontaneous facial expressions in-the-wild, accounting for factors such as face poses, lighting variations, and scales, with a specific emphasis on smile detection. Likewise, a recent study by Hassanat et al. [35] has introduced a deep convolutional neural network capable of not only identifying person identification but also recognizing age, gender, and eye smiles from facial expressions, thereby addressing multiple related tasks in an in-the-wild context.

Overall, each investigated architecture plays a distinct and contributory role in advancing the field of facial emotion recognition in-the-wild. The single stream architecture acts as the foundational FER network structure, leveraging a solitary convolutional neural network for meaningful feature extraction from facial images, thereby enabling emotion classification. Its design addresses various issues, such as noisy labels, by tailoring established CNN models like VGG13. This yields enhanced FER performance by countering label noise and employing strategies like weight freezing to curb overfitting. These adaptations bolster the accuracy of emotion recognition in diverse real-world settings. Turning to the multiple stream architecture, it tackles the inadequacy of solitary networks in capturing the diversity through distinct features necessary for optimal FER performance amidst wild conditions. By combining multiple networks within a singular framework, researchers harness different intelligent resources in unison. This approach ensures synergy between features derived from varied network architectures. For instance, crafting ensembles of networks with unique architectures empowers efficient learning and swift convergence. This architecture accommodates mechanisms like attention, feature clustering, and blends of deep learning with hand-crafted features. The amalgamation of resources from diverse networks enriches emotion recognition across challenging real-world scenarios. Cascade networks then come into play, constituted by interlinked modules, each tailored for specific tasks, yielding a more profound and potent architecture. This framework particularly suits the intricate challenges of FER in wild settings. The modular setup empowers modules to target different facets of recognition. The inclusion of multi-task CNN modules within cascade networks leads to remarkable performance improvements. These architectures surmount challenges like intricate topology, overfitting, and data scarcity, by partitioning each module to handle specific features, be it local textures or unique expression attributes, thus elevating emotion recognition in diverse real-world contexts. Finally, multi-task networks serve the purpose of integrating complementary tasks alongside FER in the same framework. Recognizing the limitations of singular FER focus, researchers incorporate tasks like facial landmark localization, facial action unit detection, and person identification, amplifying the overall performance of FER systems. These additional tasks deepen

the comprehension of emotions in their contextual backdrop. Multitask CNN architectures can serve pre-processing tasks, such as face detection and landmark localization, rendering the network more versatile in comprehending emotions within a broader framework. In summation, the range of network architectures contributes to the holistic FER process by tackling challenges unique to real-world emotion recognition. The architectures adapt CNNs and employ diverse mechanisms, such as attention, feature clustering, cascading, and multitasking to heighten precision, stability, and generalization across multifaceted environmental conditions, thus culminating in more potent and efficient facial emotion recognition in the wild.

Training Strategy

In the context of facial emotion recognition in the wild, where data scarcity and variability pose significant challenges, researchers have meticulously explored an array of training strategies to significantly enhance model performance. In fact, recognizing facial emotions in real-world scenarios presents formidable challenges due to the limited availability of data and the inherent variations present in natural environments. These challenges underscore the pressing need for innovative approaches that can bolster the accuracy and robustness of FER models. One such pivotal approach is *pretraining*, which constitutes a critical starting point for improving FER models. *Pretraining* entails the initial phase where the neural network is imbued with weights learned from established pretrained models. These models, such as the widely confirmed VGG, GoogleNet, AlexNet, and ResNet, are revered for their adeptness in capturing intricate and high-level features from extensive datasets like ImageNet. *Pretraining* is rooted in the profound idea of harnessing the wealth of insights these models have already accumulated. These pretrained models have undergone extensive training on massive datasets like ImageNet, enabling them to abstract general features that can be readily transferred to a wide array of tasks, including FER. However, the true potential of these models becomes evident when they undergo a crucial subsequent phase: *fine-tuning*. This phase entails the focused refinement of the pretrained model through exposure to task-specific data. For the specific case of FER, this involves training the model on a dataset replete with facial expressions, thereby molding the model's feature extraction capabilities to align with emotion-related attributes. This strategic orchestration offers a twofold advantage. First, it circumvents the challenges of constructing deep neural networks from scratch. Constructing a network from scratch demands a copious amount of labeled data, which often proves scarce in FER scenarios. By commencing with a pretrained model and then subjecting it to *fine-tuning*, researchers can glean benefits from the pretrained model's distilled knowledge and further adapt it to the targeted FER task using a smaller amount of task-specific data. For example, the approach demonstrated by Abbas and Chalup [1] illustrates the potency of combining a meticulously crafted CNN designed from scratch with pretrained models like VGG16 and InceptionV3 that were initially trained on the expansive ImageNet dataset. This novel hybrid approach manifests promising outcomes, surpassing the performance of individual models deployed in isolation. This hybrid amalgamation underscores the potential of integrating the robust representational power of pretrained models with the finesse of task-specific *fine-tuning*, resulting in amplified FER performance. Furthermore, *fine-tuning* strategies have not only been embraced to elevate FER performance but also adapted to confront the challenges posed by the wild and unpredictable nature of real-world scenarios. For instance, in the study conducted by Reddy et al. [65], *fine-tuning* was ingeniously applied by incorporating an additional FER dataset into the training process.

The augmentation of data enriched the model's ability to discern emotions across diverse real-world scenarios, thereby enhancing its generalizability. This signifies the adaptability of the *fine-tuning* strategy to address the wide-ranging variations in complicated conditions. To further expedite training and ensure efficient convergence, a surgical approach termed *selective fine-tuning* has been introduced. This entails training only the higher layers of the DenseNet model, optimizing the utilization of computational resources. This strategic implementation is rooted in the understanding that the lower layers of a deep neural network capture low-level features that are relatively agnostic to the task, while the higher layers encompass more task-specific attributes. By concentrating *fine-tuning* efforts exclusively on the higher layers, the model rapidly adapts to the nuances of the FER task without overfitting to the limited available data. Additionally, the approach of *fine-tuning* has been aptly demonstrated in the context of the XceptionNet architecture using the AffectNet dataset, resulting in the expedited convergence of the loss function [65]. The judicious selection of the architecture for *fine-tuning* holds considerable significance, as different architectures possess varying capabilities to capture features pertinent to FER. In summary, the employment of *pretraining* techniques and strategic *fine-tuning* strategies presents a potent arsenal for researchers striving to surmount the challenges posed by data scarcity and variability in FER. These methodologies tap into the enriched understandings embedded within pretrained models and ingeniously adapt them to the specific demands of FER tasks. As a result, the network's capability to discern emotions accurately and reliably in complex real-world conditions is markedly heightened. The profound significance of these strategies lies in their capacity to bridge the chasm between generalized feature extraction and meticulous

task-specific adaptation. This synthesis results in the cultivation of FER models that are not only more robust but also uniquely suited to excel in a diverse array of challenging real-world scenarios.

Classification Strategy

The final and crucial step in the process of recognizing facial emotions in-the-wild is the classification of the input facial images into specific emotion classes. This step holds the key to translating extracted features into meaningful emotional expressions. The choice of classification strategy greatly impacts the accuracy and the robustness of FER systems. Deep neural networks have emerged as a dominant paradigm, offering an *end-to-end* approach for feature classification, where the classification step is seamlessly integrated with the feature extraction step [24, 66]. This architectural design not only streamlines the computational flow but also enables the network to learn intricate hierarchical representations that capture both low-level details and high-level semantics, a crucial attribute for accurate emotion recognition in the wild. The versatility of convolutional neural networks has led to the exploration of diverse configurations for handling the classification task in FER. Researchers have delved into variations of CNN architectures, each geared toward capturing differentiating features that are vital for emotion recognition [1]. These architectural adaptations are devised to extract *discriminative features* that succinctly represent emotional cues, enabling more precise classification. By harnessing the power of CNNs, FER systems could be able to capture complex patterns present in facial expressions across various lighting conditions, head poses, and background clutter inherent in real-world scenarios. However, the realm of FER classification extends beyond deep neural networks.

Alternative machine learning techniques have also been scrutinized in the quest for accurate emotion recognition in-the-wild. Logistic regression and partial least squares are notable examples of such techniques that have been explored for FER [23]. These methodologies, while distinct from deep neural networks, hold their own advantages and appeal for certain scenarios. They provide insights into the efficacy of traditional machine learning algorithms in addressing the complexity of FER tasks, especially in contexts where neural networks might be challenged by limited data availability or specific data characteristics. Interestingly, researchers have ventured into innovative *alternative classification strategies* that challenge the conventional paradigms. Boughanem et al. [13] have veered from the traditional path by employing a linear Support Vector Machine (SVM) as a substitute for the conventional classification layers in deep networks. This strategic choice arises from the distinctive capacity of SVMs to excel in scenarios characterized by limited instances for classification. By replacing the conventional classification layers with an SVM, the proposed approach exhibits notable recognition rates in natural scenarios, underscoring the efficacy of SVMs, particularly when data scarcity poses a challenge. Moreover, the utility of SVMs is further demonstrated by Bargal et al. [4], who employed a *one-vs-rest* linear SVM to classify facial images with a fivefold cross-validation setup. This approach capitalizes on the ability of SVMs to effectively handle multi-class classification tasks while maintaining computational efficiency. The results underscore the utility of SVMs in FER tasks and suggest their potential to be harnessed for improving recognition accuracy in wild scenarios. The significance of these diverse classification strategies goes beyond mere technological exploration. Each approach offers a unique lens through which the complexity of facial emotion recognition is tackled. Deep neural networks provide a comprehensive end-to-end solution, leveraging hierarchical representations for capturing fine-grained emotional cues.

Traditional machine learning techniques offer insights into the role of statistical methods in the context of emotion recognition, revealing patterns and trends that may complement neural network-based methods. Alternative strategies, like SVM-based approaches, challenge conventional architectures, highlighting the potential for innovative solutions that are tailored to the unique challenges posed by FER in-the-wild. In the broader context of FER research, these distinct classification strategies offer a rich toolbox that researchers can draw upon to tailor solutions to specific real-world challenges. By understanding the nuances of these strategies and their underlying principles, researchers can make informed decisions when designing FER systems that perform accurately and robustly in diverse and challenging real-world scenarios.

Discussion and Insights

Relying solely on human expertise for effective facial emotion recognition systems is challenging, methods exclusively based on hand-crafted features have shown limited performance in uncontrolled environments [8, 70]. The shift to in-the-wild scenarios has prompted researchers to explore deep learning architectures that capture deep features for better performance. Deep learning models excel at extracting deep features and achieving superior representation. In fact, since real-world conditions like non-frontal individuals, occlusions, and imbalanced data distribution require attention, deep learning-based FER systems are meticulously engineered to tackle a myriad of challenges related to in the wild scenarios. They employ sophisticated techniques that encompass the entire spectrum from meticulous pre-processing to intricate classification strategies. In this context,

according to the proposed taxonomy in this study, recommendations and insights are provided specifically for facial emotion recognition in-the-wild. Our focus begins with the available inputs and aims to provide recommendations on how to handle input images even before the training process. Image pre-processing in FER exhibits notable differences between controlled and in-the-wild contexts due to the distinct characteristics of the data and the challenges inherent to each setting. In controlled contexts, where data collections are meticulously planned and executed, pre-processing techniques primarily concentrate on standardizing and normalizing the data to mitigate irrelevant variations. Conversely, in the wild, where images are captured under uncontrolled and diverse conditions, pre-processing techniques must address additional challenges.

First, robust face detection algorithms are necessary to handle the inherent variations encountered in in-the-wild images, such as pose variations, occlusions, multiple faces, and diverse facial expressions. These algorithms enable accurate localization and extraction of facial regions. Second, occlusion handling techniques become crucial in uncontrolled environments, as partial occlusions (e.g., sunglasses, masks, and hair covering parts of the face) are common. Preprocessing methods must employ effective strategies to handle or remove occluded regions while retaining the relevant facial information. Third, in-the-wild images often exhibit noise, compression artifacts, and/or other forms of degradation due to factors like image acquisition, quality, or digitization. Pre-processing techniques may involve denoising, artifact removal, and/or image enhancement methods to improve the overall quality of the images. Additionally, data augmentation techniques in the wild and controlled environments differ in the nature of the variations they aim to address. In controlled environments, data augmentation techniques simulate expected variations within the controlled setting, including changes in lighting conditions, pose variations, facial expressions, and controlled occlusions. The objective is to enhance the model's ability to generalize to similar variations encountered during testing. Conversely, data augmentation in the wild aims to tackle the challenges posed by uncontrolled and unpredictable real-world scenarios. In the wild, images exhibit diverse variations, such as extreme lighting conditions, occlusions, different camera angles, varied backgrounds, and facial appearance variations influenced by factors like ethnicity, age, and gender. Data augmentation techniques in the wild aim to mimic these variations, making the model more robust and capable of handling the complexities present in real-world images. To achieve this, data augmentation techniques in the wild may employ diverse and extensive transformations, including random cropping, scaling, rotation, translation, flipping, noise injection, color variations, and geometric transformations. The goal is to simulate the variability encountered in real-world images, thereby improving the model's generalization and accurate emotion recognition across individuals, environmental conditions, and other potential confounding factors.

When it comes to the input types, facial emotion recognition encompasses a diverse range of modalities that play a crucial role in enhancing its effectiveness in recognizing emotions, especially under the challenging in-the-wild context. The modalities can be classified into static images and dynamic image sequences. Static facial images have been extensively utilized, leveraging deep neural network architectures and incorporating local and global facial information to address challenges, such as occlusions and head-pose variations. However, the limited availability of in-the-wild datasets remains a challenge. Furthermore, dynamic image sequences have shown improvements in FER by capturing temporal dynamics and utilizing spatio-temporal encoding methods. Researchers have developed feature learning networks and cascaded attention networks to effectively extract spatio-temporal features that could enhance accuracy. Challenges in video-based FER include handling variations in facial expressions and achieving robustness in unpredictable scenarios. Multimodal approaches that integrate audio, text, and physiological aspects have shown promise in enhancing emotion recognition systems. These approaches have successfully utilized techniques, such as embedding acoustic information and combining different modalities, to capture affective cues. However, challenges remain in data fusion, modeling interactions between modalities, and ensuring generalizability. Overcoming these issues can advance FER systems and provide valuable insights into human emotions in real-world settings. Overall, by conducting a taxonomy based on architectures and exploring the works within each category, we were able to conclude that FER in the wild requires careful consideration of network architectures to address the challenges posed by varying conditions and limited data. Several architectures, including single stream, multiple stream, cascade, and multi-task networks, have been explored for FER in-the-wild. The single stream architecture, using a single convolutional neural network, offers a straightforward approach for FER. It has shown effectiveness in overcoming noisy label issues and achieving improved performance. However, it may not always capture the diverse range of features that are needed to handle the complexities of the wild environment. On the other hand, multiple stream networks, which combine deep learning with handcrafted features or fuse different network streams, provide the advantage of integrating both learned and predefined features. This can be beneficial for capturing diverse cues and improving FER performance in complex scenarios. The cascaded network architecture, involving the sequential combination of multiple modules, offers the potential for deeper and more effective networks. By assigning different tasks to each module, the cascaded architecture can address

specific challenges and enhance overall FER performance. Nevertheless, it requires careful design and optimization to ensure effective information flow while avoiding information loss between modules. Multitask networks, integrating additional tasks, such as facial landmark localization and facial action unit detection, have the potential to improve FER performance by leveraging shared representations and capturing complementary information. This approach can enhance the robustness as well as the accuracy of FER systems, especially when dealing with complex and diverse facial expressions in the wild.

Considering the unique strengths and challenges of each architecture, there is no one-size-fits-all recommendation for FER in the wild. The choice of architecture should depend on the specific requirements of the application and the nature of the data. It is crucial to evaluate and compare different architectures in terms of their performance, computational efficiency, and robustness to ensure the best fit for the given scenario. Therefore, the selection of network architecture for FER in-the-wild should be carefully considered. Researchers should explore and compare the performance of single stream, multiple stream, cascade, and multi-task networks, while keeping in mind the specific challenges of the wild context. By evaluating the architectures and considering the unique requirements of the application, researchers can determine the most suitable architecture to achieve robust and accurate FER performance in real-world scenarios. Based on the proposed taxonomy and the analysis of existing literature, the selection of appropriate training and classification strategies is crucial for achieving accurate facial emotion recognition in real-world environments. To address challenges such as limited data and overfitting, a combination of fine-tuning pretrained models and ensemble techniques is recommended. Fine-tuning pretrained models, including VGG, GoogleNet, AlexNet, ResNet, and others, have shown effectiveness in mitigating data scarcity and overfitting. By leveraging knowledge learned from large-scale datasets like ImageNet and adapting it to the FER task, these models provide a strong foundation for training reliable FER systems. Studies in the literature have demonstrated promising results by combining pretrained models with models trained from scratch, enhancing the overall performance. Ensemble techniques, which involve combining the outputs of multiple models, can further improve FER performance. By leveraging the diversity and complementary nature of different models, ensembles can capture a broader range of facial features and improve consequently the prediction accuracy. When it comes to the classification strategies, an end-to-end learning approach is well suited for FER in-the-wild. This approach integrates feature extraction and classification within deep neural networks, enabling a holistic representation of facial features. In fact, by learning discriminative features directly from the input data, end-to-end learning allows the network to adapt well to the complexity and the variability of real-world scenarios. Numerous studies have applied this strategy while reporting improved recognition accuracy.

Alternative strategies, such as linear SVMs, logistic regression, partial least squares, and kernel SVMs, have been also explored but may not be as suitable for FER in-the-wild. These methods often rely on hand-crafted features or assumptions that may not fully capture the complexities of real-world environments [7]. End-to-end learning offers the advantage of automatically learning relevant features and patterns directly from the data, making it more adaptable to challenging scenarios. Another alternative is extracting deep features before using them as input to traditional classifiers like SVM or KNN. While this approach can be effective, it may not fully exploit the potential of deep learning models. Deep neural networks are specifically designed to learn end-to-end representations optimized for the task at hand. Using traditional classifiers does not fully harness the discriminative power of deep learning models. In summary, for FER in-the-wild, it is recommended to combine fine-tuning pretrained models, employ ensemble techniques, and utilize an end-to-end learning strategy. This comprehensive approach allows for leveraging prior knowledge, capturing diverse features through model fusion, and automatically learning discriminative representations from the data. By incorporating these strategies, FER systems can achieve higher accuracy and robustness in real-world applications.

To sum up, when designing deep learning-based systems for FER in-the-wild, the choice of the pre-processing techniques, the network architecture, the training strategy, and the classification model, should be made based on the specific application domain and the corresponding in-the-wild context. This is due to the fact that different domains and contexts have different requirements and challenges that need to be addressed. For instance, in healthcare, FER can be used to detect pain in patients. In this domain, ROI extraction and data augmentation techniques can be employed to improve the accuracy of the FER system. ROI extraction can help to focus the system on relevant facial regions, while data augmentation can increase the amount and the diversity of the training data, thereby improving the system's robustness. Differently, in gaming and entertainment, where FER can be used to develop personalized game interfaces that fit to the user's emotional states, multiple stream architectures, and semi-supervised learning can be used to achieve this goal. Multiple stream architectures can take into account both facial expressions and contextual information, such as user inputs and game progress. Semi-supervised learning can leverage a small amount of labeled data and a large amount of unlabeled data to improve the system's generalization ability. In transportation, FER can be used to develop driver monitoring systems that detect

driver drowsiness and distraction. In this domain, cascade networks and FCNNs can be adopted to achieve high accuracy and efficiency. Cascade networks can quickly filter out non-face regions and focus on face detection, while FCNNs can accurately classify facial expressions and detect driver emotions. It is important to note that these are just few examples of the many possible applications of FER in-the-wild systems, and each application domain and in-the-wild context has its unique requirements and challenges that need to be considered. Therefore, the suitability of a particular category or subcategory of deep learning-based FER in-the-wild systems for a specific application domain and in-the-wild context should be carefully evaluated and justified based on scientific rigor. In Table 3, we have compiled a comprehensive overview that encompasses various methods employed in studies focusing on emotion recognition from real-world datasets. This compilation includes recent works as well as other relevant contributions. The table breaks down the datasets used in real-world situations, shedding light on the challenges of recognizing emotions in such contexts. The table also provides the emotion recognition rates achieved by each method, offering a clear comparison of their effectiveness.

Conclusion and Future Direction

A comprehensive overview of facial emotion recognition in-the-wild has been provided in this study, with emphasis placed on the challenges and opportunities associated with this context. The main limitations of classical methods based on hand-crafted features for FER in uncontrolled environments are highlighted, and the shift toward leveraging deep learning techniques to improve FER performance is discussed. Various datasets used in in-the-wild FER are analyzed, considering their appropriateness, challenges, emotion

Table 3 Summary of the performances of most relevant works on facial emotion recognition in the wild

Studies	Methods	Datasets	Accuracy (%)
[69]	Shallow network	FER2013	68
	Dual-branch CNN		54.64
	Pretrained CNN		71.14
[73]	DNN + Bayesian classifiers	EmotiW'17	64.68
[1]	CNN trained from scratch + pretrained CNN	EmotiW'17	72.38
[57]	DNN conceived	FERA	76.7
		SFEW	47.7
		FER2013	66.4
[47]	Dual-branch CNN	RAF-DB	88.40
		FER2013	72.81
		FERPlus	89.17
[42]	CNN + back-end emotion classifier (SVM optimized using Bayesian)	RAF-DB	82.06
[23]	Deep CNN + kernel SVM, logistic regression and partial least squares	EmotiW'16	53.9
[12]	Multiple CNN + SVM	SFEW_2.0	88.20
		FER2013	94.02
[82]	Multiple CNN	Emotion Recognition challenge	92.86
[5]	Multiple deep CNN	FER+	84.98
[88]	Attention cascade network using CNN	AFEW	53.44
[4]		EmotiW'16	59.42
[13]	Multiple CNN + SVM	SFEW	92.28
[45]	Deep CNN bilinear	SFEW_1.0	48.15
		SFEW_2.0	80.34
[72]	Traditional self-attention-based DNN	RAF-DB	88.53
	Multi-scale self-attention-based DNN	FER+	89.52
			88.62
			89.40
[81]	Multiple DNN	GroupEmoW	85.59
		SiteGroEmo	83.57

coverage, and potential applications. An expanded taxonomy of FER in-the-wild is also introduced in this study, with a focus on deep learning methods and the manufacturing steps of a facial emotion recognition system. The importance of accurate face detection and appropriate pre-processing techniques, such as occlusion handling, denoising, and image enhancement, to ensure reliable training data in challenging conditions is discussed. The significance of capturing both global and local facial features for fine-grained expression analysis is particularly emphasized. The imbalanced distribution of data within datasets in the wild context is particularly highlighted as a challenge, and how deep learning approaches can mitigate data scarcity and overfitting is discussed. Attention networks are proposed as a potential solution to enhance FER performance by focusing on unoccluded regions. The potential of multimodal approaches, integrating audio, text, and physiological aspects, can be suggested to improve emotion recognition systems, although challenges remain in data fusion and modeling interactions between modalities. Different network architectures, including single stream, multiple stream, cascade, and multi-task networks, have been evaluated while discussing their strengths and limitations for FER in-the-wild. The need for careful consideration of network architecture based on specific requirements and data characteristics is emphasized. The importance of combining fine-tuning pretrained models, ensemble techniques, and end-to-end learning strategies is also highlighted to further improve FER accuracy and robustness. Insights and recommendations tailored to different application domains and in-the-wild contexts are provided in this paper. For instance, the specific requirements and challenges of healthcare, gaming and entertainment, and transportation domains are discussed, suggesting appropriate techniques and architectures for each scenario. The need for a careful evaluation of the suitability of deep learning-based FER systems in specific application domains has been underlined. Overall, this survey paper can serve as a valuable resource for researchers interested in in-the-wild FER, providing insights into dataset compositions, specificities, and methodological approaches. The comprehensive analysis and recommendations offered in this paper can contribute to the advancement of deep facial emotion recognition, enabling more effective and robust systems for real-world applications. As facial emotion recognition technology continues to advance, the future direction of this work is to prioritize the development of standardized evaluation protocols and benchmarks for FER in-the-wild. Establishing common evaluation metrics and datasets will facilitate fair comparisons between different models and algorithms, fostering advancements in the field. By establishing a shared framework, researchers and practitioners can collaborate more effectively, exchange ideas, and collectively address the challenges of facial emotion recognition in-the-wild. This collaborative effort will contribute to the establishment of best practices and drive the field toward more reliable and universally applicable solutions.

Data Availability Statement For all the data described in this review paper, we have provided references to the original dataset sources. Please note that the availability of these datasets may be subject to certain restrictions or usage terms imposed by the original data providers.

Declarations

Conflict of Interest All authors declare that they have no conflicts of interest.

References

1. Abbas, A. and Chalup, S. K. (2017). Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 561–568.
2. AlBdairi AJA, Xiao Z, Alkhayyat A, Humaidi AJ, Fadhel MA, Taher BH, Alzubaidi L, Santamaría J, Al-Shamma O. Face recognition based on deep learning and fpga for ethnicity identification. Appl Sci. 2022;12(5):2605.
3. Altameem T, Altameem A. Facial expression recognition using human machine interaction and multi-modal visualization analysis for healthcare applications. Image Vis Comput. 2020;103:104044.
4. Bargal, S. A., Barsoum, E., Ferrer, C. C., and Zhang, C. (2016). Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 433–436.
5. Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM international conference on multimodal interaction, pages 279–283.
6. Bechtoldt MN, Beersma B, van Kleef GA. When (not) to empathize: The differential effects of combined emotion recognition and empathic concern on client satisfaction across professions. Motiv Emot. 2019;43:112–29.
7. Bejaoui H, Ghazouani H, Barhoumi W. Fully automated facial expression recognition using 3d morphable model and mesh-local binary pattern. In: Blanc-Talon J, Penne R, Philips W, Popescu D, Scheunders P, editors. Advanced Concepts for Intelligent Vision Systems. Cham. Springer International Publishing; 2017. p. 39–50.
8. Bejaoui H, Ghazouani H, Barhoumi W. Sparse coding-based representation of LBP difference for 3d/4d facial expression recognition. Multimedia Tools and Applications. 2019;78(16):22773–96.
9. Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y., and Martinez, A. M. (2017). Emotionet challenge: Recognition of facial expressions of emotion in the wild. arXiv preprint arXiv:1703.01210.
10. Bissinger, B., Märtin, C., and Fellmann, M. (2022). Support of virtual human interactions based on facial emotion recognition software. In Human-Computer Interaction. Technological Innovation: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II, pages 329–339. Springer.

11. Boughanem, H., Ghazouani, H., and Barhoumi, W. (2021). Towards a deep neural method based on freezing layers for in-the-wild facial emotion recognition. In 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA), pages 1–8. IEEE.
12. Boughanem, H., Ghazouani, H., and Barhoumi, W. (2022). Multi-channel convolutional neural network for human emotion recognition from in-the-wild facial expressions. The Visual Computer, pages 1–26.
13. Boughanem., H., Ghazouani., H., and Barhoumi., W. (2023). Ycbcr color space as an effective solution to the problem of low emotion recognition rate of facial expressions in-the-wild. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,, pages 822–829. INSTICC, SciTePress.
14. Bouzakraoui, M. S., Sadiq, A., and Alaoui, A. Y. (2019). Appreciation of customer satisfaction through analysis facial expressions and emotions recognition. In 2019 4th World Conference on Complex Systems (WCCS), pages 1–5. IEEE.
15. Bouzakraoui MS, Sadiq A, Alaoui AY. Customer satisfaction recognition based on facial expression and machine learning techniques. Advances in Science, Technology and Engineering Systems. 2020;5(4):594–9.
16. Buvaneswari, B. and Reddy, T. K. (2017). A review of eeg based human facial expression recognition systems in cognitive sciences. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pages 462–468. IEEE.
17. Castaño R, Sujan M, Kacker M, Sujan H. Managing consumer uncertainty in the adoption of new products: Temporal distance and mental simulation. J Mark Res. 2008;45(3):320–36.
18. Chen J, Ou Q, Chi Z, Fu H. Smile detection in the wild with deep convolutional neural networks. Mach Vis Appl. 2017;28:173–83.
19. Chimienti, M., Danzi, I., Gattulli, V., Impedovo, D., Pirlo, G., and Veneto, D. (2022). Behavioral analysis for user satisfaction. In 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM), pages 113–119. IEEE.
20. Cruz AC, Bhanu B, Le BT. Human automotive interaction: Affect recognition for motor trend magazine's best driver car of the year. IntechOpen: In Emotion and Attention Recognition Based on Biological Signals and Images; 2017.
21. Dhall, A., Goecke, R., Joshi, J., Sikka, K., and Gedeon, T. (2014). Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In Int Conference on Multimodal Interaction, pages 461–466.
22. Dhall A, Goecke R, Lucey S, Gedeon T. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia. 2012;19(3):34–41.
23. Ding, W., Xu, M., Huang, D., Lin, W., Dong, M., Yu, X., and Li, H. (2016). Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In Proceedings of the 18th ACM international conference on multimodal interaction, pages 506–513.
24. Dresvyanskiy D, Ryumina E, Kaya H, Markitantov M, Karpov A, Minker W. End-to-end modeling and transfer learning for audio-visual emotion recognition in-the-wild. Multimodal Technologies and Interaction. 2022;6(2):11.
25. El Hammoumi, O., Benmarrakchi, F., Ouherrou, N., El Kafi, J., and El Hore, A. (2018). Emotion recognition in e-learning systems. In 2018 6th international conference on multimedia computing and systems (ICMCS), pages 1–6. IEEE.
26. Eltenahy, S. A. M. (2021). Facial recognition and emotional expressions over video conferencing based on web real time communication and artificial intelligence. In Enabling Machine Learning Applications in Data Science: Proceedings of Arab Conference for Emerging Technologies 2020, pages 29–37. Springer.
27. Ertay, E., Huang, H., Sarsenbayeva, Z., and Dingler, T. (2021). Challenges of emotion detection using facial expressions and emotion visualisation in remote communication. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, pages 230–236.
28. Farzaneh, A. H. and Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2402–2411.
29. Fischer, M., Richter, A., Schindler, J., Plättner, J., Temme, G., Kelsch, J., Assmann, D., and Köster, F. (2014). Modular and scalable driving simulator hardware and software for the development of future driver assistence and automation systems. New Developments in Driving Simulation Design and Experiments, pages 223–229.
30. Georgescu M-I, Ionescu RT, Popescu M. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access. 2019;7:64827–36.
31. Ghosh, A., Umer, S., Khan, M. K., Rout, R. K., and Dhara, B. C. (2022). Smart sentiment analysis system for pain detection using cutting edge techniques in a smart healthcare framework. Cluster Computing, pages 1–17.
32. Gogić I, Manhart M, Pandžić IS, Ahlberg J. Fast facial expression recognition using local binary features and shallow neural networks. Vis Comput. 2020;36:97–112.
33. 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. (2013). Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing (ICONIP), pages 117–124.
34. Guerdelli, H., Ferrari, C., Barhoumi, W., Ghazouani, H., and Berretti, S. (2022). Macro- and micro-expressions facial datasets: A survey. Sensors, 22(4).
35. Hassanat AB, Albustanji AA, Tarawneh AS, Alrashidi M, Alharbi H, Alanazi M, Alghamdi M, Alkhazi IS, Prasath VS. Deepveil: deep learning for identification of face, gender, expression recognition under veiled conditions. International Journal of Biometrics. 2022;14(3–4):453–80.
36. Hossain MS, Muhammad G. Emotion-aware connected healthcare big data towards 5g. IEEE Internet Things J. 2017;5(4):2399–406.
37. Indira, D., Sumalatha, L., and Markapudi, B. R. (2021). Multi facial expression recognition (mfer) for identifying customer satisfaction on products using deep cnn and haar cascade classifier. In IOP Conference Series: Materials Science and Engineering, volume 1074, page 012033. IOP Publishing.
38. Jeong, J.-Y., Hong, Y.-G., Kim, D., Jeong, J.-W., Jung, Y., and Kim, S.-H. (2022). Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2353–2358.
39. Joshi, A., Kyal, S., Banerjee, S., and Mishra, T. (2020). In-the-wild drowsiness detection from facial expressions. In 2020 IEEE intelligent vehicles symposium (IV), pages 207–212. IEEE.
40. Kollias, D. and Zafeiriou, S. (2018a). Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770.
41. Kollias, D. and Zafeiriou, S. (2018b). Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In 2018 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.
42. Koujan, M. R., Alharbawee, L., Giannakakis, G., Pugeault, N., and Roussos, A. (2020). Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity. In

2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 24–31. IEEE.
43. Krithika LB, GG, L. P. Student emotion recognition system (sers) for e-learning improvement based on learner concentration metric. Procedia Computer Science. 2016;85:767–76.
44. Li, S., Deng, W., and Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2852–2861.
45. Li T, Chan K-L, Tjahjadi T. Multi-scale correlation module for video-based facial expression recognition in the wild. Pattern Recogn. 2023;142: 109691.
46. Li Y, Zeng J, Shan S, Chen X. Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans Image Process. 2018;28(5):2439–50.
47. Liang, X., Xu, L., Zhang, W., Zhang, Y., Liu, J., and Liu, Z. (2022). A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. The Visual Computer, pages 1–14.
48. Liu Y, Feng C, Yuan X, Zhou L, Wang W, Qin J, Luo Z. Clip-aware expressive feature learning for video-based facial expression recognition. Inf Sci. 2022;598:182–95.
49. Lopes AT, De Aguiar E, De Souza AF, Oliveira-Santos T. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. 2017;61:610–28.
50. Lotz, A., Ihme, K., Charnoz, A., Maroudis, P., Dmitriev, I., and Wendemuth, A. (2018). Recognizing behavioral factors while driving: A real-world multimodal corpus to monitor the driver's affective state. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
51. Lu J, Xie X, Zhang R. Focusing on appraisals: How and why anger and fear influence driving risk perception. J Safety Res. 2013;45:65–73.
52. Lucey, P., Cohn, J. F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., and Prkachin, K. M. (2010). Automatically detecting pain in video through facial action units. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(3):664–674.
53. Malone A, Carroll A, Murphy BP. Facial affect recognition deficits: A potential contributor to aggression in psychotic illness. Aggress Violent Beh. 2012;17(1):27–35.
54. Mega C, Ronconi L, De Beni R. What makes a good student? how emotions, self-regulated learning, and motivation contribute to academic achievement. J Educ Psychol. 2014;106(1):121.
55. Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors. 2021;21(9):3046.
56. Mohan K, Seal A, Krejcar O, Yazidi A. Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Trans Instrum Meas. 2020;70:1–12.
57. Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV), pages 1–10. IEEE.
58. Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2017;10(1):18–31.
59. Nethravathi P, Aithal P. Real-time customer satisfaction analysis using facial expressions and head pose estimation. International Journal of Applied Engineering and Management Letters (IJAEML). 2022;6(1):301–12.
60. Oden KB, Lohani M, McCoy M, Crutchfield J, Rivers S. Embedding emotional intelligence into military training contexts. Procedia Manufacturing. 2015;3:4052–9.
61. Pêcher C, Lemercier C, Cellier J-M. Emotions drive attention: Effects on driver's behaviour. Saf Sci. 2009;47(9):1254–9.
62. Pujol, F. A., Mora, H., and Martínez, A. (2019). Emotion recognition to improve e-healthcare systems in smart cities. In Research & Innovation Forum 2019: Technology, Innovation, Education, and their Social Impact 1, pages 245–254. Springer.
63. Qu X, Zou Z, Su X, Zhou P, Wei W, Wen S, Wu D. Attend to where and when: cascaded attention network for facial expression recognition. IEEE Transactions on Emerging Topics in Computational Intelligence. 2021;6(3):580–92.
64. Rathod P, Gagnani L, Patel K. Facial expression recognition: issues and challenges. International Journal of Enhanced Research in Science Technology & Engineering. 2014;3(2):108–11.
65. Reddy GV, Savarni CD, Mukherjee S. Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features. Cogn Syst Res. 2020;62:23–34.
66. Saurav S, Saini R, Singh S. Emnet: a deep integrated convolutional neural network for facial emotion recognition in the wild. Appl Intell. 2021;51:5543–70.
67. Savaş BK, Becerikli Y. Real time driver fatigue detection system based on multi-task connn. Ieee Access. 2020;8:12491–8.
68. Shang Y, Yang M, Cui J, Cui L, Huang Z, Li X. Driver emotion and fatigue state detection based on time series fusion. Electronics. 2023;12(1):26.
69. Shao J, Qian Y. Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing. 2019;355:82–92.
70. Sidhom O, Ghazouani H, Barhoumi W. Subject-dependent selection of geometrical features for spontaneous emotion recognition. Multimedia Tools and Applications. 2022;82(2):2635–61.
71. Singh, J. (2020). Learning based driver drowsiness detection model. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pages 698–701.
72. Sun N, Song Y, Liu J, Chai L, Sun H. Appearance and geometry transformer for facial expression recognition in the wild. Comput Electr Eng. 2023;107: 108583.
73. Surace, L., Patacchiola, M., Battini Sönmez, E., Spataro, W., and Cangelosi, A. (2017). Emotion recognition in the wild using deep neural networks and bayesian classifiers. In Proceedings of the 19th ACM international conference on multimodal interaction, pages 593–597.
74. Taubman-Ben-Ari O. The effects of positive emotion priming on self-reported reckless driving. Accident Analysis & Prevention. 2012;45:718–25.
75. Tischler, M. A., Peter, C., Wimmer, M., and Voskamp, J. (2007). Application of emotion recognition methods in automotive research. In Proceedings of the 2nd Workshop on Emotion and Computing–Current Research and Future Impact, volume 1, pages 55–60.
76. Tokuno, S., Tsumatori, G., Shono, S., Takei, E., Yamamoto, T., Suzuki, G., Mituyoshi, S., and Shimura, M. (2011). Usage of emotion recognition in military health care. In 2011 defense science research conference and expo (DSR), pages 1–5. IEEE.
77. Tseng S-Y, Narayanan S, Georgiou P. Multimodal embeddings from language models for emotion recognition in the wild. IEEE Signal Process Lett. 2021;28:608–12.
78. Umer, S., Rout, R. K., Pero, C., and Nappi, M. (2022). Facial expression recognition with trade-offs between data augmentation and deep learning features. Journal of Ambient Intelligence and Humanized Computing, pages 1–15.
79. Vij A, Pruthi J. An automated psychometric analyzer based on sentiment analysis and emotion recognition for healthcare. Procedia computer science. 2018;132:1184–91.
80. Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

81. Wang Y, Zhou S, Liu Y, Wang K, Fang F, Qian H. Congnn: Context-consistent cross-graph neural network for group emotion recognition in the wild. Inf Sci. 2022;610:707–24.
82. Wei, G., Jian, L., and Mo, S. (2020). Multimodal (audio, facial and gesture) based emotion recognition challenge. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 908–911. IEEE.
83. Xue, F., Tan, Z., Zhu, Y., Ma, Z., and Guo, G. (2022). Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2412–2418.
84. Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., and Zhao, G. (2016). Facial affect"in-the-wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 36–47.
85. Zhang, F., Zhang, T., Mao, Q., and Xu, C. (2018a). Joint pose and expression modeling for facial expression recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3359–3368.
86. Zhang H, Su W, Yu J, Wang Z. Identity-expression dual branch network for facial expression recognition. IEEE transactions on cognitive and developmental systems. 2020;13(4):898–911.
87. Zhang Z, Luo P, Loy CC, Tang X. From facial expression recognition to interpersonal relation prediction. Int J Comput Vision. 2018;126:550–69.
88. Zhu X, Ye S, Zhao L, Dai Z. Hybrid attention cascade network for facial expression recognition. Sensors. 2021;21(6):2003.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.