4DME: A Spontaneous 4D Micro-Expression Dataset With Multimodalities

Xiaobai Li, *Member, IEEE*, Shiyang Cheng, *Member, IEEE*, Yante Li, *Student Member, IEEE*, Muzammil Behzad, *Student Member, IEEE*, Jie Shen, *Member, IEEE*, Stefanos Zafeiriou, *Member, IEEE*, Maja Pantic, *Fellow, IEEE*, and Guoying Zhao, *Fellow, IEEE*

Abstract—Micro-expressions (ME) are a special form of facial expressions which may occur when people try to hide their true feelings for some reasons. MEs are important clues to reveal people's true feelings, but are difficult or impossible to be captured by ordinary persons with naked-eyes as they are very short and subtle. It is expected that robust computer vision methods can be developed to automatically analyze MEs which requires lots of ME data. The current ME datasets are insufficient, and mostly contain only one single form of 2D color videos. Researches on 4D data of ordinary facial expressions have prospered, but so far no 4D data is available in ME study. In the current study, we introduce the 4DME dataset: a new spontaneous ME dataset which includes 4D data along with three other video modalities. Both micro- and macro-expression clips are labeled out in 4DME, and 22 AU labels and five categories of emotion labels are annotated. Experiments are carried out using three 2D-based methods and one 4D-based method to provide baseline results. The results indicate that the 4D data can potentially benefit ME recognition. The 4DME dataset could be used for developing 4D-based approaches, or exploring fusion of multiple video sources (e.g., texture and depth) for the task of ME analysis in future. Besides, we also emphasize the importance of forming a clear and unified criteria of ME annotation for future ME data collection studies. Several key questions related with ME annotation are listed and discussed in depth, especially about the relationship between AUs and ME emotion categories. A preliminary AU-Emo mapping table is proposed with justified explanations and supportive experimental results. Several unsolved issues are also summarized for future work.

Index Terms—4D, action unit, dataset, emotion, facial expression, micro-expression, multimodality

1 Introduction

Facial expression is one major form that people convey and perceive emotions. Facial expression recognition has been a popular research topic in computer vision for over twenty years ever since Picard proposed the concept of affective computing in her book [1]. However, not all emotions are shown on the face in all occasions, and ordinary facial expressions only allow us to understand emotions on a coarse and superficial level. Under certain circumstance, people may intentionally hide their true emotion for some purpose, e.g., to avoid bad consequence or to deceit. There is one special form of facial expression, i.e., the micro-expression (ME), which may occur when people try to suppress their natural facial expressions but fail, and some are leaked out and briefly shown in the form of ME. The study of ME originated from psychology from 1960's [2] and got attention of computer vision field only from about ten years ago. Compared to ordinary facial expressions (a.k.a macro-expression), MEs are much shorter, i.e., 1/25 to 1/2 second (the precise length definition varies [3], [4], but 'no longer than 1/2 second' is commonly agreed), and the intensities of the movements are very subtle [5].

The main motivation for automatic ME analysis is that, as a fleeting subtle motion MEs are difficult for ordinary people to perceive with naked eyes [6], and it is expected that computer algorithms could help capturing and recognizing MEs and allow machines to interpret human emotions at a finer level. There are similarities between ordinary facial expression recognition and ME recognition, but the task of ME analysis is facing some special challenges: 1) lack of data, as inducing and labeling are difficult and time consuming, which both require expertise; 2) controversial data categorization; 3) brief movements with extremely low intensity, which makes the recognition task difficult. We were one of the earliest groups to work on these ME challenges, and collected the first spontaneous ME dataset in 2011, i.e., the SMIC [7]. Several other ME databases were built and shared in the following decade, including, CASME [8], CASME II [9], CASME^2 [10], SAMM [11], MEVIEW [12], and MMEW [13], which are the pillars for the progress of this research topic so far.

However, there are several constraints for the current ME databases. 1). The size of each dataset is comparatively small, e.g., of a few hundreds of ME samples. This is mainly due to

- *Xiaobai Li, Muzammil Behzad, and Guoying Zhao are with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90570 Oulu, Finland. E-mail: {xiaobai.li, muzammil.behzad, guoying.zhao}@oulu.fi.*
- *Shiyang Cheng, Yante Li, Jie Shen, Stefanos Zafeiriou, and Maja Pantic are with the Intelligent Behaviour Understanding Group (ibug), Imperial College London, SW7 2BX London, U.K. E-mail: dr.shiyang.cheng@gmail.com, yante.li@oulu.fi, {jie.shen07, s.zafeiriou, m.pantic}@imperial.ac.uk.*

Manuscript received 7 Feb. 2022; revised 1 June 2022; accepted 3 June 2022. Date of publication 0 . 2022; date of current version 0 . 2022.
This work was supported in part by the Academy of Finland (Postdoc Project 6 +E under Grant 323287, Academy Professor project EmotionAI under Grants 336116, 345122) and in part by Infotech Oulu.
This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Ethics Committee of Human Sciences of University of Oulu.
(Corresponding author: Guoying Zhao.)
Recommended for acceptance by M. Daoudi.
Digital Object Identifier no. 10.1109/TAFFC.2022.3182342

the enormous difficulties in inducing and collecting the MEs samples, as well as the tremendous efforts required for annotation. Arguably, more data are needed to develop advanced automatic ME analysis methods, especially when employing modern machine learning models that are usually data-hungry. This leads to the second problem, that is, 2) the criteria of ME labeling is ambiguous and inconsistent between different datasets which makes it difficult for data merging. Although researchers rely on the Facial Action Coding System (FACS) to label action units (AU), but there are no clear rules about mapping of AUs (or AU combinations) to ME emotion categories; 3). Current ME datasets lack data variety, i.e., most datasets only contain one form of videos, which is 2D, high-speed color video (except the SMIC which also contains near infrared videos) of frontal faces. The monotonous data format has limited the applications of existing MEs methods, which merely function in perfectly frontal faces and completely fail on near-frontal or profile faces. On the other hand, due to a lack of 4D data, fundamental research on the 3D dynamics of MEs cannot be undertaken.

Different from the situation in 4D MEs, there are several large scale multimodal 4D datasets for ordinary facial expressions Compared with traditional 2D videos, dynamic 3D videos (referred to as '4D' thereafter) could provide richer information to facilitate computer vision based analysis. With the fast development of 3D imaging technology in recent years, it is now possible to record and reconstruct high fidelity 3D facial videos with high frame rate. The number of ordinary facial expression datasets containing 4D data is increasing, such as the BP4D-Spontaneous [17], BP4D+ [18], and 4DFAB [19] datasets. Accordingly, several methods have been proposed which utilize 4D inputs or features for facial expression recognition, such as Dynamic Geometrical Image Network (DGIN) [20], Collaborative Cross-domain Dynamic Image Network (CCDN) [21], and Multi-View Transformer (MiT) [22].

Generally speaking, 4D data can facilitate the task of facial expression recognition in the following aspects. First, 4D facial data can be rendered back to the image space in arbitrary views, which can help alleviate the self-occlusion problem (e.g., facial movements are at the invisible side of the face) in traditional 2D videos. Second, 4D data also allows combining color and texture information with depth or 3D shape information, which are very helpful to deal with problems caused by head motions and lighting changes.

Compared to ordinary FE recognition, ME recognition is a more challenging task as MEs are more subtle and often only involve unilateral movements, e.g., slightly raised outer eyebrow on one side of the face, which might be completely invisible in 2D video but visible in 4D sequence (as we have an ear-to-ear reconstruction of the face). 4D data can provide possibilities to explore 4D-based approaches for achieving more robust performance for ME analysis. Meanwhile, it is also noticed that 4D data has its own special challenges, e.g., the artefacts introduced during the reconstruction process, which might hinder the analysis of such subtle movements of MEs. Therefore we cannot simply assume existing 4D-based FE recognition methods will also work well for ME recognition, and special methods need to be developed and tested on 4D data of MEs.

In this paper, we advance the research of automatic 4D MEs analysis with several contributions:

1) A new 4DME dataset is introduced, which contains multimodalities of traditional 2D grayscale videos, RGB videos, depth videos, and 4D videos of spontaneous MEs. Five categories of emotion labels and 22 AU labels are provided.
2) Several issues of ME labeling are discussed, especially about the relationship of action unit (AU) labels and ME emotion categories. A preliminary 'AU-Emo' correspondence table is summarized and proposed for ME labeling with supporting evidences.
3) We evaluate multiple methods for ME recognition on 4DME dataset, including both 2D-based methods and 4D-based methods. Preliminary findings indicate advantage of 4D-based approach. The results could work as baselines for future comparisons.

We believe this study can contribute to the ME study community not only by providing a new ME dataset, but also by initialising and promoting discussions to clarify ME categorization and its relationship with AUs, so that there will be clearer rules to follow for future ME data labeling and fusion.

2 Related Work

2.1 ME Datasets

The progress of research methods in one field is largely dependent on available datasets. Unlike the field of ordinary facial expression research in which many large scale datasets of various forms are available (e.g., CK+ [23], and BP4D [17]), current publicly released ME datasets are still limited.

One major challenge is that MEs are difficult to induce, and the earliest ME datasets are posed ones including Polikovsky's database [24] and USF-HD [25], which were collected by asking participants to act or mimic fast facial expressions. The posed datasets were helpful at the earliest stage when the topic of ME recognition was newly introduced to the computer vision field and there was no data available. However, posed expressions cannot represent the actual characteristics of spontaneous expressions occurred involuntarily in natural scenes as they differ on both spatial and temporal dimensions.

Later studies all focused on spontaneous MEs and several spontaneous ME datasets were built so far, including SMIC [14], CASME [8], CASME II [9], CAS(ME)^2 [10], SAMM [11], MEVIEW [12], and MMEW [13]. One popular approach for inducing spontaneous MEs is by asking participants to watch emotional movie clips and hide their feelings by keeping a neutral face, which was adopted by most of these datasets, except the MEVIEW dataset which contains videos of poker players on TV shows. Details of these spontaneous ME datasets are summarized in Table 1. It can be seen that most ME datasets contain 100 to 300 samples which are much smaller than the scale of ordinary FE datasets. One dataset is not sufficient especially for training deep neural networks. In the ME Grand Challenge (MEGC2019) [15] the organizers proposed the idea of composite dataset, which was essentially merging CASME II, SAMM and SMIC to

TABLE 1
Spontaneous Micro-Expression Datasets

Database	Subjects		ME video clips				Annotations	
	Number	MultiEth	Number	Resolution	Frame rate	Modality	Emotion category	AU category
SMIC [14]	16	Y	164	640 × 480	100	HS (RGB)	*Pos* (51) *Neg* (70) *Sur* (43)	None
	8		71	640 × 480	25	NIR	*Pos* (28) *Neg* (23) *Sur* (20)	
	8		71	640 × 480	25	VIS (RGB)	*Pos* (28) *Neg* (24) *Sur* (19)	
CASME [8]	19	N	195	640 × 480	60	RGB	*Hap* (5) *Dis* (88) *Sad* (6) *Con* (3) *Fea* (2) *Ten* (28) *Sur* (20) *Rep* (40)	12+
				1280 × 720	60			
CASME II [9]	26	N	247	640 × 480	200	RGB	*Hap* (33) *Sur* (25) *Dis* (60) *Rep* (27) *Oth* (102)	11+
CAS(ME)^2 [10]	22	N	57	640 × 480	30	RGB	*Pos* (8) *Neg* (21) *Sur* (9) *Oth* (19)	28
SAMM [11]	32	N	159	2040 × 1088	200	Grayscale	*Hap* (24) *Ang* (20) *Sur* (13) *Dis* (8) *Fea* (7) *Sad* (3) *Oth* (84)	ALL AUs
MEVIEW [12]	16	N	29	720 × 1280	30	RGB	*Hap* (4) *Ang* (1) *Sur* (9) *Dis* (1) *Fea* (2) *Unc* (7) Con(5)	7
MMEW [13]	36	N	300	1920 × 1080	90	RGB	*Hap* (36) *Ang* (8) *Sur* (89) *Dis* (72) *Fea* (16) *Sad* (13) *Oth* (66)	17
Composite ME [15]	68	Y	442	640 × 480	100	RGB	*Pos* (109), *Neg* (250), and *Sur* (83)	27
				640 × 480	200			
				2040 × 1088	200			
Compound ME [16]	90	Y	1050	640 × 480	100	RGB	*Neg* (233) *Pos* (82) *Sur* (70) *PS* (74) *N S* (236) *PN* (197) *NN* (158)	27
				640 × 480	200			
				2040 × 1088	200			
4DME	56	Y	1068	1200 × 1600	60	**4D**	*Neg* (127) *Pos* (34) *Sur* (30) *Rep* (6) *PS* (13) *NS* (8) *RS* (3) *PR* (8) *NR* (7) *Oth* (31)	22
				640 × 480	60	Grayscale		
				640 × 480	30	RGB		
				640 × 480	30	Depth		

[1] *Modality: RGB indicates 2D color videos; NIR indicates 2D near infra-red videos; HS indicates 2D high-speed videos.*
[2] *MultiEth: whether subjects are of multiple ethnicities.*
[3] Pos: *Positive;* Neg: *Negative;* Sur: *Surprise;* Hap: *Happiness;* Dis: *Disgust;* Rep: *Repression;* Ang: *Anger;* Fea: *Fear;* Sad: *Sadness;* Con: *Contempt;* Unc: *Unclear;* Oth: *Others;* PS: *Positively surprise;* NS *Negatively surprise;* PN: *Positively negative;* NN: *Negatively negative;*
[4] *There are inconsistent statistics in literatures. Numbers here are according to the original papers or the downloaded data files.*

generate a larger dataset for model training and evaluation. Nevertheless, the fusion of different datasets is quite difficult and ineffective, as the induced spontaneous expressions can be quite complex and the labeling criteria is often inconsistent between datasets. Zhao and Xu [16] proposed to adopt the concept of compound facial expression [26] for ME recognition, which allows and emphasizes the co-existence of multiple emotion categories of each ME, e.g., happily surprised, or fearfully surprised, and a compound ME dataset CMED was introduced which summarized five original ME datasets of SMIC, CASME, CASEME II, CAS(ME)^2, and SAMM.

2.2 2D ME Recognition Methods

Various methods have been proposed so far including both traditional feature descriptors which were explored in earlier stage, and deep neural network approaches that thrive in recent years. Traditional approaches [27], [28], [29] usually involve one or more feature descriptors plus one classifier for the ME recognition task. The most popular descriptors are spatio-temporal features, include LBP [30], HOG [31] and optic flow [32]. Several special processes were also explored and added to the approach to counter for the challenges of MEs, e.g., a temporal interpolation or normalization process [33] was used to deal with short and unequal duration of the MEs, and a video motion magnification approach [34] was introduced to magnify the subtle movements of MEs to boost the recognition performance.

As the fast progress of deep learning methods, researchers started to explore deep network-based approaches since 2016 [35], [36] for the ME recognition task. Inspired by works in ordinary FE recognition studies, most studies explored CNN or RNN based approaches. Nonetheless, due to the scarcity of ME data, early deep-based models [28], [37] struggled to compete with traditional approaches. As more efforts were made in the following years to gather more data and to specifically tailor the networks for MEs domain, several promising solutions started to emerge. First, there are large-scale ordinary facial expression datasets, and the knowledge learned from those datasets can be leverage to improve ME recognition performance by transfer learning. Currently, several transfer learning methods have been applied for robust ME recognition, including fine-tuning [38], [39], knowledge distillation [40], [41], and domain adaptation [42], [43], [44]. Second, since MEs may only involve local regional motions [45], it is crucial to selectively highlight the corresponding regions of interest (ROI) [46], [47]. Attention modules were employed in several studies [48], [49], [50], [51], [52] in various forms, which were demonstrated to be an effective solution for selecting ROIs and enhance the ME representation. Furthermore, an ME may contain multiple facial movements (AUs), and the latent semantic information among these local movements could be helpful to improve ME recognition performance. The graph convolution network (GCN) can model these semantic relationships which were explored in several studies [53], [54], [55] for ME analysis.

Although multiple approaches have been explored, they all concentrated on one source of data, i.e., 2D videos recorded with RGB cameras. The inputs could be in multiple forms, e.g., some [39], [56], [57] used static images (e.g., the apex frame), some [54], [58] used images sequences, and some others used extracted features, such as optic flow features [38], [59], facial landmarks [60], and dynamic images [61], [62], [63], but the source data are the same. The main reason is that current ME data lacks variability and only 2D RGB video data is available. 2D videos can provide clues in 2D spatial domain (mostly in frontal view) but is constrained if the motion occurs at an occluded region caused by e.g., head orientations. This problem cannot be solved by method-wise solutions but only data-wise solutions, i.e., facial videos with depth or 4D information.

2.3 4D Ordinary Facial Expression Datasets

Over the past decade, several large-size 4D facial expression datasets were released. The 4D facial point clouds allow the exploration of methods especially for fetching facial deformation patterns from dynamic 3D spatial domain for emotion

recognition. Earlier studies started from posed 4D facial expression datasets, such as BU-4DFE [64], D3DFACS [65], and Hi4D-ADSIP [66], as posed facial expressions data are comparatively easier to gather and annotate. The BU-4DFE [64] dataset contains 606 samples of posed facial expressions of six emotion categories recorded from 58 females and 43 males (18∼45 years). The videos have a frame rate of 25 frames per second (FPS), and each clip lasts for approximately $3 \sim 4$ seconds. The D3DFACS [65] is another widely used 4D dataset of posed facial expressions, which contains 519 AU sequences from ten subjects ($23 \sim 41$ years). D3DFACS was annotated with up to 38 categories of AU labels. The Hi4D-ADSIP [66] is a comprehensive 3D dynamic facial articulation database which contains 3360 facial scan sequences captured from 80 subjects of various age, gender and ethnicity. The data contains six posed facial expressions, pain, and phrase reading scenarios to facilitate both emotion recognition and diagnosis of facial dysfunctions. The above-mentioned datasets focus only on posed expressions, thereby restricting the applicability of recognition systems towards real-world applications.

Later studies also made efforts to collect spontaneous 4D facial expression datasets, and the most widely used ones include B3D(AC) [67], BP4D-Spontaneous [17], BP4D+ [18], and 4DFAB [19]. B3D(AC) [67] is the first 4D audio-visual database with spontaneous expressions and speech. The dataset contains 1109 sequences (4.67 seconds long on average) recorded from 14 subjects of 15 rated affective adjectives. The BP4D-Spontaneous [17] dataset has 328 samples from 41 subjects collected in several well-designed tasks, such as physical activities and interviews, which can evoke spontaneous expressions. The BP4D+ [18] is an extension of BP4D-Spontaneous, which incorporated different modalities such as physiological signals and thermal imaging, and 140 more subjects are included. An important characteristic of the BP4D-Spontaneous and BP4D+ datasets is that both provide AU labels which are extremely beneficial for emotion analysis. More recently, a larger size dataset, the 4DFAB [19] was released, which contains over 1.8 millions of 3D meshes (about 30 000 seconds of recordings) from 180 subjects aged from a wide range of $5 \sim 75$ years. It includes 4D data of both spontaneous and posed facial expression clips with a frame rate of 60 FPS.

2.4 Establishing Correspondence for 4D Data

Different from the static 3D data, 4D data require extra processing steps to establish correspondence between frames within a sequence. Although these steps are usually very time consuming, they are critical to the success of 4D facial expression method, because a good correspondence can help preserve the facial dynamics.

There are several approaches for this purpose, the most straight-forward approach is to directly align an universal template to every mesh in the target sequence (e.g., using Non-rigid Iterative Closest Points [68] or Active Non-rigid Iterative Closest Points [69]). In order to improve the correspondence between meshes, this step is often performed under the guidance of sparse facial landmarks. However, this approach is not computational efficient and often fails to provide temporally consistent correspondence (please refer to [70] for an in-depth explanation). Comparing with the direct 3D registration approach, non-rigid image registration in UV-space [65][71] are more favorable. This approach first *unwraps* the 3D mesh into a 2D intermediary (namely UV-space) using techniques such as cylindrical projection [72] or conformal mapping [73]. Essentially, the UV space encodes a bijective mapping from 2D positions to the corresponding 3D point in the mesh, since the mapping can faithfully represent a 3D face, establishing dense correspondence between any two UV images will automatically return us a dense 3D-to-3D correspondence for their corresponding 3D meshes. This is beneficial because it transfers the challenging 3D registration problem to the well-solved 2D non-rigid image alignment problem. The third approach to handle 4D data is comparatively simple as well as efficient, and was used in quite some deep 4D expression recognition methods [20], [21], [74]. In this approach, the 3D faces are first rigidly aligned to a common reference frame using 3D facial landmarks, so as to remove the scaling, rotation and translation effects. Next, the aligned face will be projected to 2D in single/multiple views, which can generate RGB texture or depth images for later tasks. We follow this approach for its simplicity, even though this approach cannot provide us a dense correspondence among the 4D data, the projected views are sufficient for our tasks.

2.5 4D Methods for Facial Expression Recognition

Along with the release of 4D facial expression datasets, many studies have explored 4D-based methods for recognizing ordinary facial expressions. We loosely group those methods into traditional approaches and deep learning approaches, and review them separately in the following sections. Compared with methods for static 3D data, 4D-based methods usually require an extra feature embedding (or extraction) step for the input data. For example, Cheng *et al.* [19] used 3DMM parameters instead of the 3D mesh to train their expression recognition model. In [75], 3D faces were projected into Riemannian manifold to get the radial curves for expression recognition. On the other hand, in order to capture expression dynamics, a temporal/spatial-temporal model (e.g., using LSTM [19], Res3D [53], or hidden Markov Model [74]) is also employed by 4D methods.

2.5.1 Traditional Approaches

Sun *et al.* [74] introduced a method to obtain correspondences among the dynamic sequences of 3D facial point clouds. Based on the proposed correspondences, they coined the idea of using spatiotemporal hidden Markov model (ST-HMM) for capturing the facial deformations by assessing both inter-frame and intra-frame variations. In a similar way, Yin *et al.* [64] exploited a 2D Hidden Markov Model to analyze the facial muscle movements over time for improvements in expression classification. Another study [75] explored Riemannian analysis for 4D facial expression recognition. The 3D facial meshes were mainly represented by collections of radial curves. For effectively quantifying the facial patterns of the facial expressions, a Riemannian shape analysis was applied. The authors proposed a deformation vector field and used a random forest classifier for learning the temporal dynamics of the face deformations. Sandbach *et al.* [76] proposed a method to

represent the crucial information between neighboring 3D frames as motion-based features, which was referred as Free-Form Deformation (FFD). Features were extracted from the onset and offset frames of the given expression, and then fed to GentleBoost classifiers to estimate the complete temporal dynamics of 4D expressions.

2.5.2 Deep Learning Approaches

In recent years, several deep learning based-approaches were also proposed for 4D facial expression recognition. Li *et al.* [20] proposed a dynamic geometrical image network. Geometrical images were generated by estimating the differential quantities from the given 3D facial meshes. A score-level fusion was then performed on the probability scores of different geometrical images for facial expression recognition. Behzad *et al.* [21] proposed a Collaborative Cross-domain Dynamic Image Network (CCDN) to generate cross-domain dynamic images for encoding the temporal dynamics in a single image. 3D facial meshes were projected to 2D images of multiple views, and features from various domains (e.g., texture and depth) are combined in the network to collaboratively work for 4D facial expression recognition task. In a recent study [74], an advanced method was introduced which highlights the effectiveness of sparsity-aware features. On the bases of the CCDN framework, the authors combined 3D landmarks as sparse features for capturing effective facial patterns which achieved significant performance improvement. The improved approach is not only effective for 4D emotion recognition, but also computational-efficient.

3 4DME Database Profile

We collect a 4D spontaneous ME database, i.e., the 4DME[1], which contains multimodal facial videos recorded with different cameras and both AU labels and emotion category labels are provided. The main motivation for using multiple cameras is to provide various forms of data. The 4DME dataset would be valuable to explore 1) whether 4D data can boost the ME recognition performance, and 2) whether the fusion of various data sources (e.g., RGB and depth) could facilitate the task of ME recognition. Details of the data collection and annotation are explained in below.

3.1 Equipment Setup

The data recording was held in a lab studio, and the setup is shown in Fig. 1a and 1b. The participant sits on a seat in front of a. The 4DME database contains multi-modality video data and three sets of cameras were used. First, a professional 4D imaging system, i.e., the Dimension Imaging 4D (DI4D) capturing system which contains six high-speed and high-resolution cameras (BASLER avA1600 65 k, 60 FPS, 1200 x 1600), was used for 4D data recording. The six cameras were hardware synchronized, and the grabbed frames from the six channels were used for building 4D facial data in the form of sequences of reconstructed 3D facial meshs. Each reconstructed 3D mesh contains over 50,000 vertices with a maximum edge length of 2 millimeters. Second, we used one grayscale camera (Stingray F-046B, 60FPS, 640x480) to capture

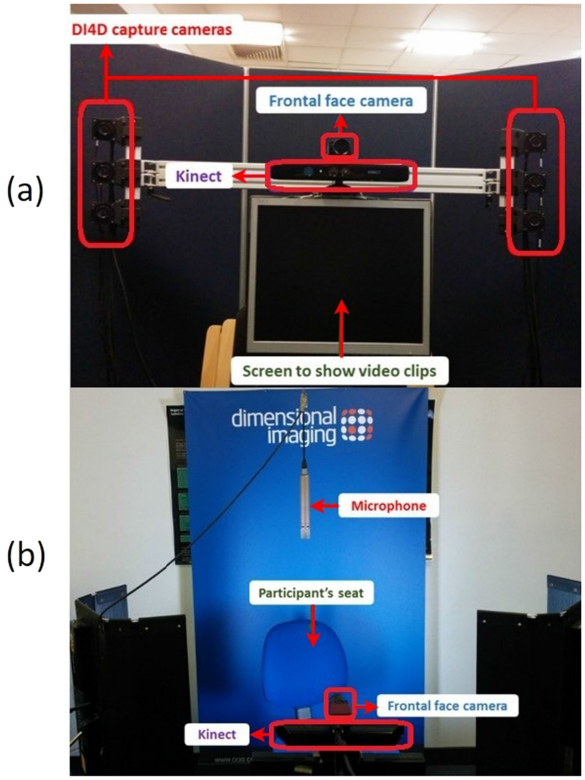


Fig. 1 Recording setup: (a) frontal view; (b) back view.

traditional 2D frontal facial videos. Third, one Kinect camera (Xbox 360, 30 FPS, 640 x 480) was used to record RGB videos and depth videos. All cameras were software-synchronized with triggers generated by the audio capturing system (the microphone as shown in Fig. 1).

3.2 Participants and Ethical Issues

All participants are volunteers recruited by posting advertisement in campus[2]. In total, 65 participants aging from 22 to 57 years (average age: 27.8 ± 3.5 years) were recruited for the data collection, of which 27 are females and 38 are males. The participants have multicultural backgrounds, i.e., 37 participants are from eastern Asia, 27 are from southern Europe (18 Greeks, four Spaniards, two Cypriots, one Serbian, one Portuguese and one French), and one is from Britain. Only one participant wears glasses. Due to an unexpected hardware failure, we were not able to reconstruct 4D data from nine participants. The rest 56 participants' data are complete and have been processed for annotation.

The research purpose and procedure were explained to each participant before the recording started, and the participants were well-aware that they can stop and quit the recording at anytime. One consent form was signed when the participant understood the contents and agreed to participate. Special questions were asked in the consent form concerning the data sharing issue, and the participants choose between two levels:1) all recorded data could be shared and used for research analysis, and facial images and videos can be published or presented for academic purposes, e.g., in paper publications, presentations, web-pages, or demos; 2)

1. https://github.com/liyantett?tab=projects (The data will be release after paper publication).

2 The study was approved by Ethics committee of human sciences, University of Oulu.

TABLE 2
Movie Clips for Inducing Emotions

	Description	Source	Time	Emotion
1	Pink flamingoes: a woman eating dog shit.	movie clip	50 s	disgust
2	Funny moments: collection of funny videos.	internet video	52 s	happy, surprise
3	There is something about Mary: a man fights with a mean dog.	movie clip	141 s	happy
4	Lion king: young Simba crying for his father's death.	movie clip	120 s	sad
5	Sea of love: doves flying by when a man sticks his head out of window.	movie clip	10 s	surprise
6	Hellraiser: a sticky, disgusting skull rising from the floor.	movie clip	87 s	disgust, fear
7	Church: collection of funny moments in church.	internet video	84 s	happy, surprise
8	Italian went to Malta: an Italian talking about experience in Malta.	cartoon video	73 s	happy
9	The thing: a beard man fires a gun with mysterious substance which can change people into monster.	movie clip	204 s	fear, disgust, surprise
10	Eye surgery: scene of conducting eye surgery.	internet video	69 s	fear, disgust
11	Natural scene: peaceful scenes with soothing music.	internet video	66 s	neutral, pleasant

all recorded data could be shared and used for research analysis, but facial images and videos cannot be published or presented, e.g., in paper publications, presentations, webpages, or demos. 30 participants agreed on level-1, and the rest 35 participants agreed on level-2.

3.3 Emotion Elicitation Procedure and Materials

It has been proved in previous studies [8], [14] that showing emotional movie clips to participants is a simple yet effective approach for inducing MEs. We adopted the same approach for the 4DME data recording. The participant was led to the seat and the height and orientation of the seat were adjusted to fit the cameras. Some participants were asked to tie up the hair or to wear a hair net to avoid occlusion of the facial parts. During the experiment, the participant was shown 11 carefully selected video clips (see Table 2) that are supposed to elicit various categories of strong emotions. There was a 1-minute break between two clips, during which the participant was asked to fill in a short survey regarding the subjective feeling of previous video. This 1-minute break was also served as a cool-down period to reset the emotion of participant. Throughout the whole experiment, the participant was required to HIDE his/her true feelings and always keep a poker face, and if failed, he/she needs to fill in a long boring questionnaire as the punishment. This setting was to create high-stake pressure and facilitate the occurrence of micro-expressions as Ekman [77] stated in his work. Before the actual recording started, there was one trial session for the participant to get familiar with the process.

3.4 Data Annotation

The 4DME dataset was annotated with both AUs and emotion categories. The annotation process was conducted in three steps. In the first step we did a rough manual segmentation. One annotator checked through all the raw videos to roughly mark out segments that may contain macro- or micro-expression movements. This step was done using an in-house video tagging software. Note that each of the marked-out segment may contain single or multiple macro-expressions and micro-expressions, as well as frames of neutral faces. The purpose of this step is to rule out the majority parts of videos in which there is no facial movement related to emotions (since we asked the participant to keep a poker face during the recording). The segments were clipped out from the long raw videos 1) for the second step of annotation, and 2) for the ME spotting task after the dataset is shared.

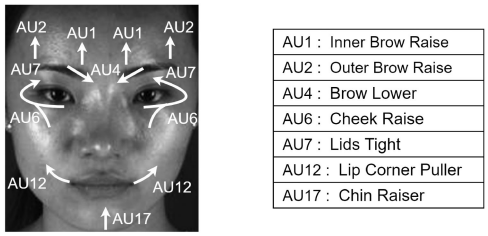


Fig. 2. The locations and motion patterns of key AUs.

In the second step we carried out fine-grained annotations (i.e., frame-by-frame) for AU labeling. Four annotators worked together for this task. The scope of AUs to be labeled were preliminarily decided by referring to related ME datasets studies, and according to the actual data we include 22 AUs in the final label book of 4DME. Fig. 2 shows the positions and motion patterns of several key AUs which have high occurrence in 4DME. Then we annotated the timestamps of AUs in all segments, specifically, the onset, apex, and offset frames of each occurred AU were marked. Three annotators worked separately and then cross checked to assure the frame-level labels. The reliability between two coders was calculated using the reliability equation proposed in [10], and difference within three frames was counted as consistent, for inconsistent cases the median of the three was selected. The average reliability of the frame coding of the three annotators is 0.79. This step focused on the timestamps while the AU categories were labeled in the next. Multiple AUs could occur at the same time, e.g., one or two main AUs (e.g., AU4 + AU7) might occur with minor ones (e.g., AU6, AU14 or AU15), which are difficult to differentiate and require professional skills. Thus two FACS [78] certified annotators examined the clips for an extra round to confirm the categories of AUs. The two annotators first worked separately and then cross checked with a reliability of 0.75.

Finally, we assigned five emotion categories to the clips, as positive, negative, surprise, repression and others. Clips shorter than 0.5 seconds (from onset to offset) were marked as micro-expressions, and clips of 0.5 to 4 seconds were marked as macro-expressions. Expressions that are static (i.e., lasting for over 4 seconds) were excluded due to lack of motion. The macro-expression cases could be used for, e.g., developing methods for joint recognition of macro- and micro-expressions, or to differentiate between these two categories which often co-occur in practical scenes. In the current study we focus on the MEs. Following the concept of compound emotions [26], we allow multiple emotion labels (maximum two) when necessary, as it was frequently encountered in our data

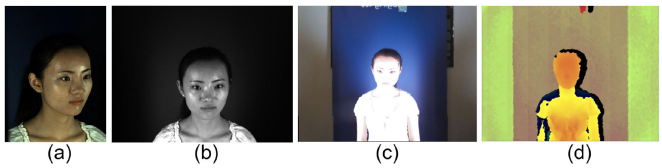


Fig. 3. Sample images of the four modalities in 4DME. (a) (one of the six) DI4D videos; (b) Grayscale videos; (c) Kinect-color videos; (d) Kinect-depth videos.

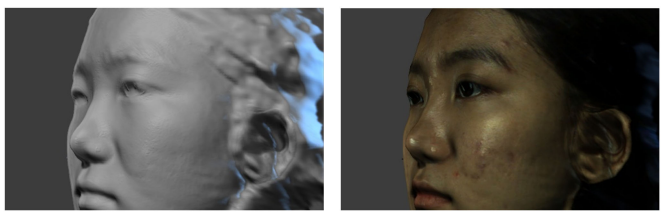


Fig. 4. A sample of reconstructed 3D mesh face. Left: without texture, right: with texture.

that multiple emotions occurred at the same time, e.g., 'happy' and 'surprised'. The emotion labels are primarily decided by the observed AUs rather than the inducing materials or self-reported emotions. More details about the relationship of AU and emotion labeling are discussed Section IV.

3.5 Data Statistics and Samples

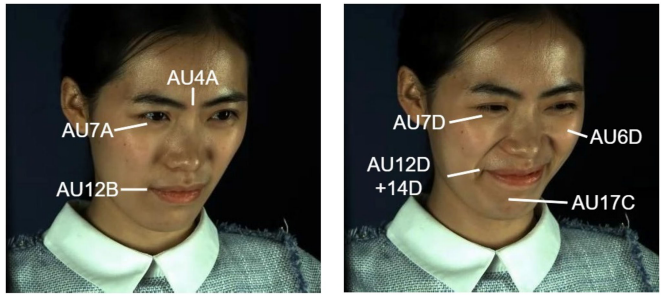
Around 5980 minutes of videos were recorded from 65 participants of four modalities, i.e., DI4D videos, frontal grayscale videos, Kinect-color videos, and Kinect-depth videos. Sample figures of the four modalities are shown in Fig. 3. After the first step of annotation, 278 segments (ranging from 0.77 to 9.82 seconds, mean duration is 2.49 seconds) were clipped out which include both micro- and macro-expressions and also some contextual frames of neutral faces. The 278 segments can be used for the ME spotting task, and the DI4D stereo images were used to reconstruct the 3D face meshes. One sample figure of the reconstructed 3D facial mesh is shown in Fig. 4: without texture on the left, and with texture on the right.

These selected segments were further labelled with micro- and macro-expressions. Note that not all the subjects displayed micro- or macro-expression, as some of them managed to keep a poker face throughout the whole recording session. In the final label book there are 267 MEs and 123 macro-expressions generated from 41 subjects, which add up to 1068 samples of MEs and 492 samples of macro-expressions of the four modalities. The clips were annotated with 22 categories of AU labels and five categories of emotion labels. Note that one clip may contain multiple AU labels and multiple emotion labels (maximum two emotions). The statistic of AU and emotion categories (of each modality) are shown in Table 3. One example of Micro-expression and one example of macro-expression from the same participant of 4DME are shown in Fig. 5.

TABLE 3
Emotion and AU Statistics of MEs (In Each Modality)

	Micro-expression						Macro-expression				
EM	AU		AU		EM	AU		AU		
P	34	AU1	46	AU15	82	P	15	AU1	33	AU15	4
N	127	AU2	54	AU17	11	N	41	AU2	37	AU17	1
S	30	AU4	117	AU20	2	S	24	AU4	47	AU20	1
R	6	AU5	8	AU24	9	R	4	AU5	7	AU24	7
RS	3	AU6	30	AU25	7	RS	1	AU6	13	AU25	0
PS	13	AU7	93	AU39	4	PS	5	AU7	38	AU39	0
NS	8	AU9	1	AU43	1	NS	5	AU9	1	AU43	1
PR	8	AU10	2	AU45	34	PR	4	AU10	3	AU45	29
NR	7	AU12	82	AU63	2	NR	6	AU12	41	AU63	0
OT	31	AU14	4	AU64	1	OT	18	AU14	7	AU64	0

P, N, S, R and OT represent Positive, Negative, Surprise, Repression, and Others respectively. EM represents emotion. EM represents emotion.



(a) Micro-expression (duration: 0.22s)

(b) Macro-expression (duration: 2.92s)

Fig. 5 Examples of a micro-expression and a macro-expression from 4DME, both belong to 'Others' emotion category. A, B, C, D and E indicate AU intensity from the lowest to the highest. E.g., 'AU4A' means AU4 (brow lower) with intensity level A.

4 AU Labeling and Emotion Categorization of Micro-Expressions

Six previous ME databases [8], [9], [10], [11], [12], [13] provide both AU labels and emotion categories (SMIC only provides emotion categories). In these database papers, general information about annotation were provided, e.g., which AUs and emotion categories are included, but explanations of how (the annotation was done) and why (certain labels were included) were insufficient in some papers. Furthermore, it lacks of a standardized or widely-accepted criteria of the ME annotation process, and the data from different datasets could be heterogeneous and some might have erroneous labels [79]). It would be beneficial for the ME research area if some of the detailed annotation problems could be further discussed, and hopefully lead towards unified and convincing solutions. In this section, we list several key problems/challenges for ME annotation, then explain our solution for the 4DME dataset annotation, and at last we point out the limitations to be sought out in future works.

4.1 Key Issues Related With ME Annotation and Proposed Solutions

We summarize three key questions related with AU labeling and three key questions related with ME emotion category labeling, and propose our solutions for 4DME.

4.1.1 AU Labeling Issues

One fundamental rule for AU labeling is to follow the instructions of the FACS. But the FACS instructions are wide

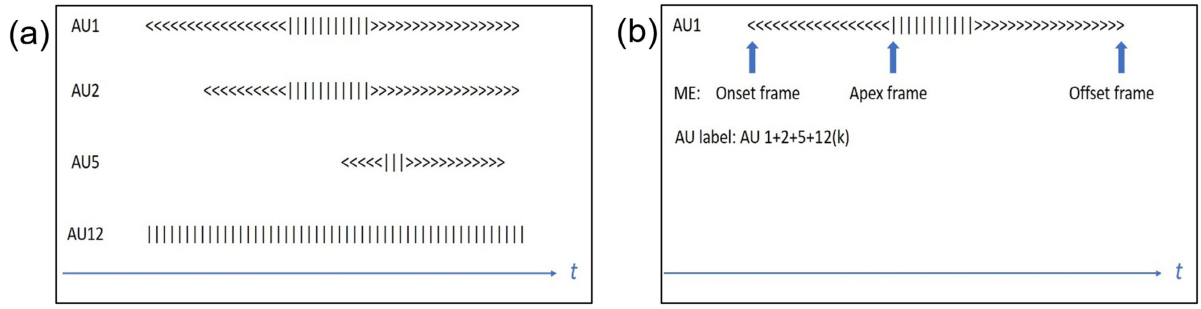


Fig. 6. (a). AU 1+2+5+12 with overlapped time spans, fully labelled with each AU's phases. '<' indicates onset phase, '|' indicates apex phase, and '>' indicates offset phase. (b). The label provided in 4DME. AU1 decides the onset, apex and offset of the ME clip, with co-occurred AU2, AU5 and AU12 in the label. '(k)' indicates AU12 is a static AU with no motion change within in the clip.

and general for annotating all possible facial movements, which need to be tailored and selected for the purpose of ME annotation.

Q1 Which AUs should be included?

The solutions of previous ME datasets [8], [9], [10], [11], [12], [13] are consistent on one side, i.e., 11 essential AUs directly related with emotional expressions were labeled in all six databases, including AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14 and AU24. But on the other side, for some minor AU categories (e.g., AU 15, AU17, AU18, AU25, AU45, and so on), previous databases took different solutions. CAS(ME)^2 and SAMM provided more AU categories than CASME, CASME II and MEVIEW.

The main motivation for providing AU labels for ME clips is to train models which can detect AUs and use them for recognizing the ME emotion category. In practice, it is reasonable to prioritize AUs with high occurrence, as AUs with too few samples are usually left out in training. In the current 4DME labeling, besides those essential AUs mentioned above, we also include AU17 (chin raiser) and AU24 (lip pressor) as we think they relate with a special emotional state of 'repression', i.e., indicates suppressing movements to prevent leaks of expressions, and we include AU45 (eye blink) as it occurs frequently and greatly interferes the detection of eye region AUs.

Q2 How to decide the duration (onset, apex, and offset) of each AU?

The six previous datasets [8], [9], [10], [11], [12], [13] all marked the onset, apex, and offset frames of each sample in their labels, but explanations about how to do it were limited in the papers.

The task is theoretically clear but difficult to conduct in practice because 1) MEs have very low motion intensity and 2) usually high-speed cameras are used for recordings and adjacent frames are quite similar. In [8], the authors described about how they chose the onset and apex frames in a footnote. Compared with the onset and the apex, the offset frames are more difficult and ambiguous to find as it is often the case that the facial muscles did not return to the relaxed state (i.e., a neutral face). In the 4DME labeling, we took two practical approaches as solutions. First, we have three annotators to check the onset, apex and offset frames of each ME clip independently, and the median frame of the three is selected for inconsistent cases to reduce personal bias. A similar approach was adopted in [10], that the average of two coders' selected frames was selected for disagreed cases. Second, we assigned an operational definition for localizing the offset frames, i.e., to find the last frame with visible offset motions, which is not necessary a complete neutral face but fixed at a stable state with no motion.

Q3 How to treat AUs with time overlaps?

Multiple AUs may occur at the same time with partially overlapped time spans. Previous databases reported combination of AUs in many samples, but the time overlap among individual AUs has not been specifically discussed. Ideally each AU could be labelled with its own starting and ending points (as illustrated in Fig. 6a), but it can hardly be achieved as that would be very time consuming. In 4DME labeling, we focus on one major AU (i.e., AU1 in the example) whose phases decide the onset, apex and offset frames of the ME clip, and then mark all AUs that occurred within the clip (Fig. 6b).

4.1.2 Emotion Category Labeling Issues

Categorical emotion labels are more prevailing than dimensional labels in facial expression datasets [80]. Although there are still ongoing debates and explorations in psychological studies about categorical emotion theories, expressions such as happy, sad, surprise, fear, anger and disgust are widely accepted by the community. But all these works are based on observations of ordinary facial expressions. Some previous ME studies [11], [12], [13] obscurely assumed that the emotion categories of MEs could be aligned with that of ordinary facial expressions, which needs to be further verified. Compared with ordinary facial expression, some special characteristics could be observed from ME data: 1) great efforts to *control* and *suppress* the true feelings, 2) very low intensity or even incomplete behaviors, and 3) could be consecutive momentary fast changes. These need to be considered when assigning emotion categories for ME data.

Q1 Which emotion categories should be included?

The emotion categories are primarily decided by the target emotion of the inducing materials (e.g., emotional movie clips). The actually induced emotions are also dependent on the task (i.e., hide true feelings and keep a neutral face) and the participant's subjective feelings (i.e., self-reports). Most previous ME datasets [8], [9], [10], [11], [13], [14] adopted similar emotion inducement method, i.e., by showing emotional movie clips (containing the six basic emotions) to participants and asking them to hide true feelings, except

MEVIEW which contains in-the-wild data of poker game videos. For the 4DME data, the inducing materials contain five categories of emotions, i.e., happiness, surprise, sadness, disgust, and fear. Considering the task of suppressing true feelings, we think it is reasonable to add one extra ME emotion category of 'repression', which indicates the suppressing movements when the subject is about to leak true feelings, e.g., tightening or pressing lips. Two previous ME datasets CASME and CASME II also included the 'repression' category based on similar reasons and observations of the data. Thus we consider the six emotion categories as the initial candidates for 4DME emotion labels which are further adjusted concerning the two following questions.

Q2 How to decide the emotion category for each ME case?

Clues for deciding the emotion category of an ME come from three sources, 1) emotions of the inducing movie clips, 2) the subject's self-reported emotions, and 3) the facial movements or the AUs. Previous ME datasets took different ways for assigning emotion labels. For SMIC, emotion categories were primarily assigned based on self-reports; For CASME, CASME II and CAS(ME)2, emotion categories were based on all three sources; other dataset papers didn't directly specify how the emotion category was decided for each ME case, although some paper [11] elaborated on questionnaires and video ratings. One limitation of the first two sources is that they both lack granularity and can only be used to summarise the whole video clip, except the case in CAS(ME)2 that each participant reviewed his/her own videos and reported on each single expression, which would be very demanding for participants. The emotional status fluctuates all the time especially when strong emotional stimulus is presented. During one movie clip, the subject may feel surprised and disgust, and try to suppress the responses, and then feel funny or happy. The emotional responses can be complex and frequently switching while the subject might only report 'disgust' in the self-report. The labeling is to assign emotion labels to each ME clip for that transient time, thus *should be primarily dependent by the occurred AUs*. This leads to the next essential question.

Q3 How to understand the relationship between AUs and emotion categories?

Mapping AUs to emotion categories for MEs is an essential research question needs to be explored in depth. One directly related reference is the Table 1 (page 136) in the FACS Investigator's Guide, which lists AU or AU combinations to the corresponding emotion categories. This should serve as a primary rule for AU and ME emotion category mapping. However, it was designed for ordinary facial expressions and might not be suitable to be directly used for ME cases. As subjects are voluntarily suppressing their facial movements, in most cases MEs only present partial or fragmented motions and it was hardly seen that a full set of AUs (e.g., AU1+2+5+25 for surprise) could all appear in one ME clip. A new table is needed for mapping AUs to ME emotions. Several AU-Emo mapping tables were proposed in previous studies [9], [10], [13], but not all tables are easy to follow. In [9], [10], only partial AUs (combinations) were listed to corresponding emotional categories and the rest were not specified. In [13] several AUs (combinations) were linked to multiple emotions, e.g., AU1+2 can be either 'surprise' or 'sadness'. We think it is more helpful if the table provides full-scope clearly defined, exclusive AU-Emo correspondences, i.e., the occurrence of AU X indicate ME emotion Y but not others.

TABLE 4
Mapping AUs to Emotion Categories of MEs

Emotion	Key AUs
Negative	AU1, AU4, AU7, AU9, AU10
Positive	AU12
Surprise	AU2, AU1+2
Repression	AU14, AU15, AU17, AU24
Others (dependent)	AU5, AU6, AU20, AU25
No emotion	AU39, AU43, AU45, AU56, AU58, AU63, AU64

We propose a preliminary AU - Emo mapping table as shown in Table 4. We start with 12 key AUs (Row 1 to Row 4) as the 'decisive' AUs. For example, if AU12 occurs, 'Positive' emotion will be labeled; or if AU4 occurs, 'Negative' emotion will be labeled. Four AUs (Row 5) are 'dependent' AUs, which means that they are related with emotions, but their occurrence is not decisive to one emotion category, i.e., they can be combined with various decisive AUs and compositely link to multiple emotions. The rest seven AUs (Row 6) have no emotional content. These mappings were carefully summarized with the premises, that *they should not conflict to the FACS Investigator's Guide table*. Besides, the following rules were also followed for 4DME labeling:

1) We choose to use five emotion categories: Positive, Negative, Surprise, Repression, and Others, which are theoretically clear and practically feasible.
2) Negative is not further divided, as it is not feasible to reliably map AUs to fine classes, e.g., AU4 and AU7 are with the highest occurrences and observed for all reported negative emotions 'fear,' 'disgust' and 'sad'.
3) We allow multi-emotion labeling, e.g., surprise + positive, with maximum two emotions.
4) Clips with very complex AU combinations. e.g., correspond to three or more emotions, are labeled as 'Others'.
5) Clips containing only 'dependence' AUs are labeled as 'Others'.
6) Clips containing key AUs for both 'Positive' and 'Negative' are assigned to 'Others' as these two are conflicting (e.g., as the examples in Fig 5).
7) Static AUs (e.g., AU12 in Fig. 6) and active AUs are both considered when assigning emotion labels.

4.2 Other Annotation Problems for Further Discussion

We summarized some preliminary rules according to our observations during data annotation, which may help in future data annotation work to achieve more unified data. They are not all well-sorted or 'ideal', and there are other problems to be further discussed in future. For example, some AUs occur at the same facial location and have similar appearance, e.g., AU12 and AU14, which are difficult to differentiate at very low intensity level even for experienced and certified

AU annotators. We can only make the best guess according to observations of the person's behaviors. Another challenging issue is about emotion categorization of complex AU combinations. In Table 4 we added one extra emotion category of 'Others' for those complicated cases, e.g., AUs for three or more or conflicting emotions. Several previous ME datasets also included the 'Others' category such as CASME II, CAS (ME)^2, SAMM, and MMEW, and MEVIEW named it as 'Unclear'. It needs to be further discussed whether such complex and transient emotions are theoretically reasonable, and how the rules of AU-emotion mapping could be further refined. One specific question is that whether the static AUs (e.g., AU12 in Fig. 6) should be considered when assigning emotion categories for the ME clip (e.g., AU 1+2+5+12).

5 Database Evaluation

The 4DME dataset contains four modalities of data including both 2D videos and 4D videos. In order to compare the effectiveness of different modalities for the task of ME recognition, we carry out separate experiments to evaluate on 2D frontal grayscale videos (Fig. 3b) and reconstructed 4D videos (Fig. 4) and report the performance as the baseline results in the two following subsections.

5.1 Evaluation on 2D Video Data

First, we carry out experiments on the 2D frontal grayscale video data for two tasks: i.e., AU detection and ME recognition. Three approaches proposed in previous works are employed for comparison, including both classic spatial-temporal descriptor of LBP-TOP [81], and deep neural network approaches of Res3D [82] and Res3D+SCA [58].

5.1.1 Method

Preprocessing. the labeled out ME clips are first preprocessed with face detection and registration before applying the three approaches. Although there are low level of rigid movements within each ME clip, registration is still needed to remove scale, rotation and translation differences across all the ME clips. To this end, we align all the faces to one pre-defined template face by using 68 facial landmarks detected with the method proposed in [31]. Then the face region are cropped to the size of 150×150 pixels according to the eye-coordinates.

Approach 1 LBP-TOP. Local binary patterns (LBP) [83] is a local binary operator which has been verified to be a powerful feature for texture classification tasks [84]. Zhao *et al.* [81] extended the LBP to LBP-TOP, which describes dynamic texture on three dimensions. LBP-TOP has been employed for ME recognition in multiple papers and has been demonstrated to be very effective. Here we use LBP-TOP feature with the SVM classifier as the first approach to provide baseline results on 2D videos of 4DME.

All facial images are first divided into 5×5 blocks, then LBP-TOP features are computed for each block and concatenated from the three orthogonal planes (XY, XT, and YT planes). The features from XT and YT planes encode the vertical and the horizontal motion patterns, respectively. Specifically, the radii in axes (X, Y, T) are set to (1,1,2). The number of neighboring points in the XY, XT and YT planes are set to 8. The features of all blocks are concatenated as one vector to represent the whole ME clip. The extracted features are fed to a one-vs-rest Linear SVM which is trained as a classifier for each emotion or AU category. The classification penalty factor C is set to 1000.

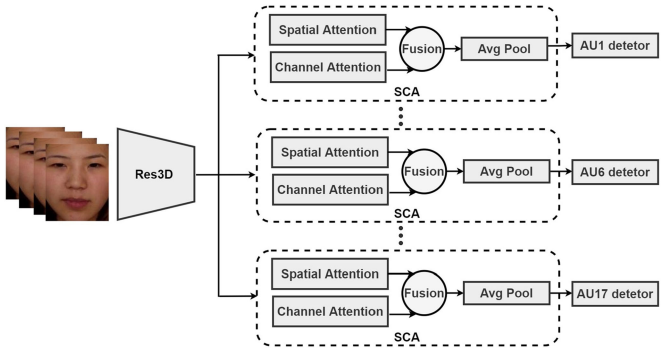


Fig. 7. The framework of Res3D+SCA (approach 3).

Approach 2 Res3D. a residual neural network (ResNet) [85] is a neural network utilizing skip connections over layers. Such a skipping mechanism can effectively simplify the network and avoid gradients vanishing by reusing activations from previous layers. ResNet has been demonstrated to be effective for discriminate features generation and achieve excellent performance on various computer vision tasks. As micro-expressions involve fast movements in the temporal domain, the ability of capturing temporal information is essential for solving the micro-expression recognition task. 3D residual network (Res3D) is able to incorporate both spatial and temporal information, which has been employed for ME recognition in previous works [53], [86], and here we test it as the second approach to provide baseline results.

Approach 3 Res3D+SCA. one common challenge shared by the two tasks of AU detection and ME recognition on ME dataset is that the involved movements are of very low intensity. To alleviate this problem, Li *et al.* [58] utilized a Spatio-Channel Attention (SCA) mechanism to better represent the subtle movements. Specifically, SCA mechanism explores the second-order correlations of spatio-wise and channel-wise features to explore the relationship information and discriminative information on various local regions, as shown in Fig. 7. Here we test the Res3D+SCA as the 3 rd approach to provide baseline results. More details of the approach can be found in [58].

For the two approaches of Res3D and Res3D+SCA, the input is ME sequential images. As the length of ME sequences vary largely, we interpolate the clips into a fixed length of 10 using the Temporal Interpolation Model (TIM) [33]. Then the interpolated clips are cropped to random patches of 112×112 for data augmentation. All models are pre-trained on Kinetics [87] and UCF-101 [88] databases. In the training process, the networks are optimized through stochastic gradient descent (SGD) with a weight decay of 0.001. The initial learning rate is set to 0.01, divided by 10 every 40 epochs until 80 epochs. All processes are implemented on Pytorch.

5.1.2 Evaluation Protocol and Metrics

A subject-independent 5-fold cross-validation protocol is employed in the following experiments. All subjects are

TABLE 5
AU Detection Performance on 2D Videos of 4DME

Metrics		AU1	AU2	AU4	AU6	AU7	AU12	AU17	AU45	Average
	Number of samples	46	54	117	30	93	82	11	34	
F1-score	LBP-TOP	0.4657	0.4713	0.4674	**0.5753**	0.5052	0.4588	0.5541	0.4526	0.4938
	Res3D	0.7130	**0.7750**	0.6881	0.4681	**0.6494**	0.5968	0.4816	0.7750	0.6434
	Res3D+SCA	**0.7307**	0.7418	**0.7362**	0.4865	0.6360	**0.6714**	**0.6452**	**0.7752**	**0.6779**
Accuracy (%)	LBP-TOP	71.53	66.66	56.92	84.26	57.67	66.66	94.38	79.40	72.18
	Res3D	**86.14**	**86.51**	69.28	**88.01**	**69.66**	63.67	92.88	90.63	80.84
	Res3D+SCA	**86.14**	85.01	**74.15**	85.76	68.91	**73.03**	**96.62**	**90.26**	**82.48**

TABLE 6
ME Emotion Recognition Performance on 2D Videos of 4DME

Metrics		Postive	Negative	Surprise	Repression	Others	Average
	Number of samples	56	142	54	24	31	
F1-score	LBP-TOP	0.4809	0.4741	0.4713	0.5499	0.5025	0.4958
	Res3D	0.5730	0.7064	0.7086	0.4865	0.4984	0.5946
	Res3D+SCA	0.6501	0.7231	**0.7455**	0.5679	0.5542	0.6481
	AU-graph	**0.6692**	**0.7385**	0.7126	**0.5807**	**0.5905**	**0.6590**
Accuracy (%)	LBP-TOP	71.16	54.31	71.16	88.39	84.64	72.73
	Res3D	73.41	70.78	82.39	85.76	83.89	79.24
	Res3D+SCA	79.03	72.66	**84.27**	90.26	86.51	82.54
	AU-graph	**80.52**	**74.15**	82.77	**91.39**	**88.39**	**83.44**

randomly divided into five folds with the consideration of roughly balanced sample numbers in each fold (i.e., ME samples from each subject vary largely). For the task of AU detection, eight AU categories are considered which contain more than ten samples, while the rest AUs with too few samples are excluded. In general, the emotions and AUs are roughly balanced in each fold and every fold contains all kinds of emotions and AUs. The specific subject information of the 5-fold protocol will be released with the database.

As explained in the section of data annotation, we allow multiple emotion-labels and AU-labels for each ME clip, thus the two tasks of ME recognition and AU detection are considered as multi-label binary classification problems. Our task is to detect whether one emotion or one AU is active or not. In our experiment, both accuracy and F1-score are utilized to evaluate the performance for detecting eight AUs and classifying five emotions. For a binary classification task especially when the samples are not balanced, it is better to incorporate F1-score with accuracy to interpret the algorithm performance. We follow [89] for the computation of the two evaluation metrics

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}, \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)}, \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)}, \quad (3)$$

$$F1 - score = 2 \times \frac{Precision * Recall}{(Precision + Recall)}, \quad (4)$$

in which TP, TN, FP, FN represent true positive, true negative, false positive, and false negative, respectively.

5.1.3 Results of AU Detection

We first evaluate the three approaches for the task of AU detection using the 2D frontal grayscale video data of 4DME dataset. Eight categories of AUs with more than ten samples are considered, and the results are shown in Table 5. From the two tables it can be seen that, for most AU categories the performance of the three tested approaches is Res3D+SCA $>$ Res3D $>$ LBP-TOP for both accuracy and F1-scores. The best average accuracy is 82.48% and the best average F1-score is 0.6779, which are both achieved by using Res3D +SCA. The results are consistent with previous findings.

Besides, the performance vary for different AU categories. LBP-TOP achieves better F1-score on AU6. One possible reason might be that AU6 involves blurry motions with subtle texture change, while deep-based methods perform better on AUs involve clear motions creating lines or edges, e.g., AU1 (Inner brow raiser), AU2 (Outer brow raiser), AU4 (Brow lower), AU12 (Lip corner puller), and AU45 (Eye blink).

5.1.4 Results of ME Emotion Recognition

We then evaluate the three approaches for the task of ME emotion recognition using the 2D frontal grayscale video data of 4DME dataset. Five categories of emotions are considered, and the results are shown in Table 6. Generally, the two deep learning based methods (Res3D and Res3D+SCA) outperform the traditional LBP-TOP approach. The Res3D+SCA achieves the best performance, i.e., the average F1-score of

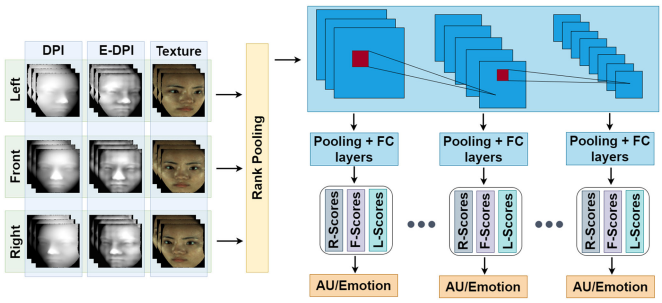


Fig. 8. A Collaborative Cross-domain Dynamic Image Network (CCDN) [21] for AU detection and ME emotion recognition on 4D data.

0.6481 and the average accuracy of 82.54% of the five emotion categories. Among the five emotions, it seems that the category of 'surprise' gets the best performance if we concern both accuracy and F1-score, while the evaluation of 'repression' and 'others' categories are dependent on the metrics due to smaller sample sizes.

5.2 Evaluation on 4D Data

Second, we carried out experiment on reconstructed 4D ME clips for the two tasks: i.e., AU detection and ME recognition. One 4D-based approach, the Collaborative Cross-domain Dynamic Image Network (CCDN) [21] which was proposed for 4D ordinary FE recognition was employed and we compare the results achieved with three single views and fused Multi-views.

5.2.1 Method

A pre-processing step is needed for the 4D data due to the reason that the 3D facial meshes may contain artefacts beyond facial regions generated during the reconstruction process. There might be noisy and unwanted mesh points in the facial regions as well. These interfering components can create problems during model training which have been found in previous studies, and it will be more severe considering the MEs are more fragile and subtle phenomenon. Henceforth, a strong pre-processing procedure is needed. Specifically, since we use three facial profiles in our baseline experiments (left, right and front), we process each 3D facial mesh to first straighten the facial posture [90], and then using annotated 3D facial landmarks [91] to rotate the 3D mesh to obtain three alignment profiles. For cropping the face, we removed the vertices beyond the facial regions. Afterwards, using 3D to 2D projection, we obtain the depth images (DPI), enhanced depth images (E-DPI) [21], and texture images for all the three profiles as shown in Fig. 8.

We duplicate the extracted set by applying Eulerian Video Magnification (EVM) [34] as it was demonstrated in [21] that motion magnification can help to improve the emotion recognition performance. A collaborative recognition strategy was employed where all the three views jointly collaborate in final predictions. As shown in Fig. 8, we obtain the rank pooling images [63] of all three profiles with their image domains. The rank pooling helps tremendously in encapsulating the temporal facial dynamics into single images which are then fed into a GoogLeNet model [92] for learning MEs. The independent predictions from the output of the deep models for each view then collaborate to yield a more robust final prediction. More details of CCDN approach can be refered to the original paper. It is expected that this method can intuitively demonstrate the importance of 4D data as it brings enriched amount of information than single view faces in 2D videos. Additionally, following prior works for using motion magnification [34] for improvements in micro-expression recognition [28], we also analyze its effect on the 4D data and compare the results.

5.2.2 Evaluation Protocol and Metrics

For fair comparisons, we followed the same evaluation protocol and used the same metrics as we used for the experiments on 2D video data.

5.2.3 Results of AU Detection

We first conduct experiments for AU detection on eight categories of AUs which have more than ten samples. Results of using three individual views and fused multi-views are shown in Table 7 in terms of both F1-score and accuracy. First, if we compare the performance of using the three individual views, it can be seen that the Front view achieved the highest performance of the three, followed by the Left view, and the Right view achieved the lowest performance. Furthermore, when using the three views collaboratively, i.e., the Multi-views, to recognize the AUs, it achieved an average F1-score of 0.7990 and an average accuracy of 86.55% which are significantly higher than any of the three single view. The results match well with our expectations, that the AU detection performance will be better when more facial areas are revealed (unblocked), i.e., Multi-views $>$ Frontal $>$ Left or Right. The 4D data carries more information and improves the system's performance.

TABLE 7
AU Detection Performance on 4D Data of 4DME

Metrics	Profiles	AU1	AU2	AU4	AU6	AU7	AU12	AU17	AU45	Average
F1-score	Left	0.6580	0.6898	0.6820	0.5468	0.7191	0.7099	0.5572	0.5367	0.6374
	Right	0.5959	0.6543	0.6683	0.4517	0.6751	0.6776	0.4241	0.6246	0.5964
	Front	0.7522	0.7833	0.8068	0.6585	0.7326	0.8000	0.6353	0.7705	0.7424
	Multi-views	**0.7941**	**0.8197**	**0.8260**	**0.7648**	**0.7773**	**0.8161**	**0.7318**	**0.8623**	**0.7990**
Accuracy (%)	Left	70.34	72.46	68.64	65.25	72.46	72.46	74.58	61.02	69.65
	Right	64.83	69.92	66.95	52.97	68.64	69.92	53.81	69.07	64.51
	Front	82.63	83.47	80.93	78.39	74.15	80.93	84.75	85.59	81.36
	Multi-views	**87.29**	**86.86**	**83.05**	**87.29**	**78.81**	**83.05**	**93.22**	**92.80**	**86.55**

TABLE 8
ME Emotion Recognition Performance on 4D Data of 4DME

Metrics	Profiles	Positive	Negative	Surprise	Repression	Others	Average
F1-score	Left	0.5971	0.6639	0.6040	0.5398	0.5804	0.5970
	Right	0.5249	0.6601	0.5900	0.5404	0.5739	0.5778
	Front	0.6367	0.6766	0.6313	0.7059	0.7298	0.6760
	Multi-views	**0.7443**	**0.8347**	**0.8034**	**0.7966**	**0.7750**	**0.7908**
Accuracy (%)	Left	66.10	66.53	66.95	65.68	69.07	66.86
	Right	61.02	66.10	64.83	66.53	68.22	65.34
	Front	69.07	68.22	67.80	82.63	83.90	74.32
	Multi-views	**80.08**	**83.47**	**85.59**	**91.10**	**87.71**	**85.59**

5.2.4 Results of ME Emotion Recognition

We then evaluate the CCDN method for the task of ME emotion recognition using the 4D data. All five categories of emotions are considered, and the results are shown in Table 8 for both metrics of the accuracy (%) and the F1-score. Similar performance patterns like the AU detection task could be observed here. First, for the three individual views, the Front view achieved the best performance of the three followed by the Left view, and the Right view achieved the lowest performance. Second, the Multi-views outperformed the three individual views and achieved an average F1-score of 0.7908 and an average accuracy of 85.59 %, which again demonstrated the advantage of 4D data for ME recognition task. The advantage of using 4D data by fusing Multi-views is consistent through all five emotion categories.

5.2.5 Discussion of Results on 4D and 2D Video Data

If put under the strictest rules, we think the results achieved on 4D data and on 2D videos are not directly comparable as the data and process approaches used are all different.

Results in Tables 7 and 8 show that, fusing clues from multiple views can work more efficiently than any of the single views. Multi-view videos can be obtained from 4D data while a single view is like a 2D video, thus it serves as a form of direct comparisons between performance of 4D and 2D video.

The results demonstrate our hypothesis that *4D data has potential advantages for the ME recognition task*, as MEs are very subtle movements that might only occur on one small region of the face, which are not always visible in a 2D single view video.

If we directly compare results of 2D videos and 4D, the advantage of 4D results can be observed on both metrics (about 2.1 % difference for accuracy values, and over 12 % difference for F1 scores) of the average results. Note that 4D based methods are not as well-developed as 2D methods for ME analysis due to lack of data. We replicated state-of-the-art 2D methods for ME recognition, but the 4D method CCDN was designed for ordinary facial expression analysis. There is no 4D method available yet specifically for ME recognition. We hope with our 4DME dataset, new 4D methods could be designed specifically for ME recognition.

5.3 Learning AUs for ME Emotion Recognition

In the third experiment, we would like to further verify the relationship between AUs and emotion categories. As discussed in previous sections, we assigned ME emotion labels depending on observed AUs, as we think this is a more objective and reliable way for annotation. Theoretically, there should be a fixed correspondence between the AUs and ME emotions, and we would like to demonstrate this in two steps. First, we explore the relationships between the activation maps of AUs and emotions learnt by a network. Second, we explore whether learning AU information would help a neural network to better recognize ME emotion categories. We use the 2D video data and 2D-based approaches for this part of experiment.

5.3.1 Relationships of the Activation Maps

In Section V.A, the Res3D model was trained separately either to learn different AU classes, or to learn different emotion classes. It would be interesting to know the specific activation regions that the model has learnt for each AU or emotion class, and the relationships between the activation maps. We adopt the Grad-Cam [93] approach to compute the Class Activation Maps (CAMs) for each AU and emotion class. The average CAMs for each AU and emotion is achieved by averaging the CAMs of all samples in the form of a 112 by 112 matrix, as shown in Figures 9.

The CAMs indicate corresponding facial regions that the network learnt that are important (assigned higher weights) for recognizing one AU or one emotion. From Fig. 9 it can be seen that the activated regions are related with the location of AUs, e.g., the CAM of AU4 is mostly activated in the upper half, and the CAM of AU17 is mostly activated in the lower half. The CAMs of emotions are more diffused.

We also computed the Pearson correlation coefficients of the CAMs in order to validate the proposed theoretical AU-Emo

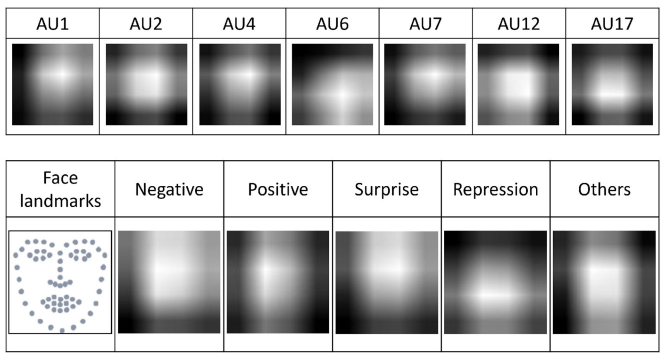


Fig. 9. Visualization of averaged CAMs for AUs and emotions. Lighter color indicates higher associated weight.

TABLE 9
Correlation Coefficients of CAMs: AU versus AU, and AU versus Emotion

	AU1	AU2	AU4	AU7	AU6	AU12	AU17
AU1	1	0.81	0.88	0.96	0.41	0.45	0.58
AU2	0.81	1	0.96	0.90	0.42	0.81	0.88
AU4	0.88	0.96	1	0.93	0.40	0.77	0.80
AU7	0.96	0.90	0.93	1	0.47	0.57	0.69
AU6	0.41	0.42	0.40	0.47	1	0.57	0.65
AU12	0.45	0.81	0.77	0.57	0.57	1	0.92
AU17	0.58	0.88	0.80	0.69	0.65	0.92	1
Negative	0.87	0.83	0.83	0.86	0.10	0.40	0.58
Positive	0.63	0.83	0.84	0.64	0.14	0.77	0.77
Surprise	0.90	0.79	0.83	0.76	0.04	0.33	0.47
Repression	0.37	0.65	0.58	0.47	0.79	0.89	0.92
Others	0.68	0.83	0.86	0.70	0.48	0.83	0.87

pairs in Table 4. The correlation coefficients are listed in Table 9. A larger value of the coefficient (range $[-1, 1]$) indicates stronger correspondence of activated regions. If a stronger correspondence could be observed between the learnt CAMs of one AU-Emo pair, e.g., AU4-Negative, then it could work as a supporting evidence for the proposed AU-Emo mapping. From the lower part of Table 9, it can be seen that the results match well the AU-Emo pairs we proposed in Table 4, e.g., AU1, AU4 and AU7 for Negative, AU12 for Positive, AU1+2 for Surprise, AU17 for Repression, all have high correlation coefficients as marked in red.

One thing to notice is that the CAMs analysis only concerns the locations. Some AUs activate similar regions as they occur at the same (or adjacent) location, thus have higher inter-correlations between AU pairs, such as *AU1 - AU2 - AU4 - AU7*, and *AU12 - AU17*, as marked in grey in the top part of Table 9. Although the activated regions are similar, the model learns different features depending on the motions. The results in the table should not be deduced in the opposite way, i.e., a high correspondence does not necessarily mean the AU is 'decisive' for that emotion. For example, the coefficient is 0.89 for AU12-Repression which might because AU12 and AU17 activated similar regions.

The CAM visualization and analysis indicate that, although the Res3D is not yet working perfectly for AU or ME recognition, it does capture the important facial regions for each class. The AUs and emotions were trained separately, i.e., when trained for emotions the model has no info about AU labels, but the learnt CAMs show consistent AU-Emo relationship patterns as we proposed in Table 4, which provide supportive evidence for our arguments.

5.3.2 Learning AUs for ME Emotion Recognition

Since the CAM analysis supports the AU-Emo mapping, we further explore whether learning AUs could help the network to achieve better performance for ME emotion recognition. We adopt the Res3D+SCA model and add a graph convolutional network (GCN) module [94] to form an end-to-end framework, referred as *AU-graph*, which can learn AUs for the task of ME emotion recognition, as shown in Fig. 10. First, the SCA is utilized to detect AUs. Then the detected AUs are passed through a GCN to recognize the emotion of ME.

Specifically, the detected AU probability represents each node in the node matrix of the graph. The adjacency matrix

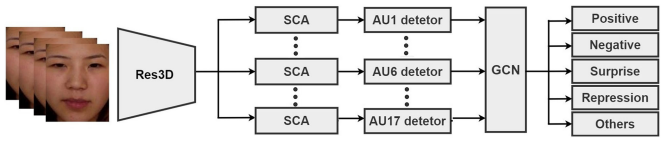


Fig. 10. The framework of AU-graph model, which uses detected AUs for ME emotion recognition.

$A_{AU} \in R^{8\times8}$ is built based on the occurrence relationship between the AUs via a data-driven approach. These two components are fed into the GCN of one layer for feature learning. The final loss L_{MEAUs} is composed of the AU detection loss L_{AUs} and the ME emotion loss L_{ME}

$$L_{MEAUs} = \alpha L_{AUs} + (1.1 - \alpha) L_{ME}, \tag{5}$$

where α is the weight balancing the two losses. As we want to explore the effectiveness of AUs for ME emotion recognition, the training focus on the AUs at first. The α is initialized as 1.0, and then divided by 10 after every 30 epoch. The total training epoch is 90. The results are shown at the last row in Table 6, referred as 'AU-graph'. It can be seen that compared with the Res3D+SCA model, *the AU-graph model increased the average F1-score by 1.68% and increased the average accuracy by 1.09%*. The results demonstrated that learning AUs could facilitate the model for ME emotion recognition. The improvement is not large as the learnt AUs (detected by SCA) are not 100% accurate which can be potentially improved.

6 Conclusion

We introduced a new spontaneous ME dataset, the 4DME. 4DME contains multimodal facial videos, including reconstructed dynamic 3D facial meshes, grayscale 2D frontal facial videos, Kinect-color videos, and Kinect-depth videos. Both micro- and macro-expression clips are labeled out, and AU labels and emotion categories are annotated. Experiments were carried out using three 2D-based methods (LBP-TOP, Res3D, Res3D+SCA) and one 4D-based method (CCDN) to provide baseline results. Preliminary findings support our hypothesis that 4D data can benefit the task of ME recognition. The previous ME datasets lack data variability, and we think that our proposed 4DME dataset is valuable for handling this by: 1) exploring 4D-based methods, and 2) exploring fusion of various modalities for ME recognition study in the future.

Besides, several key questions about ME annotation were summarized and discussed, especially about the relationship of AU labels and ME emotion categories. A preliminary AU-Emo mapping table was proposed with justified explanations and supportive experimental results. Future ME study needs more high quality data from multiple contributors. Unified data annotation rules would allow better data fusion, while data with free-form labels (or erroneous labels) are hard to use, which would be a waste of efforts. Arguably, more works are needed to tackle unsolved questions before we can reach a clearly-defined and widely-accepted criteria for ME annotation. We hope the current work can draw attention of the research community to focus on the issues and join in future discussion.

ACKNOWLEDGMENTS

The authors would like to thank CSC Finland for providing computational resources. Xiaobai Li and Shiyang Cheng Contributed equally.

REFERENCES

[1] R. W. Picard, *Affective Computing*, Cambridge, MA, USA: MIT Press, 2000.
[2] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*, Berlin, Germany: Springer, 1966, pp. 154–165.
[3] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, 2013.
[4] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation Emotion*, vol. 35, no. 2, pp. 181–191, 2011.
[5] S. Porter and L. T. Brinke, "Reading between the lies identifying concealed and falsified emotions in universal facial expressions," *Psychol. Sci.*, vol. 19, no. 5, pp. 508–514, 2009.
[6] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, 2003.
[7] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1449–1456.
[8] 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. IEEE Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–7.
[9] W. J. Yan *et al.*, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS One*, vol. 9, no. 1, 2014, Art. no. e86041.
[10] F. Qu, S. Wang, W. 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.–Dec. 2018.
[11] 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.–Mar. 2018.
[12] P. Husák, J. Cech, and J. Matas, "Spotting facial micro-expressions "in the wild"," in *Proc. 22nd Comput. Vis. Winter Workshop*, 2017, pp. 1–9.
[13] X. Ben *et al.*, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3067464.
[14] 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.*, 2013, pp. 1–6.
[15] 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.*, 2019, pp. 1–5.
[16] Y. Zhao and J. Xu, "A convolutional neural network for compound micro-expression recognition," *Sensors*, vol. 19, no. 24, 2019, Art. no. 5553.
[17] X. Zhang *et al.*, "BP4D-spontaneous: A high-resolution spontaneous 3 d dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.
[18] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3438–3446.
[19] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5117–5126.
[20] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4D facial expression recognition using dynamic geometrical image network," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 24–30.
[21] M. Behzad, N. Vo, X. Li, and G. Zhao, "Automatic 4D facial expression recognition via collaborative cross-domain dynamic image network," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 110.
[22] M. Behzad, X. Li, and G. Zhao, "Disentangling 3 d/4d facial affect recognition with faster multi-view transformer," *IEEE Signal Process. Lett.*, vol. 28, pp. 1913–1917, 2021.
[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK) : A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 94–101.
[24] 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, pp. 1–6.
[25] 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.*, 2011, pp. 51–56.
[26] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci.*, vol. 111, no. 15, pp. E1454–E1462, 2014.
[27] X. Huang, G. Zhao, X. Hong, and W. Zheng, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016.
[28] 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.–dec. 2018.
[29] S. Wang, W. Yan, X. Li, and G. Zhao, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2014, pp. 4678–4683.
[30] W. J. Yan, S. J. Wang, Y. H. Chen, G. Zhao, and X. Fu, "Quantifying micro-expressions with constraint local model and local binary pattern," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 296–305.
[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[32] Y.-J. Liu, J.-K. Zhang, W.-J. 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.–Dec. 2016.
[33] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 137–144.
[34] H. 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, 2012.
[35] D. H. 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.
[36] V. Mayya, R. M. Pai, and M. M. Pai, "Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences," in *Proc. Int. Conf. Adv. Comput., Commun. Inform.*, 2016, pp. 699–703.
[37] Y. Han, B. Li, Y.-K. Lai, and Y.-J. Liu, "CFD: A collaborative feature difference method for spontaneous micro-expression spotting," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 1942–1946.
[38] B. Song *et al.*, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184537–184551, 2019.
[39] 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.
[40] 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.–Jun. 2020.
[41] Q. Zhao, J. Dong, H. Yu, and S. Chen, "Distilling ordinal relation and dark knowledge for facial age estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3108–3121, Jul. 2021.
[42] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–4.
[43] 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*, 2020, pp. 2936–2944.
[44] B. Xia and S. Wang, "Micro-expression recognition enhanced by macro-expression from spatial-temporal domain," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1186–1193.
[45] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 367–374.

[46] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "Local relationship learning with person-specific shape regularization for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11917–11926.

[47] M. Zhang, Z. Huan, and L. Shang, "Micro-expression recognition using micro-variation boosted heat areas," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2020, pp. 531–543.

[48] L. Wang, J. Jia, and N. Mao, "Micro-expression recognition based on 2D -3D CNN," in *Proc. 39th Chin. Control Conf.*, 2020, pp. 3152–3157.

[49] 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, 2021, Art. no. 106566.

[50] M. Bai, "Detection of micro-expression recognition based on spatio-temporal modelling and spatial attention," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 703–707.

[51] M. F. Hashmi *et al.*, "LARNet: Real-time detection of facial micro expression using lossless attention residual network," *Sensors*, vol. 21, no. 4, pp. 1098, 2021.

[52] Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key facial components guided micro-expression recognition based on first and second-order motion," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[53] 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*, 2020, pp. 2237–2245.

[54] H. Xie, L. Lo, H. Shuai, and W. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2871–2880.

[55] L. Lo, H. Xie, H. Shuai, and W. Cheng, "MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 79–84.

[56] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 3094–3098.

[57] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–7.

[58] Y. Li, X. Huang, and G. Zhao, "Micro-expression action unit detection with spatial and channel attention," *Neurocomputing*, vol. 436, pp. 221–231, 2021.

[59] 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. 2019.

[60] A. Kumar, R. Jain, and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1511–1520.

[61] 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 Integration Data Sci.*, 2020, pp. 75–81.

[62] 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.

[63] 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.

[64] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–6.

[65] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2296–2303.

[66] B. J. Matuszewski *et al.*, "Hi4D-A DSIP 3-D dynamic facial articulation database," *Image Vis. Comput.*, vol. 30, no. 10, pp. 713–727, 2012.

[67] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-D audio-visual corpus of affective communication," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 591–598, Oct. 2010.

[68] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[69] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic, "Statistical non-rigid ICP algorithm and its application to 3D face alignment," *Image Vis. Comput.*, vol. 58, pp. 3–12, 2017.

[70] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Int. J. Comput. Vis.*, vol. 126, pp. 233–254, Apr. 2017.

[71] A. Patel and A. P. WilliamSmith, "3D morphable face models revisited.," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1327–1334.

[72] J. Booth and S. Zafeiriou, "Optimal UV spaces for facial morphable model construction," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4672–4676.

[73] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *Proc. IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Hum.*, vol. 40, no. 3, pp. 461–474, May 2010.

[74] M. Behzad, N. Vo, X. Li, and G. Zhao, "Towards reading beyond faces for sparsity-aware 3D/4D affect recognition," *Neurocomputing*, vol. 458, pp. 297–307, 2021.

[75] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-D facial expression recognition by learning geometric deformations," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2443–2457, Dec. 2014.

[76] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image Vis. Comput.*, vol. 30, no. 10, pp. 762–773, 2012.

[77] P. Ekman, "Lie catching and microexpressions," *Philosophy Deception*, vol. 1, no. 2, pp. 118–133, 2009.

[78] W. V. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[79] T. Varanka, "Facial micro-expression recognition with noisy labels," Ph.D. dissertation, Master's thesis, Faculty Inf. Technol. Electr. Eng., Univ. Oulu, Oulu, Finland, 2020.

[80] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul.–Sep. 2017.

[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] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6546–6555.

[83] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, 1994, pp. 582–585.

[84] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, Cybern. Syst., Part C (Appl. Rev.)*, vol. 41, no. 6, pp. 765–781, Nov. 2011.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[86] 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.*, 2019, pp. 1–6.

[87] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[88] K. Soomro, R. Z. Amir, and S. Mubarak, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[89] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3391–3399.

[90] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," *Int. J. Comput. Vis.*, vol. 126, no. 2, pp. 233–254, 2018.

[91] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3 D dense face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 152–168.

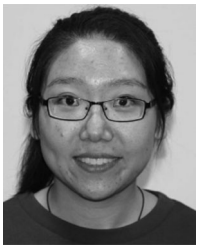
[92] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[94] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, and A. Aspuru-Guzik, "Convolutional networks on graphs for learning molecular fingerprints," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2224–2232.

Xiaobai Li (Member, IEEE) received the BSc degree in psychology from Peking University, the MSc degree in biophysics from the Chinese Academy of Science, and the PhD degree in computer science from the University of Oulu. She is currently an assistant professor with the Center for Machine Vision and Signal Analysis, University of Oulu. Her research of interests includes video analysis for behavioral understanding, remote physiological signal measurement from facial videos, and related applications in affective computing, health care, and biometrics. She has authored and co-authored 47 papers with 3900+ citations and Google H index of 27. She is an associate editor for journal *IEEE Transactions on Circuits and Systems for Video Technology* and *Image and Vision Computing*. She was the chair of several workshops and challenges held on CVPR, ICCV, FG and ACM Multimedia.

Shiyang Cheng (Member, IEEE) received the BS degree from Northeastern University (China), and the MSc and PhD degrees in computer science from Imperial College London. He is currently a researcher with Samsung AI Center, Cambridge. His research interest lies in computer vision and machine learning, especially in the area of automatic 2D and 3D face and body modelling and analysis.



Yante Li (Student Member, IEEE) received the BS degree in communication engineering from the China University of Petroleum (East China), Shandong, China, in 2014, and the master's degree in computer science and engineering from the China University of Petroleum (East China), Shandong, China, in 2017. She is currently working toward the PhD degree with the University of Oulu, Oulu, Finland. Her current research interests include micro-expression analysis and facial action unit detection.

Muzammil Behzad (Student Member, IEEE) received the BS degree with distinctions from COMSATS University Islamabad (CUI), Pakistan, and the MS degree from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, both in electrical engineering. He is currently a PhD researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Currently, he is visiting University College London, U.K.. He is the winner of the prestigious three-minutes PhD thesis competition organized by the IEEE International Conference on Image Processing 2020. His research interests lie around classical signal and image processing, affective computing, and deep learning and its applications.

Jie Shen (Member, IEEE) received the BEng degree in electronic engineering from Zhejiang University, in 2005, the MSc degree in advanced computing, and the PhD degree from Imperial College London, in 2008 and 2014, respectively. He is a research scientist with Meta AI and an honorary research fellow with the Department of Computing, Imperial College London. He was a researcher with Samsung AI Centre, Cambridge from 2018 to 2020. His research interests include facial analysis, affective computing, and social robots.

Stefanos Zafeiriou (Member, IEEE) is currently a professor of machine learning and computer vision with the Department of Computing, Imperial College London, London, U.K. He has more than 17K+ citations to his work, H-index 60. He was an associate editor and the guest editor in more than eight journals. He was the guest editor of more than ten journal special issues and co-organised more than twenty workshops or special sessions on specialised computer vision and machine learning topics in top venues, including CVPR or FG or ICCV or ECCV or NeurIPS. He was the recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011, the President's Medal for Excellence in Research Supervision for 2016, Google Faculty Research Awards, and an Amazon AWS ML Research Award. He is an EPSRC Early Career research fellow.

Maja Pantic (Fellow, IEEE) is currently a professor in affective and behavioural computing with the Department of Computing, Imperial College London, U.K.. She is currently an AI scientific research lead with Meta AI, and she was the research director of Samsung AI Centre, Cambridge, U.K. from 2018 to 2020. She currently serves as an associate editor for both the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *IEEE Transactions on Affective Computing*. She has received various awards for her work on automatic analysis of human behaviour, including the Roger Needham Award 2011. She is a fellow of the U.K.'s Royal Academy of Engineering, and the IAPR.

Guoying Zhao (Fellow, IEEE) (Fellow, IEEE) received the PhD degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an academy professor and full professor (tenured in 2017) with the University of Oulu. She is a member of Finnish Academy of Sciences and Letters, IAPR Fellow and AAIA Fellow. She has authored or coauthored more than 280 papers in journals and conferences with 19120 + citations in Google Scholar and H-index 65. She is panel chair for FG 2023, was co-program chair for ACM International Conference on Multimodal Interaction (ICMI 2021), co-publicity chair for FG2018, and has served as area chairs for several conferences and is associate editor for Pattern Recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, and *Image and Vision Computing Journals*. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.