INVITED PAPER

Facial Micro-Expressions: An Overview

This article investigates facial expressions on the "micro"-level, providing a detailed survey on a psychological and computer science perspective and the state-of-play in this exciting subfield.

By Guoying Zhao, *Fellow IEEE*, Xiaobai Li, *Senior Member IEEE*, Yante Li, *Member IEEE*, and Matti Pietikäinen, *Life Fellow IEEE*

ABSTRACT | Micro-expression (ME) is an involuntary, fleeting, and subtle facial expression. It may occur in high-stake situations when people attempt to conceal or suppress their true feelings. Therefore, MEs can provide essential clues to people's true feelings and have plenty of potential applications, such as national security, clinical diagnosis, and interrogations. In recent years, ME analysis has gained much attention in various fields due to its practical importance, especially automatic ME analysis in computer vision as MEs are difficult to process by naked eyes. In this survey, we provide a comprehensive review of ME development in the field of computer vision, from the ME studies in psychology and early attempts in computer vision to various computational ME analysis methods and future directions. Four main tasks in ME analysis are specifically discussed, including ME spotting, ME recognition, ME action unit detection, and ME generation in terms of the approaches, advance developments, and challenges. Through this survey, readers can understand MEs in both aspects of psychology and computer vision, and apprehend the future research direction in ME analysis.

KEYWORDS | Affective computing; computer vision; machine learning; micro-expression (ME); survey.

Manuscript received 31 August 2022; revised 21 April 2023 and 28 April 2023; accepted 5 May 2023. Date of publication 5 June 2023; date of current version 11 October 2023. This work was supported in part by the Academy of Finland for Academy Professor project EmotionAI under Grant 336116 and Grant 345122, in part by the University of Oulu & The Academy of Finland Profi 7 under Grant 352788, and in part by the Ministry of Education and Culture of Finland for AI Forum project. *(Corresponding author: Guoying Zhao.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

The authors are with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland (e-mail: guoying.zhao@oulu.fi).

Digital Object Identifier 10.1109/JPROC.2023.3275192

I. Introduction

Emotions are neurophysiological responses to external and/or internal stimuli [1], [2], [3]. They are associated with feelings, thoughts, behavioral responses, and pleasure or displeasure [3], which influence human cognition, decision-making, perception, learning, and so on [4], [5]. Thus, emotions play a crucial role in everyday human life. However, emotional expression and perception are not easy jobs for some people, such as those who suffer from psychological disorders, e.g., alexithymia [6].

In recent years, research on emotions has grown significantly in interdisciplinary fields from psychology to computer science. In the beginning, the affects were mainly studied by psychologists. The concept of affective computing was introduced by Picard [7] in 1997, proposing to automatically quantify and recognize human affects based on psychophysiology, biomedical engineering, computer science, and artificial intelligence. Affective computing aims to endow computers the human-like capabilities to observe, understand, and interpret human affects, referring to feeling, emotion, and mood [7], [8], [9].

Psychological research demonstrates the body language that we use, specifically our facial expressions, and relates to 55% of messages when people perceive others' feelings [10], [11]. To this end, facial expressions are a major channel that humans use to convey emotions. Analyzing facial expressions is meaningful and important, which can be seen from a wide study of facial expressions. However, people may try to conceal their true feelings under certain conditions when people want to avoid losses or gain benefits [12]. In this case, facial micro-expressions (MEs) may occur.

Recent research illustrates that, besides ordinary facial expressions, affect also manifests itself in a special format of MEs. MEs are spontaneous subtle and fleeting facial movements reacting to emotional stimulus [13], [14]. MEs

are almost impossible to control through one's willpower. Learning to detect and recognize MEs is critical for emotional intelligence and has various potential applications, such as clinical diagnosis, education, business interactions, and interrogation. Due to the practical importance of ME analysis in our daily life, researchers have increasing interest in ME analysis recently.

In this article, we review the ME studies from psychology to computer science, focusing on its research progress in computer vision and machine learning. Although there have been a few ME surveys [15], [16], [17], [18], [19], [20], [21] published, they only focus on machine learning approaches of ME spotting and recognition, which are for researchers and professionals in the related field. Different from previous surveys, this overview introduces the general development of ME analysis from psychological study to automatic ME analysis in the computer vision field. The goal is to provide a tutorial that can serve as a reference point for all people who are interested in MEs. We start from the discovery of the ME phenomenon and explorations in psychological studies and then track the studies in cognitive neuroscience about the neural mechanism beneath the behavioral phenomenon. After that, we introduce the technological studies of ME in the computer vision field, from the early attempts to advanced machine learning methods for recognition, spotting, related AU detection tasks, and ME synthesis or generation. Finally, open challenges and future directions are identified.

The rest of this article is organized as follows. Section II presents ME studies in psychology. Section III introduces the early attempts of the computer vision study. Section IV discusses spontaneous ME datasets. Section V reviews computational methods for ME analysis. The open challenges and future directions are discussed in Section VI.

II. ME STUDIES IN PSYCHOLOGY

The research of MEs can be traced back to 1966 when Haggard and Isaacs [22] first reported finding one kind of short-lived facial behavior in psychotherapy that is too fast to be observed with the naked eyes. In their report, these short-lived facial behaviors were referred to as micromomentary facial expressions. This phenomenon was also found by Ekman and Friesen [23] one year after and named it micro-facial expression. Ekman and Friesen studied clinical depression patients who claimed they have recovered but later committed suicide. When examining the films of one patient in slow motion, although the patient appeared to be happy most of the time, a fleeting agony look lasting only 1/12 s was found, which reveals strong negative feelings of the patient. Upon the doctor's questioning, the patient confessed that she was trying to hide her plan to commit suicide. This finding illustrates the existence and essence of MEs that are important behavioral clues revealing human's hidden true emotions.

MEs have different appearance characteristics compared to ordinary facial expressions [referred to as macro-expressions (MaEs)]. MaEs can involve multiple facial muscles with different levels of intensities (see subtle expressions [24] for MaEs with low intensities) and last between 0.5 and 4 s [25], [26], while MEs involve fewer facial muscles (usually only one or two) with very low intensity and short duration (the criteria of ME duration varies according to different researchers, but it is commonly agreed that an ME should be at least less than 500 ms [27], [28]). This is due to the fact that MEs occur under special conditions [23], e.g., when people attempt to cover their true feelings under high-stake situations, some may involuntarily leak in the form of ME. With such constrained conditions and strong intentions to inhibit and disguise, it is natural that MEs present in a condensed and fragmented way. One pair of example figures is shown in Fig. 1 to illustrate the differences between MaEs and MEs.

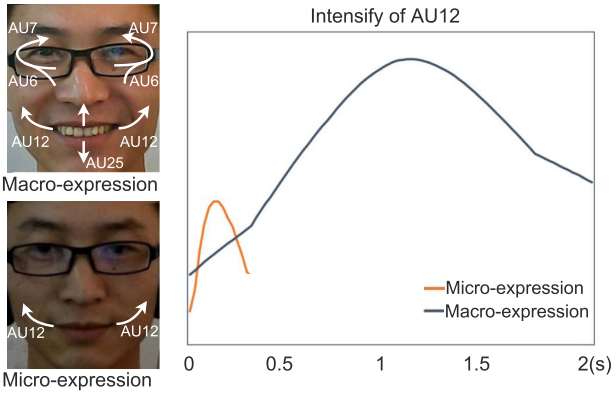


Fig. 1. *Examples of MaEs and MEs from the same person. Compared to an MaE, an ME involves fewer muscles [action units (AUs)] with much lower intensity and shorter duration. AU6, AU7, AU12, and AU25 represent cheek raiser, lid tightener, lip corner puller, and lips part, respectively.*

Ordinary persons can recognize MaEs effortlessly, but it is very challenging, if not completely impossible, to recognize MEs with naked eyes. According to psychological studies [29], [30], without special training, people can only perform slightly better than the guessing chance on ME recognition (MiX). On the other side, MEs are very important behavioral clues for lie detection, and some special occupations (law enforcement and psychotherapy) could be benefited if there is any way to train their staff to better detect and recognize MEs. In 2002, a micro-expression training tool (METT) [30] was developed for such a purpose. It was reported [14] that, after a training of 1.5 h, the trainee's MER ability can be improved by 30%–40%. One limitation of METT is that it is composed of "man-made" ME clips, e.g., by inserting one happy face image into a sequence of neutral faces, which are different from real, spontaneous MEs. Besides, Matsumoto and his colleagues also developed training tools[1] for both MiX and subtle expression recognition (SubX) as it was reported [31] that the ability to read subtle expressions is also related to MER and lie detection. Such tools have been

[1]https://www.humintell.com/

commercialized for both academic research and business purposes.

Another region of psychological research that is closely related to ME is the facial action coding system (FACS). FACS is a comprehensive tool to measure facial movements objectively [32]. FACS encodes and taxonomizes each visually discernible facial muscle movement according to human face anatomical structure, which is called AU. FACS provides detailed descriptions and coding criteria for 28 main code AUs, together with dozens of side codes about eye and head movements. Each action/movement can be coded both with the category (e.g., AU4 and AU12) and the intensity level (A-E: A for the weakest and E for the strongest). FACS and AUs are very important for studies of FER and MER although not all AUs are related to emotions. In FACS Investigator's Guide, an AU-emotion table (page 136) was provided which maps AU combinations to corresponding emotion categories according to observation evidence achieved from psychological studies. Such AU combinations are considered as the "prototypes" of facial expressions, e.g., AU6 + 12 for happiness and AU1 + 2 + 5 + 25 for surprise. However, it is worth mentioning that the mapping table is for MaE emotions but not directly for MEs. Although, in practice, most current ME datasets followed FACS and adopted the same AU-mapping rules as MaEs based on a premise that "ME and MaE share the same emotion categories and appearances," such premise was not systematically verified yet, and the correspondences between AU and ME emotion categories are still open for discussion [33].

The neural mechanisms of MEs are also explored and explained. Two neural pathways [34] originating from different brain areas are involved to mediate facial expressions. One pathway originated from the subcortical areas (i.e., the amygdala), which drives involuntary emotional expression, including facial expressions and other bodily responses, while the other pathway is originated from the cortical motor strip, which drives voluntary facial actions. According to Matsumoto and Hwang [35], when people are feeling strong emotions but try to control/suppress their expressions in high-stake situations, the two pathways meet in the middle and engage in a neural "tug of war" over the control of the face, which may lead to fleeting leakage of MEs.

Over the years, ME study has gained the interest of researchers from various fields. It even inspired the award-winning American crime drama television series "Lie to Me." The show invited the ME researcher to analyze each episode's script and teach actors and staff the science of deception detection. Many episodes of "Lie to Me" referred to the true experiences of MEs.

III. EARLY ATTEMPTS FROM COMPUTER VISION STUDY

The ME analysis topic was first introduced to the computer vision field around the year 2009. Early research on automatic ME analysis mainly concerns ME spotting and MiX. The ME spotting task aims to detect and locate ME occurrences from context clips, and the recognition task aims to classify MEs into emotional categories. ME data are needed to research the two tasks, but no ME dataset was available by that time. It is not easy to induce and collect spontaneous MEs as they only occur on very special occasions. To overcome this obstacle, some early works attempted to build posed ME datasets [36], [37] and used them for evaluating proposed methods for ME spotting and recognition tasks.

Polikovsky et al. [36] collected the first posed ME dataset by asking participants to act facial expressions as fast as possible. Eleven participants were enrolled, and the data were recorded with a pixel resolution of 640 × 480, at a frame rate of 200 frames/s (fps) in a laboratory environment. Each frame was labeled with AUs, and a 3-D-gradient orientation histogram descriptor was proposed for AU analysis in specific facial regions, which was related to the recognition of MEs. However, the authors did not mention the length of such posed MEs, nor did they concern the temporal progress of an ME in their experiments.

Meanwhile, Shreve et al. [37], [38] also conducted work on posed ME data. They built one dataset called USF-HD, in which ME examples were shown to the participants and the participants were asked to mimic the ME motions. Via this way, they collected 100 posed ME samples with a pixel resolution of 720 × 1280 at 29.7 fps. According to the authors, the USF-HD dataset contains both MEs and MaEs (a detailed description of the data is limited), and they proposed a spotting method using spatiotemporal strain, which can spot 74% of all MEs and 85% of all MaEs in the USF-HD dataset.

These studies represent the early attempts of building computing algorithms for automatic ME analysis, which contributed in the way that they helped draw more attention of computer vision researchers to the topic of ME. On the other side, the limitation of these studies is obvious, i.e., the posed MEs are different from real, naturally occurred MEs on their appearances in both spatial and temporal domains; thus, it is questionable whether the methods trained on posed ME data could be helpful in detecting and recognizing real MEs in practice. Early posed ME datasets are not used in current ME studies anymore, and multiple spontaneous ME datasets were built and shared with the research community later, which are introduced in Section IV.

IV. DATASETS

In recent years, several spontaneous ME datasets (SMIC [39] and its extended version SMIC-E, CASME [41], CASME II [42], CAS(ME)^2 [43], SAMM [44], MEVIEW [46], CAS(ME)^3 [47], micro-and-macro expression warehouse (MMEW) [20], and 4DME [33]) have been built, and details of these datasets are summarized in Table 1.

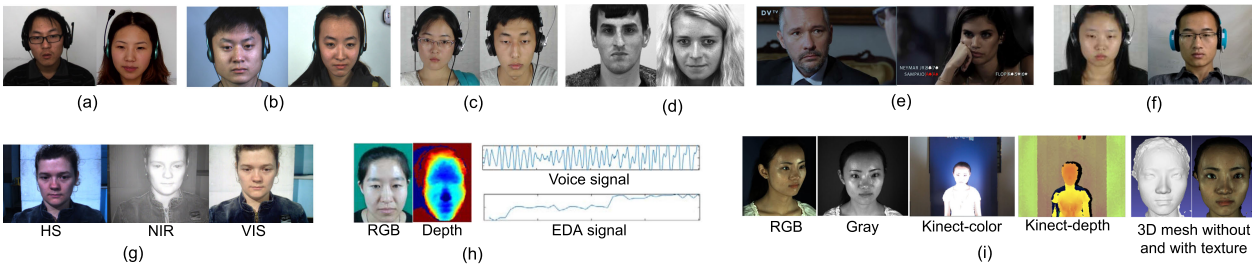


Fig. 2. ***Sample figures of the most popular ME datasets. (a) CASME. (b) CASME II. (c) CAS(ME)^2. (d) SAMM. (e) MEVIEW. (f) MMEW. (g) SMIC. (h) CAS(ME)^3. (i) 4DME.***

One common way to induce spontaneous MEs is to use movie clips with strong emotional clips. Pfister et al. [48] first proposed a protocol for inducing and collecting spontaneous MEs. They asked each participant to watch movie clips with strong emotional content in a monitored lab and asked them to try their best to keep a poker face not to reveal their true feelings. If failed (found by the experimenter via the monitoring camera), there will be a punishment, e.g., filling out a long questionnaire. Many of the later ME datasets followed the same protocol as it was demonstrated to be effective. However, this emotion-eliciting approach also has limitations, and the setup of "movie watching" is simplified compared with practical scenes in daily life. Some studies considered different scenarios for ME data collection. For example, in real life, MaEs and MEs often occur during interpersonal interactions. Husák et al. [46] collected an in-the-wild ME dataset MEVIEW. The samples were from TV series of poker games. Poker games are one kind of high-stake scenario when the players need to hide or disguise their true emotions from their opponents to achieve a win so that MEs are likely to occur. The MEVIEW dataset brought a new sight into the ME study, but, on the other side, there are also constraints of the dataset. First, the videos were recorded for TV shows but not for research, so there were frequent scene changes and also other factors, such as side views and occlusions. Second, the dataset is very small with only 31 video clips from 16 persons, which further limits its usage. Recently, inspired by the paradigm of mock crime in psychology, Li et al. [47] collected one new ME

Table 1 Spontaneous ME Datasets

Database	Subjects		ME video clips				Annotation		Task
	Num	Eth	Num	Resolution	FR	Modality	Emotion	AU	
SMIC [39]	16 8 8	Y	164 71 71	640 × 480 640 × 480 640 × 480	100 25 25	HS (RGB) NIR VIS (RGB)	*Pos* (51) *Neg* (70) *Sur* (43) / *Pos* (28) *Neg* (23) *Sur* (20) / *Pos* (28) *Neg* (24) *Sur* (19)	-	*R,G*
SMIC-E-Long [40]	26	Y	162	640 × 480	100	RGB	-	-	*S*
CASME [41]	19	N	195	640 × 480 1280 × 720	60 60	RGB	*Hap* (5) *Dis* (88) *Sad* (6) *Con* (3) *Fea* (2) *Ten* (28) *Sur* (20) *Rep* (40)	12+	*R,A,G*
CASME II [42]	26	N	247	640 × 480	200	RGB	*Hap* (33) *Sur* (25) *Dis* (60) *Rep* (27) *Oth* (102)	11+	*R,A,G*
CAS(ME)^2 [43]	22	N	57	640 × 480	30	RGB	*Hap* (51) *Neg* (70) *Sur* (43) *Oth* (19)	28	*R,S,A,G*
SAMM [44]	32	Y	147	2040 × 1088	200	Grayscale	*Hap* (24) *Ang* (20) *Sur* (13) *Dis* (8) *Fea* (7) *Sad* (3) *Oth* (84)	ALL	*R,A,G*
SAMM-LV [45]	32	Y	79	2040 × 1088	200	Grayscale	-	ALL	*S*
MEVIEW [46]	16	N	29	720 × 1280	30	RGB	*Hap* (6) *Ang* (2) *Sur* (9) *Dis* (1) *Fea* (3) *Unc* (13) Con(6)	7	*R,A,G*
MMEW [20]	36	N	300	1920 × 1080	90	RGB	*Hap* (36) *Ang* (8) *Sur* (80) *Dis* (72) *Fea* (16) *Sad* (13) *Oth* (102)	17	R,A,G
CAS(ME)^3 [47]	247	N	1059	1280 × 720	30	RGB Depth PS Audio	Hap (992) Dis (2528) Fear (892) Ang (619) Con (401) Sur (1208) Sad (635) Oth (251)	ALLE	*R,S,A,G*
4DME [33]	56	Y	1068	1200 × 1600 640 × 480 640 × 480 640 × 480	60 60 30 30	4D Grayscale RGB Depth	*Neg* (127) *Pos* (34) *Sur* (30) *Rep* (6) *PS* (13) *NS* (8) *RS* (3) *PR* (8) *NR* (7) *Oth* (31)	22	*R,S,A,G*

[1] Modality: RGB indicates 2D color videos; NIR indicates 2D near infrared videos; HS indicates 2D high-speed videos; PS indicates physiological signal.
[2] FR: frame rate
[3] Eth: whether subjects are of multiple ethnicities.
[4] ALL: all observed AUs; ALLE: all observed AUs except eye blinking.
[5] *Sur*: Surprise; *Pos*: Positive; *Neg*: Negative; *Dis*: Disgust; *Hap*: Happiness; *Rep*: Repression; *Sad*: Sadness; *Fea*: Fear; *Ang*: Anger; *Con*: Contempt; *Oth*: Others; *Unc*: Unclear; *NS* Negatively surprise; *PS*: Positively surprise; *PN*: Positively negative; *NN*: Negatively negative;
[6] Task: *R* indicates ME recognition; *S* indicates ME spotting in long videos containing MaEs and MEs; *A* indicates ME-AU detection; *G* indicates ME generation.

dataset CAS(ME)^3, which was a new approach to induce MEs.

The spontaneous ME datasets also update and evolve in the data form. Earlier spontaneous ME datasets only contain frontal 2-D videos as they were relatively easy to collect and analyze, which leads to the fact that most existing ME methods can only analyze frontal faces and are incapable of dealing with challenges in real-world applications, such as illumination variation, occlusion, and pose variations. One research [49] on facial expression recognition illustrated that dynamic 3-D videos with richer information could facilitate facial expression analysis and alleviate the self-occlusion, head motions, and lighting changes problems. Currently, the fast technological development of 3-D scanning makes recording and reconstructing high-quality 3-D facial videos possible. Li et al. collected a 4-D ME dataset (4DME) [33]. Moreover, the 4DME consists of multimodal videos, including Kinect-color videos, Kinect-depth videos, gray-scale 2-D frontal facial videos, and reconstructed dynamic 3-D facial meshes since leveraging multiple modalities is able to provide complementary information to improve the robustness of the analysis. Also, CAS(ME)^3 also consists of depth information, physiological signals, and voice signals in addition to 2-D color videos. Fig. 2 shows the examples in ME datasets.

The MMEW [20], CAS(ME)^2 [43], CAS(ME)^3 [47], and 4DME [33] contain both MEs and MaEs, which can be used to further identify the ME and MaEs. Moreover, the 4DME [33] dataset annotates the cases when MEs are mixed with MaEs based on AUs. This dataset provides the possibility to analyze the co-occurrence and relations of MEs and MaEs.

The above-discussed ME datasets are mostly constructed for MER. For researching ME spotting, some ME datasets have been extended by including non-micro-frames before and after the annotated ME samples to generate longer videos, such as the extended versions of CASME, CASME II, and SMIC. However, the video lengths in these datasets are still quite short, which means that we are only concerning the spotting task under a simplified scenario. Later, CAS(ME)^2 [43] was released which raised the challenge level of the spotting task by introducing long video clips (148 s on average) that include both MaEs and MEs. Another dataset, the SMIC-E-Long [40], was also established for the same purpose, i.e., extend the SMIC-E clips into much longer clips by adding context frames from the original recordings. For such long clips, the challenge of ME spotting lies in multiple aspects, not only that the duration of videos is longer but also other impacts such as eye blinks, head movements, and rotations, MaEs, all become more significant and complex if compared to that within a narrow observation window. Moreover, SAMM-LV [45] and CAS(ME)^3 provide long videos containing annotated intervals of MaEs and MEs according to the AUs. However, the above datasets mainly focus on separated MEs and MaEs even though, in practical situations, the MEs may occur with MaEs. The 4DME dataset emphasizes the MEs and MaEs coexisting situation, which is suitable for studying spotting ME and MaE simultaneously in realistic situations.

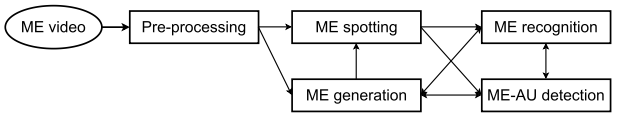


Fig. 3. ***Pipeline of ME analysis.***

V. COMPUTATIONAL METHODS FOR ME ANALYSIS

A general pipeline for ME analysis is shown in Fig. 3. Given raw collected videos, preprocessing is usually the first step to be performed, and then, the ME clips can be detected and located through ME spotting. After that, MiX and ME-AU detection can be carried out as separate or joint tasks. ME generation can synthesize either long videos, including mixed MEs and MaEs or short clips with just MEs, which is expected to benefit different ME analyses. In this section, we first introduce preprocessing steps in ME analysis and discuss different inputs. Four main tasks in nowadays ME analysis with computational methods are discussed, including ME spotting, MER, ME-AU detection, and ME generation.

A. Preprocessing

Given one ME dataset, there are multiple interfering factors of the raw facial videos, which need to be dealt with first before the actual ME analysis could be carried out, e.g., background removal, head pose change, and face size/shape variation. As subtle as MEs are, the ME analysis performance might be significantly impeded if these problems are not solved properly. Like most facial video analysis tasks, ME analysis methods include two "common" preprocessing steps, i.e., face detection and face alignment. The former removes the background and keeps only the facial region, and the latter reduces the variation of facial shapes and poses by aligning corresponding facial landmarks. Moreover, there are two other "special" preprocessing steps that are commonly used in many ME methods as they are specifically helpful for ME analysis. The first one is *motion magnification*. Since MEs have very low intensity, motion magnification can help to enlarge the motion, thus facilitating ME analysis. The second one is *temporal interpolation*. Since MEs are fleeting phenomena with very short duration and various clip lengths, temporal interpolation can help to generate more frames or normalize the video length. The four preprocessing steps, i.e., face detection, face alignment, motion magnification, and temporal interpolation, are each elaborated on in the following.

1) Face Detection: Face detection is the first step in an ME analysis system. Face detection finds the face

location and removes the background. The Viola–Jones face detector [50] is one of the most widely used face detection algorithms. It can achieve robust face detection on near-frontal faces with cascaded weak classifiers. Moreover, it is computationally efficient and could run in real time, so it is widely employed in many face analysis scenarios. However, the Viola–Jones detector does not work well with significant pose variations and occlusions [51]. Wu et al. [52] proposed to utilize probabilistic and pose-specific detector approaches for handling these problems. Later, with the development of deep learning methods, deep learning-based face detection methods have been presented to deal with scale and pose variations [53], [54], [55]. A lightweight deep convolutional network was proposed for robust face detection by using incremental facial part learning [54]. HyperFace proposed in [55] exploited synergy among multiple tasks of landmark localization, face detection, gender recognition, and pose estimation for better face detection performance. Currently, deep learning-based face detection methods have been integrated into popular open- source libraries, such as OpenCV and Dlib, to facilitate facial video analysis tasks, including ME spotting and recognition.

2) Face Registration: MEs are very subtle movements, which could be easily affected by head pose variations. Face registration aims to align each detected face to a reference face according to key facial landmarks so that to alleviate pose variation and head movement problems. Therefore, face registration is an essential preprocess needed for MER. There are multiple methods available for face registration, such as discriminative response map fitting (DRMF) [56], active appearance models (AAMs) [57], and active shape models (ASMs) [58], which are all widely adopted in related studies. AAM [57] is able to match faces with different expressions rapidly, and DRMF is good at handling occlusions in complex backgrounds efficiently in real time [56]. Similar to face detection, deep learning is also exploited for face registration. For example, a deep cascaded framework was proposed by Zhang et al. [59] to exploit the correlation between alignment and detection to further enhance the alignment performance in unconstrained environments, i.e., when various poses, illuminations, and occlusions are involved.

3) Motion Magnification: Motion magnification aims to enhance the intensity level of subtle motions in videos, e.g., the invisible trembling of a working machine. It was found to be helpful for ME analysis tasks and employed as a special preprocessing step. The Eulerian video magnification (EVM) method [60] is one of the most popular used magnification methods [61]. The original method can be used to magnify either color or motion content of an input frame sequence. There is one adjustable parameter for the magnified level, i.e., a larger amplification value leads to a larger scale of motion amplification. For ME magnification, it is not the case that the larger magnification the better. One issue to be concerned with is that, for a very large magnification level, bigger artifacts and more noises are introduced at the same time. According to empirical studies [61], [62], [63], a suitable amplification factor should be about 5–30 depending on different data. Besides EVM, Oh et al. [64] presented a learning-based motion magnification method that was employed to magnify ME in [65]. One advantage of the learning-based motion magnification method is that it causes fewer noises compared to EVM.

4) Temporal Interpolation: Besides low intensity, another challenge of ME analysis is that ME clips are very short and with varying lengths, which is not good for clip-based analysis, especially when recording MEs with a relatively low-speed camera. Temporal interpolation solves this issue by interpolating sequences into a designated length. Specifically, temporal interpolation can be used to upsample ME clips [61] with too few frames to get longer and unified ME clips for stable spatial–temporal feature extraction. Also, extending ME clips and subsampling to multiple short clips with temporal interpolation can be used for data augmentation [66]. The temporal interpolation model (TIM) [67] is one of the most popular methods used in ME analysis, which characterizes the sequence structure by a path graph. Moreover, Niklaus and Liu [68] designed a network for temporal interpolation, which extracts information from pixelwise contextual information of the input frames in order to calculate a high-quality intermediate frame for interpolation. Their method can perform well in complex temporal interpolation scenarios in reality when large motions and occlusions are involved.

B. Inputs

The characteristics of low intensity and short duration make MER very challenging. There can be different inputs for MER in the form of images, videos, or optical flow.

1) Images: A large number of facial expression recognition studies are based on static images because of the convenience of image processing and the availability of massive amounts of facial images. However, different from MaEs, MEs involve subtle facial movements. To this end, Li et al. [62], [69] studied using magnified apex frames for ME analysis. Their experimental results demonstrated that MER with one single apex frame could achieve comparable performance to that of using the whole ME clip. Following [62], [69], Sun et al. [70] further studied apex frame-based MER, which could leverage massive images in MaE databases and turned out to be able to obtain better performance than employing the whole videos.

One motivation for researchers to develop single apex frame-based ME analysis methods is that processing single apex frames reduces computational complexity. However, on the other side, temporal information is lost when using one single frame. Some works [71], [72] proposed to use multiple key ME frames as inputs. ME sequences recorded with high-frame rates (e.g., 200 fps) might contain redundant information for ME analysis, and Liong

Table 2 Comparisons of Inputs for ME Analysis

Input modality		Strength	Shortcoming
Images	Apex [62],[69],[70]	Low computation; Leverage the facial images	Require magnification and apex detection; No temporal information
	Frame aggregation [71],[72],[73],[74]	Leverage key temporal information	Rely on frame selection stratagy
	Dynamic image [76],[77],[78],[79]	Embedding spatio-temporal information	Require dynamic information computation
Video [80],[81], [82],[83],[84],[85],[86]		Process directly	Information redundancy; Various lengths
Optical flow [96],[97],[98]		Remove identity to some degree; Movement considered	Optical flow computation is necessary

and Wong [72] demonstrated that using only the onset and apex frames of one ME as the input could provide sufficient spatial and temporal information for the analysis. Furthermore, Kumar and Bhanu [73] and Liu et al. [74] designed strategies for automatic key frame selection, and the selected key frames are aggregated as the input for MiX.

Embedding dynamic information of a video to a standard image can form dynamic image input, which has been proposed for action recognition [75]. Considering that the dynamic image can summarize appearance and subtle dynamics into an image, multiple MER methods [76], [77], [78], [79] employed dynamic image as input and achieved promising performance. The dynamic image can simultaneously consider spatial and temporal information, and keep computational efficiency by processing only one image.

2) Video Input: ME video clips are commonly utilized input for ME analysis [80], [81], [82], [83], [84], [85], [86]. It considers spatial and continuous temporal information simultaneously and can be processed directly without extra operations. However, the ME videos have short and varying duration. Many approaches employed TIM to interpolate the ME clips to longer and/or the same length [61]. There are methods using spatial–temporal descriptors, such as local binary patterns from three orthogonal planes (LBP-TOP) [81], long short-term memory (LSTM), and some methods utilizing LSTM and recurrent neural networks (RNNs) [87], which aim to process the time-series data with various duration. However, there is redundancy in ME sequences, and the computational cost is relatively high [88].

3) Optical Flow: Another widely used input is optical flow. Optical flow estimates the local movement between images by computing the direction and magnitude of pixel movement in image sequences [89]. Optical flow has been verified as effective for movement representation. In recent years, various optical flow computation methods have been proposed [90], [91], [92], [93], [94], such as Lucas-Kanade [90], Farnebäck's [93], TV-L1 [94], and FlowNet [95]. Inspired by the effectiveness of optical flow, many ME analysis methods utilized optical flow to represent the micro-facial motion [96], [97], [98]. Furthermore, the optical flow can reduce the identity characteristic to some degree [96]. MER approaches based on optical flows often outperform MER approaches based on appearances [96], [97]. However, current optical flow-based MER approaches mainly employ the traditional optical flows with complicated feature operations leading to slow computation.

Moreover, considering the strengths of different inputs, some works combined multiple inputs to learn multiview information to further improve the performance [84], [97], [99], [100]. The summarization of the input strengths and shortcomings is shown in Table 2.

C. ME Spotting

As earlier discussed, ME spotting is one of the main tasks in automatic ME analysis, which identifies temporal locations (sometimes also the spatial places in faces) of MEs in video clips. Three keyframes are included in a complete process of one ME, i.e., the onset, the apex, and the offset. The onset is the first frame in which the ME motion is first discriminable. The frame with the highest motion intensity in the ME clip is the apex frame. The offset is the frame marking the end of the motion. ME spotting could be identifying the keyframes in long clips. Spotting is the very first step before many other ME analysis tasks, such as MER and ME action unit detection but even more challenging. There are human test experiments in the research of [61], which indicate that automatic MER technology can outperform humans, while the performance substantially decreases when including the ME spotting step in the completed ME system.

The ME community has carried out preliminary research in ME spotting. The second Facial Micro-Expression Grand Challenge (MEGC 2019) proposed challenges for ME spotting in long videos [101]. The long videos contain a lot of non-ME movements, such as eye blinking, weak head rotation, swallowing, and MaEs, which are similar to practical situations. Furthermore, the third MEGC 2020 [102] developed the challenge to spot both MaE and ME from long videos. In this section, we discuss ME spotting in terms of heuristic and machine learning-based methods followed by a discussion.

1) Heuristic ME Spotting: Traditional algorithms are usually training-free, heuristic, and spot MEs by comparing feature differences in a sliding window with fixed-length time [61], [62], [111]. The location of MEs can be determined by a thresholding method. LBP [61], [111], HOG [61], and optical flow [103], [112], [113], as shown in Fig. 4, are the commonly used features for ME spotting,

Table 3 Comparison of Heuristic and Machine Learning-Based ME Spotting Methods

	Method	Year	Input	feature	Spotting	SMIC-E-HS	CASME II	CAS(ME)^2	SAMM
Heuristic	[103]	2015	OF	Flow vector	Threshold	AUC: 95	-	-	-
	[104]	2016	OF	MDMO	Threshold	-	-	F1: 0.3348	-
	[61]	2018	Video	LBP	Threshold	AUC: 83.32	AUC: 92.98	-	-
	[105]	2020	OF	Flow vector	Multi-Scale Filter	-	-	F1 ME: 0.0547 F1 MaE: 0.2131	F1 ME: 0.1331 F1 MaE: 0.0725
Machine learning based	[106]	2016	Video	Landmark	Adaboost (5-folds)	AUC: 86.93	-	-	-
	[46]	2017	Video	Intensity change	SVM (LOSO)	AUC: 88	AUC: 97	-	-
	[107]	2018	Video	CNN	Feature matrix processing (5-folds)	-	ACC: 71.97*	-	
	[108]	2019	Video	HOG	LSTM (LOSO)	F1: 0.62	F1: 0.86	-	-
	[109]	2020	Video	LBCNN	SVM (LOSO)	-	-	F1: 0.0595	F1: 0.0813
	[110]	2021	Video	2+1D CNN	Classification regression (LOSO)	-	-	F1: 0.026	F1: 0.049

[1] OF: Optical flow.
[2] F1: F1-score; AUC: Area under ROC curves; ACC: Accuracy.
* Spotting the apex frame in long videos.

which are specifically discussed in this section. Li et al. [61] proposed a training-free ME spotting approach based on LBP difference contrast in 6×6 blocks and a sliding window. Moreover, baseline results of ME spotting in long videos were provided in this study [61]. On the other hand, Patel et al. [103] located the apex, onset, and offset frames in ME clips by computing the motion amplitude shifted over time in optical flow. Main directional maximal difference analysis (MDMD) was proposed [104], [113] to spot MEs based on the magnitude of maximal difference in the main direction of optical flow. Similar to [61], MDMD also utilized a sliding window and block-based division. Ma et al. [114] employed the oriented optical flow histogram to further improve the apex frame spotting performance. To distinguish MaEs and MEs in long videos, Zhang et al. [105] proposed to disentangle head movement by computing the mean optical flow of the nose region and utilizing a multiscale filter to increase the ability to spot MEs and MaEs.

All of the above methods spot MEs in the spatial–temporal domain. However, MEs have rapid and low-intensity spatial movements, which are not obvious in the spatial–temporal domain but can lead to large changes in the frequency domain. To this end, frequency-based ME spotting methods [62], [69] were presented to spot the apex frame in the ME sequence by exploiting information in the frequency domain, which can reflect the rate of facial changes.

The strength of feature difference-based approaches is that the approaches consider the temporal characteristic according to the size of the sliding window, and the spotting results can be simply obtained by setting thresholds. However, as they are heuristic, mainly based on practical experience regarding, e.g., a threshold in the feature difference value, the spotting results can be easily influenced by the other facial movements with similar intensity or duration, such as eye blinks. Thus, it is hard to distinguish MEs from other similar facial movements with feature-difference-based ME spotting methods.

2) Machine Learning-Based ME Spotting: As the heuristic methods based on thresholds are weak to distinguish MEs from other facial movements, machine learning-based methods were developed to tell apart different facial movements, which regards ME spotting as a binary classification of non-ME and ME frames. In general, these methods first extract features of each ME frame, and a classifier is utilized to recognize the ME frames.

Husák et al. [46] computed an image intensity change descriptor on RoIs and applied an SVM classifier to spot MEs. Xia et al. [106] and Borza et al. [115] both used an Adaboost model to classify ME frames for ME spotting. Specifically, the former utilized geometric features of landmarks of face shapes, and the latter employed motion descriptors based on absolute frame differences.

Inspired by the successful application of deep learning in action detection [116], Zhang et al. [107] first proposed to search the ME apex frame in long ME videos thorough adopting a convolutional neural network (CNN) with feature matrix processing. Later, LSTM networks were introduced in [108] and [40] to spot MEs in long videos due to their strength in processing sequences with various lengths. Moreover, Wang et al. [110] presented an end-to-end deep framework consisting of three modules: clip proposal, $2 + 1$D CNN, and classification regression, to extract features, propose ME clips, and classify MEs, respectively. Moreover, a local bilinear structure-based network was proposed to extract local and global features in a fine-grained way to identify MEs and MaEs [109].

3) Discussion: In general, the heuristic ME spotting methods are mostly based on thresholds to classify MEs and non-MEs, which are weak to distinguish the MEs from other facial movements, such as eye blinks. The machine learning-based methods can recognize different facial movements by training classifiers. However, the performance of ME spotting is restricted by the small-scale ME datasets and unbalanced ME and non-ME samples.

With the increase in ME spotting research, various evaluation protocols have been proposed, using different training and testing sets, and various metrics, such as the area under the curve (AUC), the receiver operating characteristic (ROC) curve, mean absolute error (MAE), accuracy, recall, and F1-score [18]. It causes inconsistencies and makes fair comparisons very hard, as shown in Table 3. Thus, in the future, it is essential to design standard

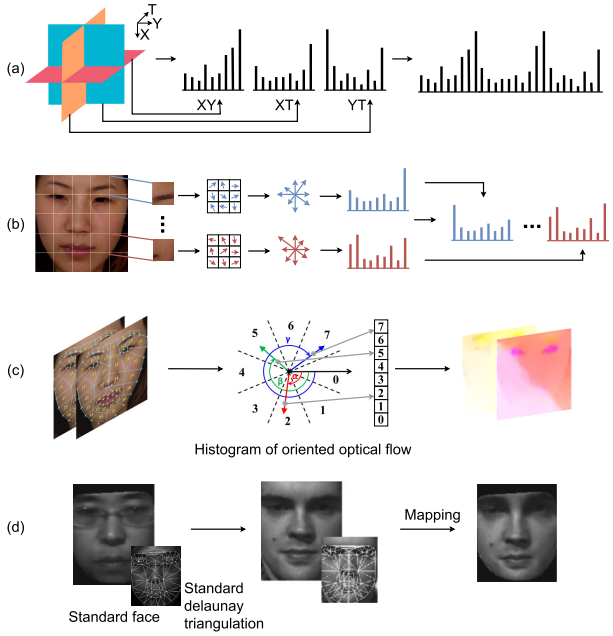


Fig. 4. ***Handcrafted features for ME analysis. (a) LBP-TOP [48]. (b) HOG [117]. (c) Optical flow [118]. (d) Delaunay triangulation [119].***

evaluation protocols for ME spotting as the first step has been taken in Tran et al.'s [40] benchmark work. In addition, multiple studies attempted to explore spotting MEs and MaEs simultaneously in long videos. Due to the presence of noise, irrelevant movements, and mixed MEs and MaEs, it is very challenging to learn discriminative features on limited datasets and accurately locate the various MEs and MaEs that should be further studied.

D. MER

The spontaneous MER research with computational technology can be traced to the work of [48]. Pfister et al. [48] proposed to use spatiotemporal local texture descriptors combined with a TIM. Later, various methods were developed for efficient MER. In the beginning, most of the MER methods are based on traditional handcrafted features. In recent years, the fast development of deep learning technology enables deep learning-based methods to archive the state-of-the-art performance in MER. In this section, we discuss the MER methods in terms of traditional learning and deep learning methods.

1) Traditional Learning Methods: Due to most ME datasets with limited samples, handcrafted features are widely researched in MER. The handcrafted features represent the image details without explicit semantic knowledge/meaning, such as intensities [120] and gradients [61]. Basic classifiers such as KNN and SVM are employed to classify the features. In this section, we mainly discuss the LBP-TOP and its variants, gradient variants, and optical flow-based approaches, which are widely used in MER, as shown in Table 4.

The most popular appearance feature utilized for MER is LBP-TOP [81]. LBP-TOP is a texture descriptor thresholding the neighbors of each pixel with a binary code in spatial and temporal dimensions. Most of the MER research employs the LBP-TOP as their baseline due to its computational simplicity, as shown in Fig. 4(a). Later, several LBP-TOP variants were proposed to meet the different needs of MER [120]. Wang et al. [120] presented a spatiotemporal descriptor to enhance the efficiency of MiX by suppressing the redundancy information of LBP-TOP with six intersection points (LBP-SIP). Then, a more compact variant with less computational time, LBP-MOP, was proposed. LBP-MOP concatenates the LBP features from the temporal pooling results of image sequences in three orthogonal planes [121]. Huang et al. [122] designed a spatiotemporal completed local quantized pattern (STCLQP), which considers not only the pixel intensity but also the sign, magnitude, and orientation components.

Aside from LBP, another widely used feature for MER is the gradient-based feature. High-order gradients could represent the structure information of an image in detail [123]. The histogram of gradients (HOG) is one of the most widely used features for its ability to describe the edges in an image with geometric invariance [123], as shown in Fig. 4(b). Li et al. [61] developed the histogram of image gradient orientation (HIGO) ignoring the magnitude weighting of the first-order derivatives to depress the influence of illumination affect. Moreover, both HOG-TOP and HIGO-TOP were utilized together in [124] to further improve the performance of MER.

Considering the strength of optical flow discussed in Section V-B, several works designed feature descriptors based on optical flow for MER. Liu et al. [118] proposed a main directional mean optical flow (MDMO) that considers both local information and its location through regions of interest (RoIs), as shown in Fig. 4(c). The MDMO only exploits the dominant direction of optical flow in each RoI. However, facial motions spread progressively because of the elasticity of the skin. Allaert et al. [125] presented to extract the coherent movement of the face from dense optical flow to better describe facial movement. Inspired by the strength of optical strain in capturing small facial deformation, Liong et al. [126] proposed to apply optical strain to the MER task. Optical strain computes the shear and normal strain tensor components of optical flow. To reduce the dimensionality and enhance computational efficiency, Liong et al. [126] resized and max-normalized the strain maps to a relatively low resolution to keep consistency across the database. To effectively learn ME information from the active regions, optical strain weighted (OSW) features were presented to weight local LBP-TOP features according to the temporal mean-pooled optical strain map [127]. Moreover, Liong and Wong [72] designed a biweighted oriented optical flow (BI-WOOF) descriptor adding local and global weighting to HOOF by optical strain and magnitude values, respectively, in order to reduce the noisy optical flows. In contrast to the

above works [72], [126], [127], Happy and Routray [128] proposed a fuzzy histogram of optical flow orientation (FHOFO) collecting the motion directions and ignoring motion magnitudes due to the low intensity of MEs. A fuzzy membership function was utilized to map the directions of motion into angular bins to create smooth histograms for motion representation.

Besides the above features, there are descriptors representing MEs in other views, such as color [129] and facial geometry [119], [130]. Wang et al. [129] proposed a tensor-independent color space (TICS) method. In TICS, the RGB color is transformed into independent color components and combined with dynamic texture to increase MER accuracy further. Lu et al. [119] designed a Delaunay-based temporal coding model (DTCM) to normalize the ME sequences temporally and spatially based on Delaunay triangulation to suppress the personal appearance influence, which is not relevant to MEs, as shown in Fig. 4(d). Furthermore, a facial dynamics map (FDM) [130] was proposed to handle subtle face displacements by characterizing the movements of an ME in different granularity.

Once features are extracted, classifiers are used to categorize the MEs. Classification involves two stages: training and testing. In the training stage, classifiers learn to recognize MEs based on the labels and extracted features. In the testing stage, the trained classifier's performance is evaluated by evaluation metrics, such as accuracy and F1 score. Various supervised classification methods have been used for MER, e.g., support vector machine (SVM) [131], Adaboost [132], random forest [133], k-Nearest Neighbor (kNN) [128], [134], and linear discriminant analysis (LDA) [135]. SVM is the most widely used classifier because of its robustness, accuracy, and effectiveness especially when the training samples are limited.

2) Deep Learning Methods: In recent years, deep learning has achieved excellent performance in many research fields, such as facial expression recognition [138], object detection [139], and image classification [140]. Several researchers have attempted to explore MER with deep learning. However, deep learning is a data-driven method that requires a large amount of data to learn a robust representation. MEs have small-scale datasets and low intensity, which makes MER based on deep learning hard. Patel et al. [141] first attempted to utilize facial expression and object-based CNN models and selected relevant deep features for representing MEs. However, its MER accuracy on CASME II is 47.3%, much lower compared to handcrafted descriptors. Following Patel's work [141], various deep learning-based methods have been proposed to improve MER. To date, deep learning-based MER has achieved state-of-the-art performance by leveraging massive facial images and designing effective network structures and special blocks. In the following, we will first discuss the MER methods of taking advantage of existing data through fine-tuning, learning from multiview data, and transfer learning. Then, effective structures and blocks designed for learning discriminative ME features are specifically discussed. Finally, we introduce the losses applied in MER.

In the beginning, most MER works adopted well-designed classical convolutional networks, such as ResNet family [140], [142], [143], Inception network [144], [145], [146], AlexNet, and VGG-FACE [62], [147]. The effectiveness of these networks has been verified on common tasks, such as image classification and face recognition. Furthermore, these networks are pretrained on datasets with a large number of images, such as ImageNet and VGG-FACE datasets [147]. Fine-tuning deep networks pretrained on large datasets can effectively avoid the overfitting problem caused by small-scale ME datasets [62], [74], [77].

To further leverage the information on the limited ME samples, multiple works adopted multistream network structures to extract multiview features from various inputs. The optical flow features from the apex frame network (OFF-Apex-Net) [148] built a dual-stream CNN for MER based on optical flow-derived components, the strain along the horizontal and vertical directions. Khor et al. [149] proposed a dual-stream shallow network (DSSN) based on heterogeneous features. Moreover, other works developed multiple substreams to extract features from frame sequences, static images, optical flow, or RoIs [85], [150], [151], [152]. Song et al. [97], [136] designed three-stream CNN (TSCNN) models extracting features from the apex frames, local facial regions, and optical flow between onset, apex, and offset frames to leverage the information of the static spatial, local, and dynamic temporal information, as shown in Fig. 5(a). In addition, other works developed multiple substreams to extract features from frame sequences and optical flow, or RoIs. She et al. [152] employed three RoIs and global regions, and designed a four-stream model to explore local and global information. To further explore the temporal information of MEs, several works [88], [153], [154] cascaded CNN and RNN or LSTM to extract features from individual frames of ME sequence and capture the facial evolution of MEs.

Recent research demonstrates that taking advantage of the information on relevant tasks could also benefit facial expression recognition [139]. Inspired by this finding, to make full use of the information on faces, multiple research developed multitask learning for better MER by leveraging different side tasks [76], [155]. Nie et al. [76] designed a GEnder-based MER (GEME) incorporating gender detection task with MER, as shown in Fig. 5(b). Furthermore, Zhou et al. [146], [156] proposed to recognize AUs and MEs and further aggregated AU representation into ME representation to improve MER performance. Other methods leverage the knowledge of other tasks through transfer learning. Directly fine-tuning on a pretrained model is the simplest. Besides fine-tuning, knowledge distillation is also widely applied to MER [70].

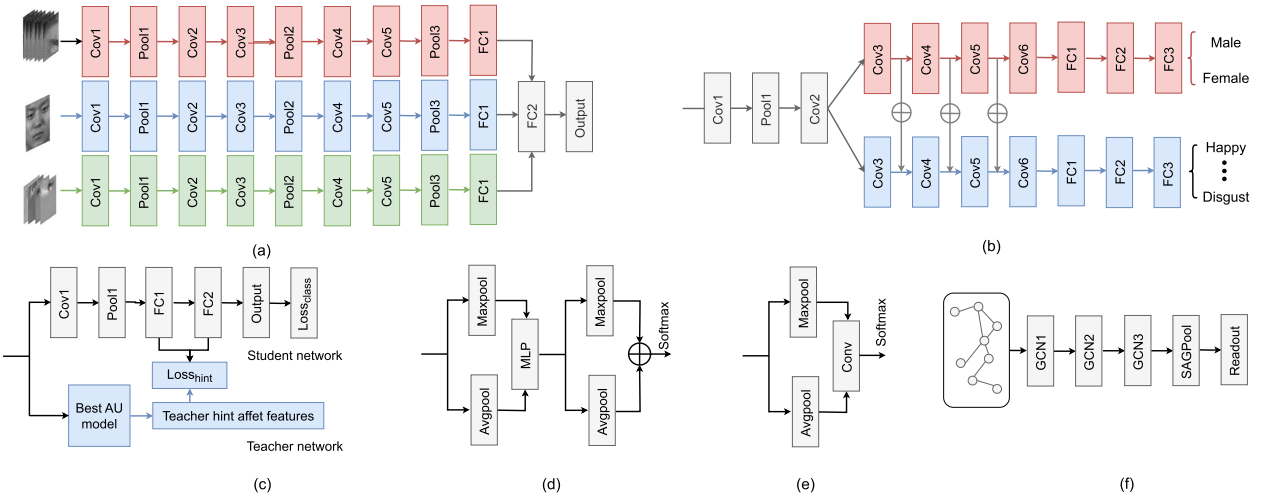


Fig. 5. ***Demonstration of improving MER performance through different methods. (a) TSCNN employing a multistream structure to leverage multiview inputs [97], [136]. (b) Knowledge distillation [70]. (c) GEME with multitask learning framework [76]. (d) Channel attention module [137]. (e) Spatial attention module [137]. (f) Graph on AU representation [83]. Different colors in the figure represent different paths in the network.***

Knowledge distillation utilizes pretrained high-capacity networks to guide the training of small and fast networks [157]. Sun et al. [70] guided the shallow network learning for MER by mimicking the intermediate features of a network trained for AU detection and facial expression recognition. Fig. 5(c) illustrates the knowledge distillation process. However, it is not reasonable to directly mimic the MaE representation, as the appearances of MEs and MaEs have differences. To this end, instead of features, Zhou et al. [158] proposed to transfer attention to improve MER. Another effective transfer learning approach is domain adaptation that obtains domain invariant representations through embedding domain adaptation into the deep learning pipeline. The adversarial learning strategy is adopted in MER to narrow down the gap between the MEs and MaEs, and leverage massive MaE images to boost the MER performance [159], [160], [161].

Besides utilizing more data, many studies designed effective shallow networks for MER to avoid overfitting [82], [162], [163]. Zhao and Xu [164] designed a six-layer CNN and utilized a 1×1 convolutional layer to process the ME input to increase the nonlinear representation. Liong et al. [165] designed a Shallow Triple Stream Three-dimensional CNN (STSTNet) with two layers to learn features from optical flow features computed from the onset and apex frames in each ME video clip. Other works trim multiple convolutional layers of the deep network to achieve a shallow network [149], [166].

Since MEs involve fewer facial muscles (usually only one or two) with low intensity, MEs are related to changes [167] in RoIs. In order to emphasize learning on RoIs and reduce the influence of information unrelated to MEs, multiple works introduced attention modules [168], [169], [170], [171], [172], [173]. Inspired by the squeeze-and-excitation blocks [150] adaptively learning the weights of each feature channel, channel attention was employed in MER with spatiotemporal attention to improve the representational ability of MEs [66], [137], [174], [175], as shown in Fig. 5(d) and (e).

In addition, MEs perform as specific combinations of multiple facial AUs. The latent semantic information among facial changes has an important contribution to MER. The graph convolutional network (GCN) has been verified to effectively model the semantic relationships. Inspired by the successful application of GCN in face analysis tasks, the research [65], [83], [146], [176] applied the GCN to model the relationship between the local facial movements. Specifically, Lei et al. [65], [177] designed graphs based on the ROIs along facial landmarks, while [73], [83], [146], and [176] built graphs on AU-level representations to infer the AU relationship and boost MER performance, as shown in Fig. 5(f).

Deep networks apply loss functions to perform end-to-end classification. The loss function penalizes the deviation between the predicted and true labels during the learning process. Most ME analysis works directly utilize softmax cross-entropy loss that is widely used in classification tasks [178]. However, ME datasets suffer from low interclass differences due to the low intensity of MEs. Contrastive loss [179], triplet loss, and center loss [180] were introduced to MER to increase intraclass compactness and interclass separability of MEs [69], [160]. In addition, the samples in ME datasets have imbalanced distribution since some MEs, such as fear, are difficult to trigger. The focal loss was employed to alleviate the issue by focusing on misclassified and hard samples [76], [96], [146].

3) Discussion: MEs are involuntary, rapid, and subtle facial movements. The main challenge for robust MER is how to effectively extract discriminative representations.

Table 4 Traditional Learning-Based MER on SMIC, CASME, and CASME II

Method	Year	Pre-processing		Input	feature	Classifier	Protocol	SMIC			CASME			CASME II		
		Mag	TIM					Cate.	F1	ACC	Cate.	F1	ACC	Cate.	F1	ACC
[120]	2014	-	-	Video	LBP-SIP	SVM	LOVO	3	-	66.40	-	-	-	5	-	62.80
[132]	2014	-	✓	Video	LBP-TOP+STM	Adaboost	LOSO	3	0.4731	44.34	-	-	-	5	0.3337	43.78
[121]	2015	-	-	Video	LBP-MOP	SVM	LOSO	3	-	50.61	-	-	-	5	-	43.72
[122]	2016	-	-	Video	STCLQP	SVM	LOSO	3	0.6381	64.02	4	0.5	57.31	5	0.5836	58.39
[61]	2017	✓	✓	Video	HIGO	SVM	LOSO	3	-	68.29	-	-	-	5	-	67.12
[118]	2016	-	✓	OF	MDMO	SVM	LOSO	3	-	80.8	4		68.86	5		64.83
[126]	2014	-	-	OF	OS	SVM	LOSO	3	-	53.56	-	-	-	-	-	-
[128]	2017	-	✓	OF	FHOFO	SVM	LOSO	3	0.5182	51.22	4	0.5409	65.99	5	0.5197	55.86
[129]	2014	-	-	OF	TICS	SVM	LOSO	3	-	59.79	4	-	61.86	5	-	60.82
[130]	2017	-	✓	OF	FDM	SVM	LOSO	3	0.5380	54.88	4	0.2401	42.02	3	0.2972	41.96

[1] Mag: Magnification; TIM: Temporal interpolation model.
[2] OF: Optical flow;
[3] Cate: Category; F1: F1-score; ACC: Accuracy.

Handcrafted features are low-level representations that are able to effectively describe the texture, color, and so on while being weak in extracting high-level semantic information. In contrast, deep learning-based features are abstract high-level representations.

As shown in Tables 4 and 5, in the beginning, MER works used handcrafted features. In recent years, with the development of deep learning, most of the current MER methods are based on CNNs, and the deep-based MER achieves state-of-the-art performance. The performances of MER are influenced by various factors, such as preprocessing, features, and network structure. It is difficult to directly compare the methods of every step. However, from the experimental results, the general trends of MER can be found.

In general, the preprocessing step could benefit the MER for both traditional learning and deep learning-based MER approaches. Fig. 6 shows a comparison of the performance with TIM and magnification. From Fig. 6, we could draw a conclusion that magnification and TIM can benefit MER. However, the suitable magnification and temporal interpolation factor should be further studied.

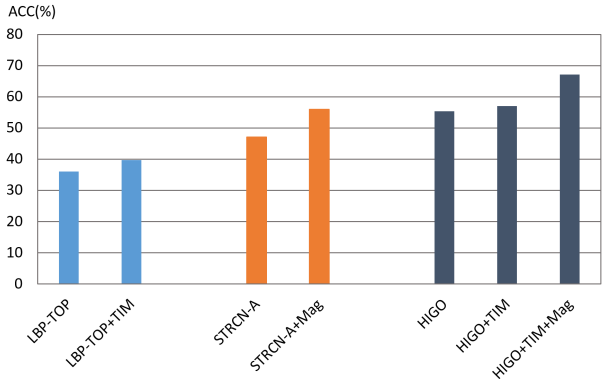


Fig. 6. *Comparisons of MER performance with/without TIM and/or magnification. Specifically, results are based on LBP-TOP and LBP-TOP + TIM on CASME II [181], STRCN-A and STRCN-A + Mag on CASME II [96], and HIGO, HIGO + TIM, and HIGO + TIM + Mag on SMIC [61].*

Deep learning is a data-driven method. The small amount of MEs is far from enough to train a robust network. Current MER research designed shallow networks or leveraged massive MaE images to solve the data limitation problem. The shallow network and transfer learning have achieved big development for MER on small-scale ME datasets, such as STSTNet [165], MiMaNet [161], and DIKD [70], as shown in Table 5. However, the performance is still far from satisfying for real-world applications. For the former approach, more effective blocks and structures should be developed to learn discriminative ME features with fewer parameters in the future. For the latter approach, considering the appearance difference between MEs and MaEs, transfer learning methods should be further studied to solve the domain shift problem. Leveraging information from other related tasks, such as age estimation and identity classification, could be considered. In addition, unsupervised learning and semisupervised learning [182] [183] are promising future directions for MER, as they could leverage the massive unlabeled images.

E. ME-AU Detection

ME analysis is a relatively new topic. Currently, most research focused on ME spotting and recognition [186], [187], [188], [189]. The study of facial expression recognition indicates that AU detection is able to facilitate complex facial expression analysis, and developing facial expression recognition with AU analysis simultaneously could boost the facial expression recognition performance [190], [191].

Inspired by the AU contribution to facial behavior analysis, researchers started to study AU detection in MEs. However, compared to MaEs, AU detection becomes more challenging due to the low intensity of MEs and small-scale ME datasets. AU detection is a fine-grained facial analysis that is complicated. Common facial AU datasets contain a large number of facial samples and identity diversity [192], e.g., Aff-Wild2 [193] (564 videos/2 800 000 frames of hundreds of subjects). In contrast, an ME dataset may only contain thousands of images, e.g., CASME containing around 2500 images of 19 subjects. Moreover,

Table 5 Deep Learning-Based MER on SMIC, CASME II, and SAMM

Method	Year	Pre-processing		Input	Network architecture	pre-training	Protocol	SMIC			CASME II			SAMM		
		Mag	TIM					Cate.	F1	ACC	Cate.	F1	ACC	Cate.	F1	ACC
ELRCN [184]	2018	-	✓	OF	4S-CNN+LSTM	✓	LOSO	-	-	-	5	0.5	52.44	-	-	-
OFF-ApexNet [148]	2019	-	-	OF	2S-CNN	-	LOSO	3	0.6709	67.68	3	0.8697	88.28	3	0.5423	68.18
TSCNN [97]	2019	✓	-	OF+Apex	3S-CNN	✓	LOSO	3	0.7236	72.74	5	0.807	80.97	5	0.6942	71.76
LEARNet [78]	2019	-	-	DI	CNN	-	LOSO	3	-	81.60	7	-	76.57	-	-	-
3D-FCNN [85]	2019	-	✓	OF	3S-CNN	-	LOSO	3	-	55.49	5	-	59.11	-	-	-
STSTNet [165]	2019	✓	-	OF	3S-3DCNN	-	LOSO	3	0.6801	70.13	3	0.8382	86.86	3	0.6588	68.10
3D-CNN+LSTM [154]	2019	-	-	Video	3DCNN+LSTM	-	LOSO	-	-	-	5	-	62.5	-	-	-
Graph-TCN [65]	2020	✓	-	Apex	TCN+GCN	-	LOSO	-	-	-	5	0.7246	73.98	5	0.6985	75.00
AU-GACN [83]	2020	-	-	Video	3DCNN+GCN	-	LOSO	-	-	-	3	0.355	71.2	3	0.433	70.2
CBAMNet [137]	2020	✓	✓	Video	3DCNN	✓	10-fold	-	-	-	3	-	69.92	-	-	-
DIKD [70]	2020	-	-	Apex	CNN+KD+SVM	-	LOSO	3	0.71	76.06	4	0.67	72.61	4	0.83	86.74
AffectiveNet [185]	2020	✓	-	DI	4S-CNN	-	LOSO	-	-	-	4	-	68.74	-	-	-
GEME [76]	2021	-	-	DI	2S-CNN+ML	-	LOSO	-	-	-	5	0.7354	75.20	5	0.4538	55.88
LR-GACNN [73]	2021	✓	-	OF+Landmark	2S-GACNN	-	LOSO	-	-	-	5	0.7090	81.30	5	0.8279	88.24
GRAPH-AU [177]	2021	✓	-	Apex	2S-CNN+GCN	-	LOSO	-	-	-	5	0.7047	74.27	5	0.7045	74.26
DSTAN [175]	2021	-	✓	OF+Video	2S-CNN+LSTM+SVM	-	LOSO	-	-	-	5	0.73	75	-	-	-
KFC [172]	2021	-	-	OF	2S-CNN	-	LOSO	3	0.78	77	5	0.7375	72.76	5	0.5709	63.24
AMAN [173]	2022	✓	✓	Video	CNN	✓	LOSO	3	0.77	79.87	5	0.71	75.40	5	0.67	68.85

[1] Mag: Magnification; TIM: Temporal Interpolation Model.
[2] OF: Optical flow; DI: Dynamic image.
[3] nS-CNN: n-stream CNN; ML: Multi-task learning; DA: Domain adaption; KD: Knowledge distillation.
[4] Cate: Category; F1: F1-score; ACC: Accuracy.

the AUs in MEs have imbalanced distribution, e.g., there are 129 AU4, while only 13 AU4 are in CASME II. Existing ME-AU detection research proposed to utilize the MaEs [194] or specific ME characteristics, such as subtle local facial movements [66], [195], [196], which are discussed in more detail in the following, to overcome these issues.

In order to overcome the lack of ME data, Li et al. [194] proposed a dual-view attentive similarity-preserving (DVASP) knowledge distillation to utilize the facial images in the wild to achieve robust ME-AU detection. Considering that one of the key factors for successful knowledge distillation is a generalized teacher network, DVASP utilized a semisupervised dual-view cotraining approach [197], [198] to construct a generalized teacher network by exploiting the massive labeled and unlabeled facial images in the wild. To address the appearance gap between the MEs and MaEs, an attentive similarity-preserving distillation method was proposed to break the domain shift problem by transferring the correlation of important activations instead of directly mimicking the features. In order to overcome the lack of ME data, Li et al. [194] proposed a DVASP to utilize facial images in the wild to achieve robust ME-AU detection. Considering that a generalized teacher network is one of the key factors for successful knowledge distillation, DVASP utilized a semisupervised dual-view cotraining approach [197], [198] to build a generalized teacher network by exploiting the massive unannotated facial images in the wild. To address the appearance gap between the MEs and MaEs, an attentive similarity-preserving distillation method was proposed to break the domain shift problem by transferring the correlation of important activations instead of directly mimicking the features.

Other ME-AU research focuses on modeling subtle AUs [66], [195] based on ME characteristics. An intra-contrastive and intercontrastive learning method was proposed to enlarge and utilize the contrastive information between the onset and apex frames to obtain the discriminative representation for low-intensity ME-AU detection [195]. To effectively learn local facial movement and leverage relationship information between different facial regions to enhance the robustness of ME-AU detection, a spatial and channel attention module was designed to capture subtle ME-AUs by exploring high-order statistics [66]. On the other hand, Zhang et al. [196] proposed a segmentation method based on AUs to extract features on key facial regions and utilized multilabel classification to classify the AUs.

The specific information on the abovementioned AU detection methods is shown in Table. 6. We can see that the work and performance of ME-AU detection are limited. AU detection is a fine-grained detection identifying different facial movements. The low intensity of MEs increases the difficulty of AU detection. Moreover, the ME-AU detection suffers from small-scale and extremely unbalanced datasets as some AUs coexist and the occurrence of some AUs is very low. In the future, more effective AU detection approaches should be explored to better study MEs.

F. ME Generation

As the discussion in Section IV implies, it is challenging to collect MEs compared to ordinary facial expressions. Also, annotating MEs needs certified FACS coders to check videos frame by frame several times, which is time-consuming and labor-intensive. These issues lead to limited samples and imbalanced distribution in ME analysis. Synthesizing MEs is an option to solve these problems. Recently, with the development of generative adversarial network (GAN) [138], [200], [201], image and video generations have been widely applied for data augmentation and image translation [202], [203] and achieved distinctly improved performance in various fields, such as face generation [204], [205] and style transfer [206], [207]. Recently, the ME researchers started to explore

Table 6 AU Detection on CASME II and SAMM

Method	Pre-processing		Input	Method	Protocol	CASME II		SAMM	
	Mag	TIM				Num	F1	Num	F1
LKFS [196]	-	-	Video	LBP-TOP on key regions	5-fold	26	0.535	26	0.513
SCA [66]	-	✓	Video	3D-CNN	4-fold	8	0.668	4	0.505
DVASP [194]	✓	-	Apex	2S-CNN+KD	4-fold	8	0.726	4	0.520
IICL [195]	✓	-	Onset-Apex	3S-CNN	4-fold	8	0.708	4	0.516

[1] Mag: Magnification; TIM: Temporal interpolation model.
[2] nS-CNN: n-stream CNN; DA: Domain adaption; KD: Knowledge distillation.
[3] Num: the number of AUs ; F1: F1-score.

utilizing GAN to generate facial images. However, the MEs are subtle and rapid, and straightforwardly utilizing GANs cannot generate satisfying MEs. The workshop about MEGC 2021 started to include ME generation task [208], leading to increased interest in ME generation. Current ME generation methods mainly leverage AUs or facial key points.

Since facial expressions are constituted by AUs [209], [210], AU-ICGAN [83], FAMGAN [211], and MiE-X [199] introduced GANs based on AUs to generate MEs. Xie et al. [83] proposed the AU Intensity Controllable GAN (AU-ICGAN) to synthesize subtle MEs. Considering that the ME has rapid change and temporal information plays an important role, AU-ICGAN simultaneously evaluated the image quality and video quality to generate nearly indistinguishable ME sequences, which effectively improves the MER performance. Xu et al. designed fine-grained AU modulation (FAMGAN) to eliminate the noise and deal with the asymmetrical AUs. Super-resolution was incorporated into FAMGAN to enhance the quality of the generated ME images. In addition, Liu et al. [199] synthesized a large-scale and trainable ME dataset (MiE-X, which includes 5000 identities and 45 000 samples in total) based on the relationship between AUs and expression categories. The experiments demonstrated that generated MEs could help improve the MER performance. As shown in Fig. 7, the performance of ApexME [62] and Branches [159] pretrained on MiE-X is improved by 3.1% and 3.2% on MMEW and 5.4% and 3.0% on SAMM in terms of accuracy compared to pretrained on ImageNet.

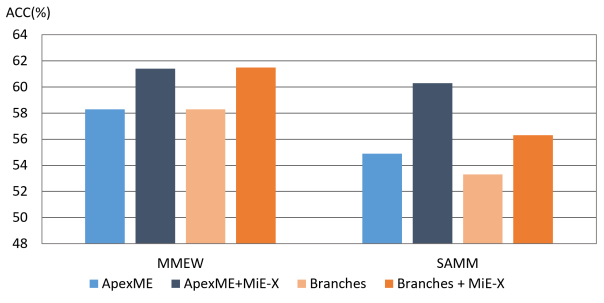


Fig. 7. ***Comparisons with the methods pretrained with ImageNet or MiE-X established by ME generation method [199]. The ApexME and Branches are pretrained on ImageNet, while ApexME + MiE-X and Branches + MiE-X are pretrained on the synchronized dataset MiE-X.***

Other works estimated facial motion using key points to generate ME images [212], [213]. Specifically, Fan et al. [212] developed a deep motion retargeting (DMR) network to capture subtle ME variants by learning key points. Zhang et al. [213] combined the motion model based on key points and local affine transformations with facial prior knowledge and achieved first place in MEGC 2021. Besides, Yu et al. [214] proposed an Identity-aware and Capsule-Enhanced GAN (ICE-GAN) to synthesize MEs with the discriminator detecting the image authenticity and expression categories. Moreover, instead of generating ME images, Liong et al. [215] synthesized optical-flow images of MEs to improve the MER performance based on computed optical flow.

ME generation is a new direction in ME analysis. The subtle and rapid facial movements make ME challenging. Currently, the quality of the generated MEs is not realistic enough. However, with further investigation, it is expected that they could be not only helpful in other ME analyses, such as MER, ME spotting, and ME AU detection, but also in ME synthesis for augmented reality, HCI, and so on.

VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

Knowing how others feel is an important element in social interactions, but it is not always an easy task. Sometimes, people may intentionally express their emotions in the form of, e.g., facial MaEs in order to deliver messages or attitudes, and sometimes, people may hide their true feelings for different reasons. Computational methods developed for ME analysis, including spotting, recognition, and generation, could help in multiple use cases e.g., convert emotion understanding and expert training.

Covert Emotion Understanding: ME analysis tools can aid doctors and therapists in better understanding people's covert emotions for emotion well-beings. Just as how the ME phenomenon was first found, an important application case is a mental assessment and tutoring, especially for young and disordered people. About one in ten young people is affected by mental health problems [216], which can cause wide-ranging effects. They can be long-lasting, and there are well-identified increased physical health problems associated with mental health [217]. Therapists often check recorded videos to examine and review patients'

conditions, which is very time-consuming, and ME analysis methods can be implemented as a tool aiding the process, e.g., to locate and tag suspect moments for review. Such fine-level video analysis can also help doctors find covert emotions in assessments, which are valuable for diagnosis and treatment. Also, the technique can also be applied in other scenarios, such as law enforcement for investigations, online education, and intelligent human–computer interaction, in which MEs should be concerned for fine-level, accurate emotion interpretation.

Expert Training: Besides working as a tool aiding experts in multiple scenarios, ME analysis tools (especially ME generation) can also help in training experts to improve their abilities to read covert emotions. One thing worth noticing is that sample diversity in terms of age, gender, and race should be considered and balanced in designing such tools to insure fair usage of the technology. The ability to read others' emotions is essential for some occupations, such as medical, security, and others. A law enforcement officer or a doctor interacts with a large number of people daily, and they must give judgments and decisions depending on observations within a few minutes. Training tools that can improve their ability to read people's emotions (even the covert ones) will benefit their performance at work. One study illustrated that medical students' communication skills were significantly improved after being trained with the ME training tool METT [218]. However, as mentioned in Section II, METT is composed of "man-made" ME clips that are different from real MEs. Using real spontaneous MEs would have better training effects but is not possible to achieve a large number of real samples on all categories and designated identities, as required for developing a training tool. The technique of ME generation could provide another option. Robust generation models can learn characteristics from real ME samples and generate a large number of MEs on designated model faces, which could provide better and richer training materials.

Even though MEs have been a hot topic with great potential in multiple application scenarios, there are noteworthy challenges from both technological and ethical perspectives, which need to be addressed in future studies.

A. Challenges From the Technological Perspective

1) Small-Scale and Imbalanced Data: Data are the central part of ME research. Although multiple datasets have been collected and released, the scale of most current datasets is still limited, i.e., a few hundred samples. Data annotation is one key issue that hinders the development of large-scale ME datasets as it requires certified expertise and is very time-consuming. Moreover, some emotion, such as fear, is difficult to be evoked, which causes data imbalance issues. Data-driven methods tend to classify test samples to the majority class leading to poor classification performance. Thus, lacking large-scale, well-annotated, and balanced ME data is still a big barrier to ME research. Since it is challenging to induce and label MEs from scratch, leveraging vast online videos with computer–human-cooperated annotations and synthesizing ME samples for imbalanced categories could be possible solutions for the data issues [219].

2) Compound MEs: Many MER studies are built on the assumption of a simplified scenario that each ME appearing at one single time window corresponds to one emotion, e.g., happiness, anger, surprise, sadness, fear, and disgust. However, that is not always the case in practice. Psychological studies [220], [221] show that people can produce mixed expressions in their daily life when two or more "basic" emotions are felt. For instance, surprise can occur with either happiness or fear at the same time and expressed on the face as "happily surprised" or "fearfully surprised." Such mixed expressions are referred to as compound expressions; 17 compound expressions had been identified in [221], suggesting that MaEs and emotions are more complex than previously believed. It is reasonable to assume that there could also be compound MEs, as MEs are initiated the same way as MaEs driven by people's felt emotions. So far, only a few works [222] have concerned compound emotions for ME analysis. Compound MEs could be rare and more challenging to study, but, as they reflect the specific emotional states that practically exist, they should be considered and not ignored in future ME studies. On the other side, one should always be cautious when inferring complex status as compound emotions based on the observation of such a short time window of an ME, as there are open discussions [223] about whether recognizing emotions without context is reliable.

3) Multimodal MER: Psychological research illustrates that there are various ways to express emotions. Visual scenes, voices, bodies, other faces, cultural orientation, and even words shape how emotions are perceived in a face [223]. Leveraging complementary information from multiple modalities can also enhance ME analysis for a better understanding of human's covert emotions. With the rapid development of social media, a large amount of data including texts, videos, and audio is shared online, which could be employed for multimodality research. Moreover, videos recorded from various sensors provide different forms of visual information, such RGB, depth, thermal, and 3-D meshes, which might contribute to the task of ME analysis in different ways. Two lasted ME datasets (i.e., 4DME and CAS(ME)^3) have already considered this in their data building. It would be interesting and valuable to explore integrating multichannels and multimodalities for ME analysis in the future.

4) MEs and MaEs: Most previous facial expression studies explored MaEs and MEs separately, i.e., when exploring the MER task, concerning ME clips and ignoring MaE cases. However, in practice, it is natural and often that MaEs and MEs coexist and even overlap [23] with each other. In the MEGC challenges (MEGC 2019 and MEGC 2020), the organizers posed one track to spot both MEs and MaEs

in long video clips. So far, the challenge data only consider a simpler situation where MaEs and MEs coexist but occur separately. Future studies should dig deeper to explore more challenging situations where MaEs and MEs overlap with each other, which would be a substantial step toward the accurate understanding of human emotion in realistic situations.

5) MEs in Realistic Situations: So far, most existing ME studies are still restricted to analyzing MEs collected in lab environments. They usually concern frontal view facial images without big head movements, illumination variations, or occlusion. However, in realistic situations, it is impossible to avoid these noises. ME analysis methods built on the basis of constrained settings might not generalize well to wild environments. Effective algorithms for analyzing MEs in unconstrained settings with pose changes and illumination variations must be developed in the future.

Moreover, considering the facts that: 1) there is too little training data and 2) numerous "handcrafted" traditional features and "handcrafted" neural architectures have been proposed, more research could be dedicated to specific applications (e.g., in the medical field, HCI, security, and business), in which application-dependent constraints dealing, e.g., with illumination, viewing angles, and types of facial expression, can be used to simplify the problem. Collecting enough training samples and use of multimodal data are also well-motivated and natural.

B. Ethical Issues

1) Privacy and Data Protection: Data are one of the most valuable assets in ME studies. ME data contain facial videos that are sensitive data that must be considered for privacy protection to avoid potential leakage of the participants' personal information. Data protection laws, such as the EU General Data Protection Regulation (GDPR) [224] and the California Consumer Privacy Act (CCPA) [225], have been established to protect the privacy of personal data, referring to international data protection agreements, transfer of participant names, record data, and so on.

In data collection for research purposes, participants are usually gathered in a voluntary way, and they will sign a consent form before any data are collected. The consent form explains issues related to data collection procedure, data processing, and data sharing, and lists all rights and options that participants have. For example, a participant has the right to withdraw his/her own data at any time. Moreover, since people's faces include sensitive biometric information, a consent form should also concern and specify proper usage in various application scenes. Besides defining rules to regulate data usage, another aspect worth attention is privacy-preserving data sharing protocols and techniques, e.g., to remove sensitive information (e.g., the identity), while preserving facial movement properties for ME analysis [226].

2) Fairness and Diversity Among General Population: New technology should consider its fairness and validity among the general population with diverse ages, gender, culture, ethnicity, and so on. For ME studies, this issue needs to be addressed from four aspects.

First, ME datasets should be more diverse. Most existing ME datasets contain samples from young college students of 18–35 years old from Asia and/or Europe due to the sites of research and availability of participant recruitment, while data from older people or Africans or Latinos are lacking.

Second, the fairness and reliability of ME data labeling should be considered. MEs (and AUs) are difficult to label. The current standard for labeling is that two or more professional annotators will work together and cross-check their labels, and the FACS system plays an important role as one annotator should pass the FACS test to get his/her certificate to become a qualified annotator. The FACS test helps improve the reliability among different annotators, but more factors need to be considered. One is the cultural background of the annotators (e.g., Asian or Caucasian) that might impact their judgment. It is hard to tackle as the overall number of certified FACS annotators is very few, and it is already hard for a research group to gather two or more for cross-checking. The other issue might be addressed or improved in the future, i.e., the FACS training materials only contain face images from a few Caucasians but not from other ethnicities. It is not known whether this impacts the labeling of Asian or African faces, but this could be addressed with helps from psychologists and the FACS developers by adding more diverse faces to the training materials.

Third, the fairness of the MiX models should be concerned in terms of outputs on different populations. As current data are biased toward young Caucasian and Asian people, it is not known whether the trained models can generalize well to other population groups, or whether the outputs might be significantly biased on a diverse sample set.

Fourth, ME generation methods could be a helpful tool for improving the diversity of samples. As the generation models can synthesize ME movements with any given face, we can select and generate samples on a balanced face set covering multiple age, gender, and ethnical groups. Such a balanced and diverse set of synthesized ME samples could serve better in applications, e.g., training experts for recognizing covert emotions.

3) Regulated Usage of ME Technology: MEs provide important clues to people's true feelings and, thus, are useful in many potential applications. Meanwhile, there are risks if such technologies [227], [228], [229] are misused for malicious purposes. In both research communities and practical applications, the right-to-privacy and right-to-know should be respected, and consent agreements should be made in any scenario where human participants are involved. People have the right to know that such technology is applied when they are entering a certain area, and they should also have the right to opt-out unless

in law-enforced scenarios. Legislation should be further developed to define specific rules to regulate the use of ME data and technologies.

VII. CONCLUSION

In conclusion, micro-expressions, being involuntary, subtle, and rapid facial expressions, possess the ability to unveil individuals' genuine emotions. The field of computer vision holds significant promise for automatic micro-expression analysis, presenting numerous potential applications and impacting our daily lives. This article offers a comprehensive review of the development of micro-expressions within the realm of computer vision. Instead of solely focusing on the introduction of machine-learning techniques for micro-expression detection and recognition, this overview encompasses the exploration of micro-expression analysis from its roots in psychology and early endeavors in computer vision to the diverse range of contemporary computational methods. The survey not only addresses the current state of the field but also highlights open challenges and outlines future directions, aiming to provide a tutorial-like reference point to anyone with an interest in micro-expressions.

Acknowledgment

This article was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ, USA. ■

REFERENCES

[1] J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions*. London, U.K.: Oxford Univ. Press, 2004.
[2] W. James, "What is emotion?" 2023 Amer. Psychol. Assoc., Washington, DC, USA, Tech. Rep., 1948, pp. 290–303.
[3] P. E. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. London, U.K.: Oxford Univ. Press, 1994.
[4] H. Okon-Singer, T. Hendler, L. Pessoa, and A. J. Shackman, "The neurobiology of emotion–cognition interactions: Fundamental questions and strategies for future research," *Frontiers Hum. Neurosci.*, vol. 9, p. 58, Feb. 2015.
[5] L. Pessoa, "On the relationship between emotion and cognition," *Nature Rev. Neurosci.*, vol. 9, no. 2, pp. 148–158, 2008.
[6] P. E. Sifneos, "The prevalence of 'alexithymic' characteristics in psychosomatic patients," *Psychotherapy Psychosomatics*, vol. 22, nos. 2–6, pp. 255–262, 1973.
[7] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.
[8] J. Tao and T. Tan, "Affective computing: A review," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Cham, Switzerland: Springer, 2005, pp. 981–995.
[9] M. A. Hogg and D. Abrams, "Social cognition and attitudes," Univ. Kent, Canterbury, U.K., Tech. Rep., 2007, pp. 684–721.
[10] A. Mehrabian and M. Wiener, "Decoding of inconsistent communications," *J. Pers. Social Psychol.*, vol. 6, no. 1, p. 109, 1967.
[11] T. T. Amsel, *An Urban Legend Called: 'The 7/38/55 Ratio Rule'*, vol. 13, 2nd ed. Warsaw, Poland: Sciendo, De Gruyter Poland Sp. z o.o., Jun. 2019, pp. 95–99.
[12] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, Jan. 2006.
[13] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.
[14] P. Ekman, "Lie catching and microexpressions," in *The Philosophy of Deception*. Oxford, U.K.: Oxford Univ. Press, 2009, pp. 118–133.
[15] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "An overview of facial micro-expression analysis: Data, methodology and challenge," 2020, *arXiv:2012.11307*.
[16] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, "Micro-expression recognition: An updated review of current trends, challenges and solutions," *Vis. Comput.*, vol. 36, no. 3, pp. 445–468, Mar. 2020.
[17] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: Facial micro-expression recognition," *Multimedia Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, Aug. 2018.
[18] Y.-H. Oh, J. See, A. C. L. Ngo, R. C. -W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods, and challenges," *Frontiers Psychol.*, vol. 9, p. 1128, Jul. 2018.
[19] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2028–2046, Oct. 2022.
[20] X. Ben et al., "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5826–5846, Sep. 2022.
[21] L. Zhou, X. Shao, and Q. Mao, "A survey of micro-expression recognition," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104043.
[22] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*. Cham, Switzerland: Springer, 1966, pp. 154–165.
[23] P. Ekman and W. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.
[24] D. Matsumoto and H. Hwang. (2011). *Reading Facial Expressions of Emotion*. [Online]. Available: https://www.apa.org/science/about/psa/2011/05/facial-expressions
[25] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *J. Nonverbal Behav.*, vol. 6, no. 4, pp. 238–252, 1982.
[26] E. Svetieva and M. G. Frank, "Empathy, emotion dysregulation, and enhanced microexpression recognition ability," *Motivat. Emotion*, vol. 40, no. 2, pp. 309–320, Apr. 2016.
[27] U. Hess and R. E. Kleck, "Differentiating emotion elicited and deliberate emotional facial expressions," *Eur. J. Social Psychol.*, vol. 20, no. 5, pp. 369–385, Sep. 1990.
[28] C. M. Hurley, A. E. Anker, M. G. Frank, D. Matsumoto, and H. C. Hwang, "Background factors predicting accuracy and improvement in micro expression recognition," *Motivat. Emotion*, vol. 38, no. 5, pp. 700–714, Oct. 2014.
[29] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *Proc. Annu. Meeting Int. Commun. Assoc.* New York, NY, USA: Sheraton, 2009, pp. 1–35.
[30] P. Ekman, "Microexpression training tool (METT)," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2002.
[31] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *J. Nonverbal Behav.*, vol. 33, no. 1, pp. 59–69, Mar. 2009.
[32] W. V. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," Consulting, Palo Alto, CA, USA, Tech. Rep. 22, 1978, vol. 3.
[33] X. Li et al., "4DME: A spontaneous 4D micro-expression dataset with multimodalities," *IEEE Trans. Affect. Comput.*, early access, Jun. 14, 2022, doi: 10.1109/TAFFC.2022.3182342.
[34] W. E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions," *Psychol. Bull.*, vol. 95, no. 1, pp. 52–77, 1984.
[35] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivat. Emotion*, vol. 35, no. 2, pp. 181–191, Jun. 2011.
[36] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention*, 2009, p. 16.
[37] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Proc. IEEE Conf. Workshops Autom. Face Gesture Recognit.*, Mar. 2011, pp. 51–56.
[38] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro- and micro-expression spotting in video using strain patterns," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–6.
[39] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.
[40] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, "Micro-expression spotting: A new benchmark," *Neurocomputing*, vol. 443, pp. 356–368, Jul. 2021.
[41] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.
[42] W.-J. Yan et al., "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.
[43] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, Oct. 2018.
[44] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018.
[45] C. H. Yap, C. Kendrick, and M. H. Yap, "SAMM long videos: A spontaneous facial micro- and macro-expressions dataset," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 771–776.
[46] P. Husák, J. Cech, and J. Matas, "Spotting facial micro-expressions 'in the wild,'" in *Proc. 22nd Comput. Vis. Winter Workshop (RETZ)*, 2017, pp. 1–9.
[47] J. Li et al., "CAS(ME)^3: A third generation facial

spontaneous micro-expression database with depth information and high ecological validity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2782–2800, Mar. 2023.
[48] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1449–1456.
[49] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, Oct. 2012.
[50] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, May 2004.
[51] A. A. Salah, N. Sebe, and T. Gevers, "Communication and automatic interpretation of affect from facial expressions," in *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*. Hershey, PA, USA: IGI Global, 2011, pp. 157–183.
[52] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real AdaBoost," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 79–84.
[53] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, nos. 5–6, pp. 555–559, Jun. 2003.
[54] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3560–3565.
[55] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[56] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.
[57] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[58] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.
[59] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[60] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, Aug. 2012.
[61] X. Li et al., "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, Oct. 2018.
[62] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?" in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3094–3098.
[63] Z. Xia, W. Peng, H. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 8590–8605, 2020.
[64] T.-H. Oh et al., "Learning-based video motion magnification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 633–648.
[65] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-TCN with a graph structured representation for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2237–2245.
[66] Y. Li, X. Huang, and G. Zhao, "Micro-expression action unit detection with spatial and channel attention," *Neurocomputing*, vol. 436, pp. 221–231, May 2021.
[67] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proc. CVPR*, Jun. 2011, pp. 137–144.
[68] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1701–1710.
[69] Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 249–263, 2021.
[70] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1037–1043, Apr. 2022.
[71] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process., Image Commun.*, vol. 62, pp. 82–92, Mar. 2018.
[72] S. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 534–537.
[73] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1511–1520.
[74] J. Liu, W. Zheng, and Y. Zong, "SMA-STN: Segmented movement-attending spatiotemporal network for micro-expression recognition," 2020, *arXiv:2010.09342*.
[75] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3034–3042.
[76] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "GEME: Dual-stream multi-task gender-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, Feb. 2021.
[77] T. T. Q. Le, T. Tran, and M. Rege, "Dynamic image for micro-expression recognition on region-based framework," in *Proc. IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Aug. 2020, pp. 75–81.
[78] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet: Dynamic imaging network for micro expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 1618–1627, 2020.
[79] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2799–2813, Dec. 2018.
[80] S.-J. Wang et al., "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, Dec. 2015.
[81] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[82] V. Mayya, R. M. Pai, and M. M. M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 699–703.
[83] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2871–2880.
[84] D. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 382–386.
[85] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, pp. 1331–1339, Nov. 2018.
[86] M. Peng et al., "Recognizing micro-expression in video clip with adaptive key-frame mining," 2020, *arXiv:2009.09179*.
[87] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, pp. 64–67, Dec. 2001.
[88] M. Bai and R. Goecke, "Investigating LSTM for micro-expression recognition," in *Proc. Companion Publication Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 7–11.
[89] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.
[90] B. D. Lucas, "Generalized image matching by the method of differences," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 1986, p. 163.
[91] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[92] T. Senst, V. Eiselein, and T. Sikora, "Robust local optical flow for feature tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1377–1387, Sep. 2012.
[93] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2003, pp. 363–370.
[94] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L^1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*. Cham, Switzerland: Springer, 2009, pp. 23–45.
[95] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[96] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 626–640, Mar. 2020.
[97] B. Song et al., "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.
[98] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis—A practical evaluation study," 2019, *arXiv:1904.11592*.
[99] N. Liu, X. Liu, Z. Zhang, X. Xu, and T. Chen, "Offset or onset frame: A multi-stream convolutional neural network with CapsuleNet module for micro-expression recognition," in *Proc. 5th Int. Conf. Intell. Informat. Biomed. Sci. (ICIIBMS)*, Nov. 2020, pp. 236–240.
[100] B. Sun, S. Cao, J. He, and L. Yu, "Two-stream attention-aware network for spontaneous micro-expression movement spotting," in *Proc. IEEE 10th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Oct. 2019, pp. 702–705.
[101] J. See, M. H. Yap, J. Li, X. Hong, and S. Wang, "MEGC 2019—The second facial micro-expressions grand challenge," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[102] J. Li, S. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "MEGC2020—The third facial micro-expression grand challenge," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 777–780.
[103] D. Patel, G. Zhao, and M. Pietikäinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Cham, Switzerland: Springer, 2015, pp. 369–380.
[104] S.-J. Wang, S. Wu, and X. Fu, "A main directional maximal difference analysis for spotting micro-expressions," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 449–461.
[105] L. Zhang et al., "Spatio-temporal fusion for macro- and micro-expression spotting in long video

sequences," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 734–741.
[106] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Comput. Vis. Image Understand.*, vol. 147, pp. 87–94, Jun. 2016.
[107] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, "SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos," *IEEE Access*, vol. 6, pp. 71143–71151, 2018.
[108] T.-K. Tran, Q.-N. Vo, X. Hong, and G. Zhao, "Dense prediction for micro-expression spotting based on deep sequence model," *Electron. Imag.*, vol. 31, no. 8, pp. 401-1–401-6, Jan. 2019.
[109] H. Pan, L. Xie, and Z. Wang, "Local bilinear convolutional neural network for spotting macro- and micro-expression intervals in long video sequences," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 343–347.
[110] S. Wang, Y. He, J. Li, and X. Fu, "MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos," *IEEE Trans. Image Process.*, vol. 30, pp. 3956–3969, 2021.
[111] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1722–1727.
[112] S.-T. Liong, J. See, K. Wong, A. C. L. Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 665–669.
[113] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, Mar. 2017.
[114] H. Ma, G. An, S. Wu, and F. Yang, "A region histogram of oriented optical flow (RHOOF) feature for apex frame spotting in micro-expression," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 281–286.
[115] D. Borza, R. Danescu, R. Itu, and A. Darabant, "High-speed video system for micro-expression detection and recognition," *Sensors*, vol. 17, no. 12, p. 2913, Dec. 2017.
[116] E. Vahdani and Y. Tian, "Deep learning-based action detection in untrimmed videos: A survey," 2021, *arXiv:2110.00111*.
[117] P. Carcagnì, M. D. Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: A comprehensive study," *SpringerPlus*, vol. 4, no. 1, pp. 1–25, Dec. 2015.
[118] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, Oct. 2016.
[119] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A Delaunay-based temporal coding model for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 698–711.
[120] Y. Wang, J. See, R. C. Phan, and Y. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 525–537.
[121] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition," *PLoS ONE*, vol. 10, no. 5, May 2015, Art. no. e0124674.
[122] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, Jan. 2016.
[123] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[124] Y. Zhang, H. Jiang, X. Li, B. Lu, K. M. Rabie, and A. U. Rehman, "A new framework combining local-region division and feature selection for micro-expressions recognition," *IEEE Access*, vol. 8, pp. 94499–94509, 2020.
[125] B. Allaert, I. M. Bilasco, and C. Djeraba, "Consistent optical flow maps for full and micro facial expression recognition," in *Proc. VISAPP*, vol. 5. Setúbal, Portugal: SciTePress, 2017, pp. 235–242.
[126] S.-T. Liong, R. C.-W. Phan, J. See, Y.-H. Oh, and K. Wong, "Optical strain based recognition of subtle emotions," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2014, pp. 180–184.
[127] S.-T. Liong et al., "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Process., Image Commun.*, vol. 47, pp. 170–182, Sep. 2016.
[128] S. L. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 394–406, Jul. 2019.
[129] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4678–4683.
[130] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, Apr. 2017.
[131] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Jul. 1995.
[132] A. C. L. Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 33–48.
[133] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.
[134] Y. Guo, Y. Tian, X. Gao, and X. Zhang, "Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 3473–3479.
[135] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. Cham, Switzerland: Springer, 2013, pp. 237–280.
[136] K. Li et al., "Three-stream convolutional neural network for micro-expression recognition," *Austral. J. Intell. Inf. Process. Syst.*, vol. 15, no. 3, pp. 41–48, 2019.
[137] B. Chen, Z. Zhang, N. Liu, Y. Tan, X. Liu, and T. Chen, "Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition," *Information*, vol. 11, no. 8, p. 380, Jul. 2020.
[138] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.
[139] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.
[140] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.
[141] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2258–2263.
[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[143] M. Peng, C. Wang, T. Bi, Y. Shi, X. Zhou, and T. Chen, "A novel apex-time network for cross-dataset micro-expression recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 1–6.
[144] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.
[145] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[146] L. Zhou, Q. Mao, and M. Dong, "Objective class-based micro-expression recognition through simultaneous action unit detection and feature aggregation," 2020, *arXiv:2012.13148*.
[147] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.
[148] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019.
[149] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 36–40.
[150] L. Yao, X. Xiao, R. Cao, F. Chen, and T. Chen, "Three stream 3D CNN with SE block for micro-expression recognition," in *Proc. Int. Conf. Comput. Eng. Appl. (ICCEA)*, Mar. 2020, pp. 439–443.
[151] H. Yan and L. Li, "Micro-expression recognition using enriched two stream 3D convolutional network," in *Proc. 4th Int. Conf. Comput. Sci. Appl. Eng.*, Oct. 2020, pp. 1–5.
[152] W. She, Z. Lv, J. Taoi, B. Liu, and M. Niu, "Micro-expression recognition based on multiple aggregation networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 1043–1047.
[153] S. C. Nistor, "Multi-staged training of deep neural networks for micro-expression recognition," in *Proc. IEEE 14th Int. Symp. Appl. Comput. Intell. Informat. (SACI)*, May 2020, pp. 29–34.
[154] R. Zhi, M. Liu, H. Xu, and M. Wan, "Facial micro-expression recognition using enhanced temporal feature-wise model," in *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*. Cham, Switzerland: Springer, 2019, pp. 301–311.
[155] Q. Li, S. Zhan, L. Xu, and C. Wu, "Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow," *Multimedia Tools Appl.*, vol. 78, no. 20, pp. 29307–29322, Oct. 2019.
[156] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," 2021, *arXiv:2101.04838*.
[157] A. Romero, N. Ballas, E. K. Samira, A. Chassang, C. Gatta, and B. Yoshua, "FitNets: Hints for thin deep nets," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.
[158] L. Zhou, Q. Mao, and L. Xue, "Cross-database micro-expression recognition: A style aggregated and attention transfer approach," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 102–107.
[159] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–4.
[160] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: A micro-expression recognition framework," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2936–2944.
[161] B. Xia and S. Wang, "Micro-expression recognition enhanced by macro-expression from spatial–temporal domain," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1186–1193.
[162] Y. S. Gan and S.-T. Liong, "Bi-directional vectors from apex in CNN for micro-expression

recognition," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2018, pp. 168–172.
[163] P. Gupta, "MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network," *IEEE Trans. Affect. Comput.*, early access, Feb. 25, 2021, doi: 10.1109/TAFFC.2021.3061967.
[164] Y. Zhao and J. Xu, "Compound micro-expression recognition system," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Jan. 2020, pp. 728–733.
[165] S. Liong, Y. S. Gan, J. See, H. Khor, and Y. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.
[166] R. Belaiche, Y. Liu, C. Migniot, D. Ginhac, and F. Yang, "Cost-effective CNNs for real-time micro-expression recognition," *Appl. Sci.*, vol. 10, no. 14, p. 4959, Jul. 2020.
[167] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 367–374.
[168] L. Wang, J. Jia, and N. Mao, "Micro-expression recognition based on 2D–3D CNN," in *Proc. 39th Chin. Control Conf. (CCC)*, Jul. 2020, pp. 3152–3157.
[169] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, and J. Yearwood, "LGAttNet: Automatic micro-expression detection using dual-stream local and global attentions," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106566.
[170] M. Bai, "Detection of micro-expression recognition based on spatio-temporal modelling and spatial attention," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 703–707.
[171] M. F. Hashmi et al., "LARNet: Real-time detection of facial micro expression using lossless attention residual network," *Sensors*, vol. 21, no. 4, p. 1098, Feb. 2021.
[172] Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key facial components guided micro-expression recognition based on first & second-order motion," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.
[173] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2420–2424.
[174] V. R. Gajjala, S. P. T. Reddy, S. Mukherjee, and S. R. Dubey, "MERANet: Facial micro-expression recognition using 3D residual attention network," 2020, *arXiv:2012.04581*.
[175] Y. Wang et al., "Micro expression recognition via dual-stream spatiotemporal attention network," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Aug. 2021.
[176] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 79–84.
[177] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1571–1580.
[178] D. M. Kline and V. L. Berardi, "Revisiting squared-error and cross-entropy functions for training neural network classifiers," *Neural Comput. Appl.*, vol. 14, no. 4, pp. 310–318, Dec. 2005.
[179] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.
[180] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.
[181] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.
[182] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.
[183] X. Wang, X. Wang, and Y. Ni, "Unsupervised domain adaptation for facial expression recognition using generative adversarial networks," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–10, Jul. 2018.
[184] H. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 667–674.
[185] M. Verma, S. K. Vipparthi, and G. Singh, "AffectiveNet: Affective-motion feature learning for microexpression recognition," *IEEE Multimedia Mag.*, vol. 28, no. 1, pp. 17–27, Jan. 2021.
[186] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," *Neurocomputing*, vol. 175, pp. 564–578, Nov. 2017.
[187] X. Duan, Q. Dai, X. Wang, Y. Wang, and Z. Hua, "Recognizing spontaneous micro-expression from eye region," *Neurocomputing*, vol. 217, pp. 27–36, Dec. 2016.
[188] S.-J. Wang, W.-J. Yan, T. Sun, G. Zhao, and X. Fu, "Sparse tensor canonical correlation analysis for micro-expression recognition," *Neurocomputing*, vol. 214, pp. 218–232, Nov. 2016.
[189] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, Jul. 2014.
[190] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1841–1850.
[191] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.
[192] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.
[193] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–6.
[194] Y. Li, W. Peng, and G. Zhao, "Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.
[195] Y. Li and G. Zhao, "Intra- and inter-contrastive learning for micro-expression action unit detection," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 702–706.
[196] L. Zhang, O. Arandjelovic, and X. Hong, "Facial action unit detection with local key facial sub-region based multi-label classification for micro-expression analysis," in *Proc. 1st Workshop Facial Micro-Expression, Adv. Techn. Facial Expressions Gener. Spotting*, Oct. 2021, pp. 11–18.
[197] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–152.
[198] X. Niu, H. Han, S. Shan, and X. Chen, "Multi-label co-regularization for semi-supervised facial action unit recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 909–919.
[199] Y. Liu, Z. Wang, T. Gedeon, and L. Zheng, "Action units that constitute trainable micro-expressions (and a large-scale synthetic dataset)," 2021, *arXiv:2112.01730*.
[200] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2970–2979.
[201] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4432–4441.
[202] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[203] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.
[204] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," in *Proc. Class Project Stanford CS231N, Convolutional Neural Netw. Vis. Recognit., Winter Semester*, no. 5, 2014, p. 2.
[205] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3313–3332, Apr. 2023.
[206] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 11, pp. 3365–3385, Nov. 2020.
[207] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.
[208] J. Li et al., "FME'21: 1st workshop on facial micro-expression: Advanced techniques for facial expressions generation and spotting," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5700–5701.
[209] E. L. Rosenberg and P. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 2020.
[210] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 818–833.
[211] Y. Xu, S. Zhao, H. Tang, X. Mao, T. Xu, and E. Chen, "FAMGAN: Fine-grained AUs modulation based generative adversarial network for micro-expression generation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4813–4817.
[212] X. Fan, A. R. Shahid, and H. Yan, "Facial micro-expression generation based on deep motion retargeting and transfer learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4735–4739.
[213] Y. Zhang, Y. Zhao, Y. Wen, Z. Tang, X. Xu, and M. Liu, "Facial prior based first order motion model for micro-expression generation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4755–4759.
[214] J. Yu, C. Zhang, Y. Song, and W. Cai, "ICE-GAN: Identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis," 2020, *arXiv:2005.04370*.
[215] S.-T. Liong et al., "Evaluation of the spatio-temporal features and GAN for micro-expression recognition system," *J. Signal Process. Syst.*, vol. 92, pp. 705–725, Mar. 2020.
[216] M. Murphy and P. Fonagy, "Mental health problems in children and young people," *Our Children Deserve Better, Prevention Pays: Annual Report of the Chief Medical Officer*. Department of Health, ch. 10, 2013.
[217] *Parental Mental Illness: The Problems for Children. Information for Parents, Carers and Anyone Who Works With Young People*, Roy. College Psychiatrists, London, U.K., 2012.
[218] J. Endres and A. Laidlaw, "Micro-expression recognition training in medical students: A pilot

study," *BMC Med. Educ.*, vol. 9, no. 1, pp. 1–6, Dec. 2009.
[219] G. Zhao and X. Li, "Automatic micro-expression analysis: Open challenges," *Frontiers Psychol.*, vol. 10, p. 1833, Aug. 2019.
[220] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. E1454–E1462, Apr. 2014.
[221] S. Du and A. M. Martinez, "Compound facial expressions of emotion: From basic research to clinical applications," *Dialogues Clin. Neurosci.*, vol. 17, no. 4, pp. 443–455, Dec. 2015.
[222] Y. Zhao and J. Xu, "A convolutional neural network for compound micro-expression recognition," *Sensors*, vol. 19, no. 24, p. 5553, Dec. 2019.
[223] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions Psychol. Sci.*, vol. 20, no. 5, pp. 286–290, 2011.
[224] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Cham, Switzerland: Springer, vol. 10. 2017.
[225] L. de la Torre, "A guide to the California consumer privacy act of 2018," SSRN, Elsevier, Rochester, NY, USA, Tech. Rep., 2018, pp. 1–17.
[226] H. Proença, "The UU-Net: Reversible face de-identification for visual surveillance video footage," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 496–509, Feb. 2022.
[227] M. Kosinski, "Facial recognition technology can expose political orientation from naturalistic facial images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–7, Jan. 2021.
[228] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 145–151.
[229] S. Oviatt, "Technology as infrastructure for dehumanization: Three hundred million people with the same face," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 278–287.

ABOUT THE AUTHORS

Guoying Zhao (Fellow, IEEE) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005.

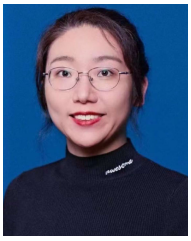
She is currently an Academy Professor and a Full Professor (tenured in 2017) with the University of Oulu, Oulu, Finland. She is also a Visiting Professor with Aalto University, Espoo, Finland. She has authored or coauthored more than 300 papers in journals and conferences with more than 23 330 citations in Google Scholar and an H-index of 72 (May 2023). Her current research interests include image and video descriptors, facial expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics.

Dr. Zhao is a member of Academia Europaea, a member of the Finnish Academy of Sciences and Letters and a Fellow of the International Association for Pattern Recognition (IAPR) and Asia-Pacific Artificial Intelligence Association (AAIA). She was the Panel Chair of International Conference on Automatic Face and Gesture Recognition (FG) 2023, the Co-Program Chair of the ACM International Conference on Multimodal Interaction (ICMI 2021), and the Publicity Chair of Scandinavian Conference on Image Analysis (SCIA) 2023 and FG 2018. She has served as the area chair of several conferences and was/is an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA, *Pattern Recognition*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Image and Vision Computing* journals. Her research has been reported by Finnish TV programs, newspapers, and *MIT Technology Review*.

Xiaobai Li (Senior Member, IEEE) received the B.Sc. degree in psychology from Peking University, Beijing, China, in 2004, the M.Sc. degree in biophysics from the Chinese Academy of Sciences, Beijing, in 2007, and the Ph.D. degree in computer science from the University of Oulu, Oulu, Finland, in 2017.

She is currently an Assistant Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. Her research interests include facial expression recognition, micro-expression analysis, remote physiological signal measurement from facial videos, and related applications in affective computing and healthcare.

Dr. Li was the Co-Chair of several international workshops in Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), International Conference on Automatic Face and Gesture Recognition (FG), and ACM Multimedia. She is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Frontiers in Psychology*, and *Image and Vision Computing*.



Yante Li (Member, IEEE) received the Ph.D. degree in computer science from the University of Oulu, Oulu, Finland, in 2022.

She currently holds a postdoctoral position at the Center for Machine Vision and Signal Analysis, University of Oulu. Her current research interests include affective computing, micro-expression analysis, and facial action unit detection.

Matti Pietikäinen (Life Fellow, IEEE) received the Doctor of Science in Technology degree from the University of Oulu, Oulu, Finland, in 1982.

From 1980 to 1981 and 1984 to 1985, he visited the Computer Vision Laboratory, University of Maryland, Baltimore, MD, USA. He is currently an Emeritus Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. He has made fundamental contributions, e.g., to the local binary pattern (LBP) methodology, texture-based image and video analysis, and facial image analysis. He has authored over 350 refereed papers in international journals, books, and conferences. His papers have over 83 500 citations in Google Scholar (H-index: 100) (April 2023).

Dr. Pietikäinen served as a member of the Governing Board of the International Association for Pattern Recognition (IAPR) from 1989 to 2007. He became one of the founding fellows of the IAPR in 1994. He is an IEEE Fellow for his contributions to texture and facial image analysis for machine vision. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for Fundamental Contributions in Computer Vision. He was a recipient of the prestigious IAPR King-Sun Fu Prize 2018 for fundamental contributions to texture analysis and facial image analysis. He received the Computer Science Leader Award from Research.com. He ranked 351 in the world and 1 in Finland. He was an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *Pattern Recognition*, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, and *Image and Vision Computing* journals. He serves as a Guest Editor for special issues of IEEE TPAMI and PROCEEDINGS OF THE IEEE. He was the President of the Pattern Recognition Society of Finland from 1989 to 1992 and was named its Honorary Member in 2014.