CAS(ME)^3: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity

Jingting Li, *Member, IEEE,* Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, *Senior Member, IEEE,* Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu, *Member, IEEE*

Abstract—Micro-expression (ME) is a significant non-verbal communication clue that reveals one person's genuine emotional state. The development of micro-expression analysis (MEA) has just gained attention in the last decade. However, the small sample size problem constrains the use of deep learning on MEA. Besides, ME samples distribute in six different databases, leading to database bias. Moreover, the ME database development is complicated. In this article, we introduce a large-scale spontaneous ME database: CAS(ME)^3. The contribution of this article is summarized as follows: (1) CAS(ME)^3 offers around 80 hours of videos with over 8,000,000 frames, including manually labeled 1,109 MEs and 3,490 macro-expressions. Such a large sample size allows effective MEA method validation while avoiding database bias. (2) Inspired by psychological experiments, CAS(ME)^3 provides the depth information as an additional modality unprecedentedly, contributing to multi-modal MEA. (3) For the first time, CAS(ME)^3 elicits ME with high ecological validity using the mock crime paradigm, along with physiological and voice signals, contributing to practical MEA. (4) Besides, CAS(ME)^3 provides 1,508 unlabeled videos with more than 4,000,000 frames, i.e., a data platform for unsupervised MEA methods. (5) Finally, we demonstrate the effectiveness of depth information by the proposed depth flow algorithm and RGB-D information.

Index Terms—Micro-Expression, Micro-Expression Databases, CASME, Depth Information, Ecological Validity, Multi-Modality.

1 Introduction

As a proverb says, you may know a person's face but not his mind. Therefore, it is challenging for a person to have an insight into other individual's states of mind. While in the field of computer vision, with the advent of deep learning technology, human parsing and face recognition have been significantly developed [1], [2], [3], [4], and its accuracy rate has reached beyond human capabilities. Furthermore, it has been widely used in complex practical applications, such as face unlocking of smartphones, face recognition access control, etc. Besides person identification, the research on face-based mind understanding is emerging for decades and is highly challenging. For instance, facial shapes could reveal the personality [5]; facial expression (FE) and color could reveal human emotion independently [6]. Personality and emotions are essential manifestations of human minds and play a crucial role in human understanding and human-computer interaction. Moreover, the research topics on emotional states that involve other complex minds, such as deception [7], depression diagnosis [8], etc., have also seen significant interest and progress recently.

In his book *The Expression of the Emotions in Man and Animals* [9], Charles Darwin revealed that the expression of human emotions is difficult to be suppressed. When an individual fails to suppress his or her expression, there will be involuntary expressions. As shown in Fig. 1, neuropsychological research has found that voluntary and involuntary expressions are controlled by two different neural pathways, i.e., the pyramidal tract and extrapyramidal tract, respectively [10], [11]. The confrontation between voluntary and involuntary expressions could produces micro-expressions (MEs) [12]. ME may be leaked due to voluntary inhibition before expressing emotions or truncated after common expressions are expressed [10]. Therefore, theoretically, ME is a brief, local and subtle FE that often appears in a high-stakes state [13], with a very short duration, less than 500ms.

The involuntary characteristic of an ME makes it an important external indicator revealing one's genuine emotions and intentions [14]. ME analysis has many applications, particularly in national security [14] and medical care [15]. Among them, deception detection based on ME stands out. Based on the cognitive psychology research, the meta-analysis found that telling lies will be accompanied

- *This paper is supported in part by grants from the National Natural Science Foundation of China (U19B2032, 61772511, 62106256, 62061136001), in part by grants from the China Postdoctoral Science Foundation (2020M680738), and in part by the National Key Research and Development Project (2018AAA0100205). (Su-Jing Wang and Xiaolan Fu are corresponding authors.)*
- *J.T Li and Z.Z Dong are with Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China.*
- *S.J Wang, S.Y Lu and C.B Huang are with the Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China, and also with the Department of Psychology, University of the Chinese Academy of Sciences, Beijing, 100049, China. E-mail: wangsujing@psych.ac.cn*
- *W.J Yan is with School of Mental Health, Wenzhou Medical University, Wenzhou, Zhejiang, 325035, China.*
- *Y.H Ma is with School of Computer science, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212100, China.*
- *Y Liu and X.L Fu are with the State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China, and also with the Department of Psychology, University of the Chinese Academy of Sciences, Beijing, 100049, China.*

by some emotional and uncontrollable reactions [16]. And when people try to hide their genuine emotions, ME may occur. Under the experimental conditions, these MEs have a stable correlation with deceptive behaviors [17].

In order to apply ME in practical applications, Ekman et al. developed A tool named Micro-Expression Training Tool (METT) [18] to train people on the detection of MEs. However, even after the METT training, the ME detection accuracy for human being remains less than random level [19]. Moreover, the accuracy of the naked-eye ME detection would be influenced by the emotional context [20]. Therefore, objective and efficient automatic ME Analysis (MEA) is required for further ME practical applications. As illustrated in Fig. 2, MEA consists of ME spotting (MES) and ME recognition (MER), respectively. MES is to detect clips when MEs occur on facial videos, i.e., to determine whether videos contain MEs, and if so, to locate the onset and offset frames of ME clips. In contrast to face detection, which is to determine a suitable closed rectangular region containing the face image on two-dimensional (2D) plane, MES is to determine a one-dimensional closed interval containing ME as appropriate as possible on the one-dimensional timeline of videos. Meantime, MER refers to classifying a given ME clip into a psychologically specified emotional category. Research on the MEA has developed since the turn of the century. Fig. 3 shows the trend of the number of MEA research articles. This number is low but is increasing. Especially in recent two years, deep learning methods are blooming in the MEA field. However, the performance has not been greatly improved because of the limitation of the small sample size (SSS) problem of ME.

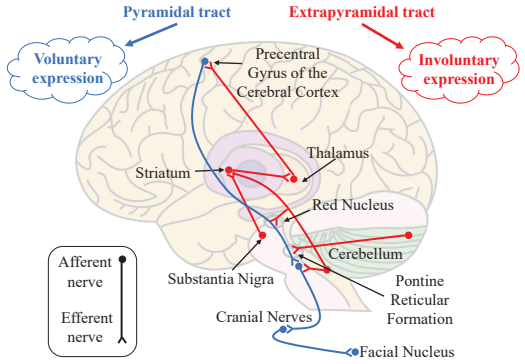


Fig. 1. Neural tracts for FE. Voluntary and involuntary expressions are controlled by the pyramidal tract (blue trajectory) and extrapyramidal tract (red trajectory), respectively. In particular, both the pyramidal system and the extrapyramidal downward transmission begin in the precentral gyrus of the cerebral cortex. In the pyramidal tract, the cortical nucleus tract (blue trajectory) is a neural pathway that controls facial muscles and bones and is responsible for voluntary expression. Meantime, for the extrapyramidal tract, the complete pathway (red trajectory) travels from the cerebral cortex through the brainstem to areas such as the red nucleus and substantia nigra, then through cranial nerves and down to the facial nucleus, responsible for unconscious expressions.

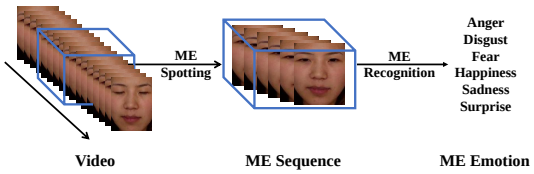


Fig. 2. ME analysis (MEA) process, including MES and MER, respectively locating the moment when the ME occurs in the video and classifying the ME video clip.

Databases are vital for the research of artificial intelligence in various fields [21], [22], [23], [24]. Similarly, for the MEA automation, ME samples are the basis of the research. Current commonly used ME databases include CASME series: CASME [25], CASME II [26], CAS(ME)^2 [27], SMIC [28], SAMM [29] and MMEW [30]. The total sample size in these databases is small, limiting the development of deep learning in MEA. Furthermore, these samples were distributed across six different databases, leading to database biases [31], including specific preferences during the construction process and inconsistent perceptions on the expression categories of annotators in each dataset. Even though differences in data collection can facilitate generalization of the feature learning process, they still impact studying MEs. Meanwhile, the CASME series, although large in scale, has less domain variation in the acquisition environment, source of human subjects (demographics/age), etc. Regarding the annotation, the sentiment categories could be uniformly labeled by specific rules [32], which can alleviate the problem of data bias and improve the performance of expression classification up to a certain point. However, still, the existence of data bias prevented this magnitude of sample size from being maximized to its fullest extent.

However, creating a ME database is a particularly challenging task, facing three major difficulties on ME elicitation, collection, and annotation.

1) **ME elicitation** is the process to elicit the emotion of the subject and leads to a leak of ME. Compared with FE databases in which most samples are posed expressions, it is more complex to induce effective MEs. Since ME appears when a person wants to hide his/her true emotion, ME generation is a spontaneous process. Therefore, collecting spontaneous ME samples is a more appropriate way to conduct studies close to real scenario applications. The most common method is the neutralization paradigm, i.e., asking subjects to watch strong emotional stimuli (see subsection 2.1 for explanation) and try to neutralize the face the whole time or suppress their FEs when they realized there is one. Therefore, the elicitation must be a professionally designed psychological process, and the requirement of FE suppression makes MEs extremely rare during the recording process. Besides, as chances of that ME appears on a neutral

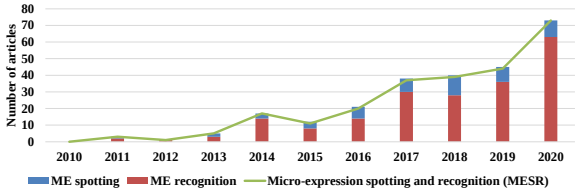


Fig. 3. MEA research trend. The number of articles on MEA is increasing yearly, mainly in the area of MER (bottom column). MES research has not yet attracted sufficient attention (top column).

face are not high in practical applications, the neutralization paradigm directly causes a low ecological validity (see subsection 2.1 for details) for ME samples, which still has a big gap between the real-life MEA.

2) **ME collection** is a complicated process due to the inherent characteristic of MEs. Unlike the distinctive facial movement of FEs, MEs and neutral faces are almost indistinguishable by a single image. Hence, MEA needs to be performed on videos, i.e., importing the modality of temporal information. Meanwhile, since ME is a subtle facial movement, environmental noise, such as illumination variations, would affect the MEA performance. Thus, current MEA methods still require a large amount of data collection in a variable-controlled environment further to enhance the ME feature extraction capability without interference.

3) **ME annotation** is a very laborious and time-consuming process. As ME is a brief, subtle local facial movement, it is not as evident as FE. It is hard to be detected in the video with naked human eyes. The coder, i.e., the person who labels the data, needs to be professionally trained. And, it usually takes half an hour to mark a one-minute video. Meantime, there are two kinds of ME labels: Action Unit (AU) and emotion class. The relation between AU and emotion for ME is still ambiguous, requiring further exploration. This uncertainty leads to that the current sample classification criteria are inconsistent. Besides, due to the eliciting paradigm, the subject may conceal the true emotion by covering it with blinking, smiling, or other facial movements, i.e., ME might be masked. It also increases the complexity of the annotation process.

These three difficulties cause two significant challenges for ME databases: the SSS problem and low ecological validity. To address these two challenges, we make the following contributions in this article. First, inspired by human visual perception, we introduce depth information modality into the ME database for the first time. We also designed the psychological experiment to demonstrate that the depth information is helpful for human visual perception to ME. Second, we recruited 216 subjects, recording 1,300 labeled and 1,508 unlabeled long videos. CAS(ME)^3 contains not only RGB images but also depth information, physiological and voice signals, i.e., CAS(ME)^3 is a multi-modal ME database. The large volume of data provides a platform for multi-modal self-supervised learning based MEA. Moreover, CAS(ME)^3 has a database volume comparable to the total number of currently available databases. Meanwhile, since samples are collected in the same environment and labeled by the same labelers, CAS(ME)^3 allows MEA methods to be validated on larger scale data while effectively avoiding database bias. Last but not least, we collected ME samples through the mock crime paradigm, building a high ecological validity ME database.

In summary, we released a third-generation spontaneous ME database: CAS(ME)^3 [1] . The cubic notation represents the introducing of depth information modality to MEA and the third generation ME eliciting paradigm (see detailed explanation in subsection 2.2) to construct a high ecological validity database.

1. To download CAS(ME)^3, please visit http://casme.psych.ac.cn.

The inability to effectively learn the subtle, brief, and localized ME feature is the pain point the research community is facing. Researchers could benefit from our database to advance deep learning-based multi-modal MEA methods and improve the performance and robustness of the network. In addition, the high ecological validity of multi-modal data allows further research on the implementation of MEA in practical applications and physiologically-based mechanisms of ME.

The article is organized as follows: Section 2 introduces the related works on ME databases and current automatic MEA methods; Section 3 investigates the influence of depth information on human visual perception through a psychological experiment; Section 4 describes the detailed information of CAS(ME)^3; Section 5 presents the benchmark methods and experimental results for MES and MER, database comparison, and multi-model analysis on part C respectively; Section 6 concludes the paper.

2 Related Works

This section first introduces the psychological terminology explanation to understand our latter psychological analysis and experiments better. Then, the published ME databases and the related works on MEA methods are presented.

2.1 Terminology Explanation

We conducted psychological studies in this article to demonstrate the effect of depth information on human visual perception and the feasibility of mock crimes on ME elicitation. This subsection explains some psychological terminologies for better readability.

Ecological validity: The functional and predictive relationship between people's performance on a set of experiment and their behavior in a variety of real-world settings [33]. In the field of computer vision, the ecological validity of the data reflects the closeness to the real data in the practical scenario.

Stimuli: The materials that subjects need to cognitively process and respond to, generally including text, sound, pictures, and video, etc. [34].

Reaction time: The time taken for the subject to complete a task, i.e., from the start of playing the stimuli until the subject responds to it. [35].

2.2 ME Databases

Ecological validity is a crucial factor in determining whether the ME database is close to or suitable for MEA in real complex scenarios. Moreover, it is directly related to the paradigm of MEs elicitation. Thus, we divided the published databases into three generations in terms of ME elicitation methods with gradually increasing ecological validity.

1) **The first generation**: In the early stages of MEA, ME samples were collected from actors who tried to pose fleeting FEs after observing standard expression samples. USF-HD [36] and Polikovsky's database [37] are two posed ME databases. However, ME is considered spontaneous and difficult to be disguised. Besides, spontaneous and posed samples have different elicitation method and theoretical neural basis. The spontaneous ME samples are collected

through the elicitation paradigm that based on the ME generation mechanism [12]. Meanwhile, the emotion is elicited in a state, including three complete emotional dimensions: the subjective experience, external expression, and physiological arousal. The process is associated with brain regions that control emotions, such as the frontal cortex. In contrast, since subjects are asked to mimic the specific expression action, posed expressions include only the external performance dimension and are controlled by the brain's motor cortex. In daily life, real expressions are an effective means for us to understand others' emotions/feelings. Thus, inducing spontaneous ME samples is necessary to improve the database's ecological validity and develop MEA studies with robustness.

2) **The second generation**: In response to the above concerns, researchers with psychological backgrounds tried to elicit MEs with emotional stimuli. It was found that watching emotional stimuli while neutralizing faces, i.e., neutralization paradigm is an effective method to elicit spontaneous MEs without many irrelevant facial movements. The MEs in databases such as CASME series [25], [26], [27], SMIC [28], SAMM [29] and MMEW [30] were collected in this way. Such ME samples have several merits, the most important of which is that they are spontaneous. Moreover, they are well-controlled in illumination, environment, and neutral faces without talking. The majority of current MEA are conducted on these samples, especially the CASME series databases, as the elicitation process was designed by professional psychologists. The CASME series has been requested by more than 600 research teams in over 50 countries, and more than 80% of the MEA articles have used at least one of these databases for method validation.
However, this eliciting paradigm still has shortcomings. The MEs in 2nd generation databases were collected in lab situations (though spontaneous) but are supposed to be different in real life. They are actually not "natural" enough if we would like to put MES and recognition in practical fields. First, the subjects were required to neutralize their faces while watching stimuli. There should be various methods to conceal FEs, e.g., masking with smiles. Second, these lab situations elicit the emotional response by presenting stimuli but not interpersonal communication. However, in real life, FEs are a social signal that are seen more often in interpersonal interactions rather than occurring alone. It is in such situations that we need to hide our genuine FEs occasionally. Third, watching emotional episodes is actually not a high-stakes situation, though the subjects may have motivations to conceal their FEs with reward/punishment configurations. The feelings were supposed to be different from lying to a person in a high-stakes situation.

3) **The third generation**: Addressing above constraints, collecting MEs samples from more ecological situations is inevitable for the further development of MEA research. In psychology, there are already well-established paradigms. For instance, eliciting paradigms through mock crime, dictator games, and prisoner's dilemma could elicit ME with high ecological validity. Husak et al. published an in-the-wild database MEVIEW [38], in which the samples are video clips from poker games and TV interviews downloaded from the Internet, with a total of 40 labeled MEs. Although these samples are from real scenarios and the ecological validity is high, there are too many uncontrollable factors, such as zooming in and out, head movement, hand-over-face occlusions, etc. Based on the current state of MEA research, ME samples still need to be collected in well-controlled laboratory scenarios. Mock crime is a high-risk stimulus, in contrast to neutralization paradigm. Therefore, it is more suitable for ME elicitation and further research related to its application in lie detection. Hence, we released a sub-set of ME video samples collected by mock crime paradigm in CAS(ME)^3, improving ecological validity and eliminating uncontrollable factors.

2.3 Automatic ME Analysis

2.3.1 MES

MES methods can be divided into video clip spotting and frame spotting; the latter focuses on locating the apex frame [39]. Nevertheless, even if the final output is different, the algorithms and the related features of these two still have commonalities.

Concerning the algorithm, feature difference (FD) represents the first attempt at MES, and algorithms combined with machine learning (ML) are the current research trend. Since 2014, many research teams [**?**], [26], [32], [40] have used FD to spot spontaneous ME in videos. The main process of this idea is first to calculate the difference between the frame features in a sliding time window; then, determine the most significant movement by setting a threshold for the entire video. However, the ability to distinguish MEs from other facial movements remains weak, especially in long videos containing many other movements and noise.

For this reason, methods combined with ML are gradually becoming the mainstream of MES. For instance, [38] and [41] used the SVM classifier to spot ME frames. Besides these, there are some MES methods that incorporate deep learning, such as MESNet [42], LSSNet [43]. However, fewer than 20 papers using ML for MES have been published, and the sample size limits these algorithms. The numbers of samples in published databases are not large enough to train a high-performing classifier.

2.3.2 MER

Unlike MES, all the MER methods use machine learning for emotion classification, and can be categorized into two main categories: handcrafted feature methods and deep learning methods. LBP-TOP [44], HOG [45], and OF [39] based methods are the most common handcraft feature methods. The MER performance has been greatly improved through continuous exploration of the spatial-temporal features suitable for MEs. Yet, due to the ME characteristics, these methods are not robust enough to realize MER in real scenes.

In recent years, MER combined with deep learning has become a major trend, e.g., STRCN [46], the low-complexity recurrent CNN [47], LEARNet [48], 3D-CNN [49], the joint local and global information learning network [50], etc. In addition, to further address the limitation of the small-sample problem on deep learning-based MER, many approaches introduce transfer learning to enhance the performance of ME feature extraction, such as [51], [52]. The MEA approaches based on deep learning are limited by the SSS problem for three main reasons. First, deep network models

involve a large number of parameters, and the SSS problem of MEs can cause overfitting problems of the model. Second, although there are many transfer-learning-related methods, the improvement of the MEA performance is not particularly significant, and the effect is limited. Finally, compared with the algorithms for expression recognition and face recognition, the number of samples and network parameters for MEA are limited by the ME sample size.

As noted above, a novel ME database with a large amount of ME samples is necessary for further development in combination with deep learning. However, capturing and especially labeling MEs is very challenging. Moreover, there have been studies to enhance the performance of model mining features through multi-modality, thus alleviating the limitations of SSS problems [53]. Therefore, while trying our best to capture the RGB modal video, we expand the ME information with more geometric features by capturing data of one more modality, i.e., depth information.

3 Human Visual Perception on ME with Depth Information

The series of cognitive processes that organize and interpret the sensation information of objects or events in the external world is known as perception. It includes visual perception, auditory perception, haptic perception, olfactory perception, taste perception, and so on. In particular, 80% of the external information acquired by humans comes from visual perception [54]. In this section, we investigated the role of depth information on the human visual perception of ME by comparing 2D perception and three-dimensional (3D) perception. The experimental results showed that depth information helped the human recognition of ME.

3.1 Method

The proposed psychological experiment was based on a within-subject design, which means that each subject participates in both 2D perception and 3D perception.

3.1.1 Subjects

To avoid the effects of cognitive ageing [55] on the depth visual perception study, we selected, we recruited thirty-one undergraduate students (9 males and 22 females; Mean (M) = 23.5 years, Standard Deviation (SD) = 1.75 years). (See supplementary file for the study on aged subjects.) They voluntarily participated in this experiment, with payment. And they all had a normal or corrected-to-normal vision and no known psychiatric disorders. Specifically, all studies involving human subjects in this article adhered to the Declaration of Helsinki and were approved by the institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences.

3.1.2 Stimuli

After weighing the emotional sample distribution, resolution and sample form (RGB or grayscale) of different databases, 30 ME samples were chosen from the CASME database [25], including six emotion types. In order to eliminate the influence of background information, only the cropped face area was retained in the 2D video stimulation. We converted 2D video stimulation to 3D video stimulation. Finally, 60 2D and 3D ME video clips were produced as the stimuli. They were presented in a 2.3 $\times$ 1.4 square meters (m^2) screen, placed 3.5 m in front of the subject, by a BenQ TH6370 projector (support 3D mode). Subjects were asked to wear BenQ 3D Active Glasses in the experiment, with 3D mode on for 3D ME videos and off for 2D videos. Before the experiment began, subjects were seated in a lab with sound insulation. The light intensity of the environment was 0.26 Lux and 288.73 Lux when the projector was turned off and on, respectively. Fig. 4 illustrates apparatus and lab environment.

3.1.3 Procedure

Subjects were required to watch 2D and 3D ME video stimuli and respond to related questions. The procedure for both tasks was the same. To exclude practice effects [34], the order of the two tasks was counterbalanced across subjects.

Before starting the formal experiment, the subjects should first complete three practice trials to get familiarized with the procedure. The three stimuli for the practice were also selected from the CASME database and were not repeated in the formal experiment. Therefore, data from practice were not included in the result analysis. Furthermore, with the random order of two tasks and the practice trials, we can eliminate the effect of subjects' curiosity or familiarity with the videos on the experiment.

In this experiment, when the subjects pressed the space bar, they were presented with a ME video clip, played only once. Then, the subjects were asked to answer three questions about emotional valence, emotional type, and emotional intensity quickly and accurately. Fig. 5 presents the experimental procedure in detail.

3.2 Result

The experimental results were analyzed using one-way Analysis of Variance (ANOVA) statistics analysis [56], as listed in Table 1. We calculated the mean and standard deviation of the subjects in 2D and 3D, respectively, and compared the differences in reaction time (RT) of emotion recognition and intensity ratings (see Fig. 6). RT was significantly shorter in 3D than in 2D conditions for both evaluations of emotion valence and emotion type. The video with 3D ME was also rated with higher intensity. Our results indicated that subjects might benefit from 3D MEs, i.e., ME videos with depth information for MER.

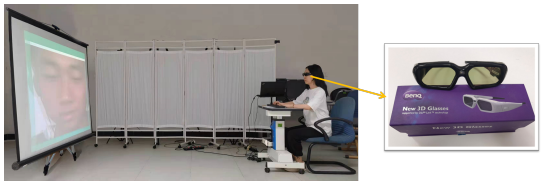


Fig. 4. Lab configuration. The subject wore 3D Active Glasses to watch 3D ME videos. The subject sat at a chair and adjusted the table to the appropriate height. A BenQ TH6370 projector was located under the table and projected to a screen that was placed 3.5 meters away from the subject. The subject used a Bluetooth keyboard to respond. The control computer was placed at the ride side of the subject.

TABLE 1
The comparison result for 2D and 3D artificial MER performance analysis. In the context of the same number of subjects, for ANOVA analysis, if p is less than 0.05, there is a significant difference between the two cases; the larger the $F(1,30)$, the more reliable the experimental results are likely to be; η_p^2 corroborates that the experimental results do not occur by chance. (See theoretical explanations of $F(1,30)$, p and η_p^2 in [57])

	Intensity					Reaction time for emotion valence					Reaction time for emotion type				
	M	SD	$F(1,30)$	p	η_p^2	M	SD	$F(1,30)$	p	η_p^2	M	SD	$F(1,30)$	p	η_p^2
2D	3.48	1.65	4.35	<0.05	0.01	2.15	1.46	15.46	<0.001	0.05	2.32	2.49	7.93	<0.01	0.75
3D	**3.64**	**1.58**				**1.83**	**1.26**				**1.91**	**1.60**			

3.3 Discussion

This study investigated the influence of human visual perception on MER by appending depth information. The results show that depth information could facilitate emotion recognition, indicated by shorter RT and a higher intensity rating for 3D videos. This may be explained by that 3D videos provide additional specific information to the cognitive process of 2D and 3D information [58]. Specifically, in eye-tracking studies, there is a significant difference in attention distribution of gazing on faces between 2D and 3D visualization [59]Moreover, electrophysiological studies have also demonstrated that viewing 3D stimuli increased the activation degree of the specific stream in the visual system, whose function is depth information perception [60]. Functional magnetic resonance imaging (fMRI) studies have also proven that the human parietal cortex that controls the above-mentioned neural stream plays an essential role in processing depth information [61]. Depth information can help people build a robust facial characterization [59], so that the face in stereoscopic vision is closer to the human face in reality. Hence, depth information makes facial features easier to be recognized.

As demonstrated above, depth information helps to enhance the human perception of ME. Inspired by this, we will introduce depth information into MEA.

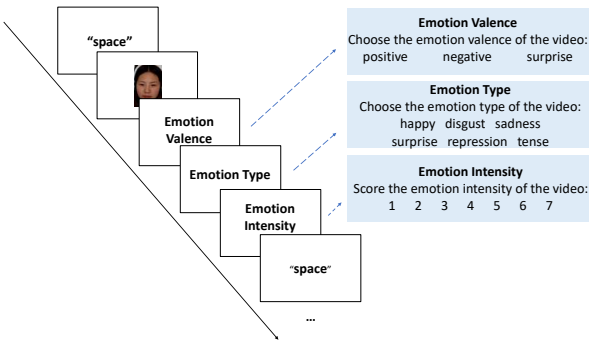


Fig. 5. Experimental procedure. First, the subjects were required to press the "space" key to start a ME video clip, which was played only once. Then the subjects answered three questions including emotional valence, emotional type, and emotional intensity according to their own feelings.The practice phase consisted of 3 trials, and the formal experiment consisted of 30 trials.

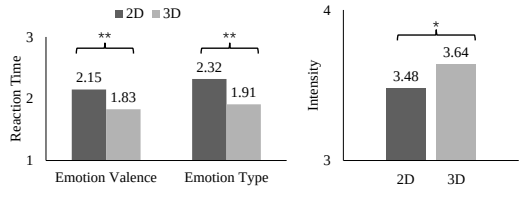


Fig. 6. Schematic diagram of results. The figure shows the mean values of emotion valence, emotion type, and emotion intensity under 2D and 3D viewing conditions. * means $p<0.05$, ** means $p<0.01$, indicating significant difference.

4 CAS(ME)^3 Database

The results of the psychological experiment in Section 3 demonstrate that depth information improves human ME cognitive behavior. Moreover, with the popularity of depth cameras, research methods based on depth information will be widely used and become an inevitable research trend in computer vision. Depth has proven to be very useful in areas such as face recognition [62], expression recognition [63], etc. RGB-D images can be used to construct a 3D model of a human face. The human face is a continuous smooth surface in all directions. Thus, these points, which have the same depth value, form a continuous curve. The curve is called the *depth contour*. The geometric change of depth contours on the facial surface could be caused by a muscular action which means the face undergoes deformation and the distance from the camera changes. As demonstrated by the depth contours in Fig. 7, the depth information changes visibly when an expression occurs on the face. (See supplementary material for a video demo showing the depth contour variation for a ME video clip.) Therefore, introducing depth information can help the MEA algorithms to detect changes in human faces more sensitively.

Based on the above research, we construct a ME database with depth information: CAS(ME)^3, consisting of parts: Part A, Part B, and Part C. The male/female ratio in the CAS(ME)^3 database is 112/135, including all three parts. The mean age of the subjects is 22.74 and the SD is 1.75. First, for the research continuity of MEA in the databases based on a second-generation elicitation paradigm, we collected ME

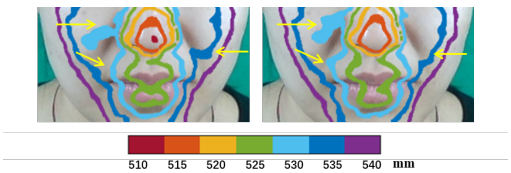


Fig. 7. Depth contours on the face during a facial ME. On the subject's face, from onset to apex frames, an AU 14 (Dimpler) occurs at the corner of the mouth. In both lower face frames (onset and apex) with depth contour, the same colors have the same depth values, varying from 510mm to 545mm. In the corners of the mouth and cheek regions, the contour lines have different shapes in the two frames (as indicated by yellow arrows). This difference confirms that the FE of the face can be reflected in the depth information.

samples in Part A and B using the same paradigm. There are 100 subjects in Part A, and each of them was asked to watch 13 emotionally stimuli and keep their faces expressionless, and each viewing was recorded. Hence, there is a total of 1,300 video clips. In these clips. 943 MEs and 3,143 MaEs are labeled by professional coders. In Part B, there are 116 subjects and 1,508 unlabeled video clips. We hope that labeled clips in Part A and unlabeled clips in Part B provide a data platform for developing ME unsupervised or self-supervised learning. In Part C, we used the third generation of ME eliciting paradigm, i.e., mock crime, to elicit ME with higher ecological validity and finally succeeded in capturing 166 MEs and 347 MaEs from 31 subjects. Compared to the high ecological validity database MEVIEW, we doubled the number of ME samples and also labeled MaEs. In addition, during the recording process, we also collect ME-related physiological signals such as heart rate and voice signals to enrich multi-modal MEA. Table 2 lists the overview of CAS(ME)[3], around 80 hours of videos in total, i.e., about eight million frames, including 1,109 MEs and 3,490 MaEs.

4.1 ME Collection Settings

As ME is a brief, subtle facial expression, the ME samples are collected in a strictly controlled environment to concentrate on MEA and avoid other disturbing factors.

The laboratory settings are shown in Fig. 8. We used four 24W LED lights. The LED lights are equipped with ballast to prevent AC power from causing the LED lights to flicker, affecting the light's stability. We use a set of 2 LED lights with a reflector umbrella to focus the light on the subject's face. Furthermore, the reflector umbrella can provide more stable and soft lighting. This lighting preparation setting can greatly avoid the influence of light strobe on the shooting during the recording process. We used an Intel® RealSense™ D415 camera to record the subjects' facial movements, with a resolution of 1280×720 pixels. We record RGB color images and the corresponding depth information simultaneously during the process of inducing ME. Due to the limitation of the device, the frame rate is 30 frames rate par second (fps). However, at 30fps, the frame rate interval of ME is in the range of 6-15 frames, which can already retain the ME temporal variation. There is a green curtain about 1.5 meters behind the subject. It provides the discriminant depth information between the background (the curtain) and the foreground (the subject). It also provides the discriminant green channel information between skin color and non-skin color [64]. The information only makes it convenient to crop facial area.

4.2 Part A: ME Data with Label

In the published databases, the introduction of deep learning did not substantially improve the MEA performance. To verify the effectiveness of MEA combined with deep learning, we collected a relatively large number of video samples in the first two parts of the database based on the second generation of ME eliciting paradigm compared to the previous databases.

4.2.1 Eliciting Paradigm

Eliciting materials: We used 13 emotional video clips, which were evaluated and selected as the stimuli to elicit MEs in CASME series databases [25], [26], [27]. Each clip belongs to a type of basic emotions[2], either disgust, fear, sadness, anger, or happiness. The number of videos for each emotion is 2, 4, 3, 2, and 2, respectively. Although there was no stimulus specifically of the emotion type surprise, some scenarios in the stimulus videos could induce the emotion of surprise in the subjects. The duration of the videos ranges from 34 to 144 seconds.

Procedure: Each subject entered a configured lab and was asked to be seated in front of a monitor where we would later present the emotional stimulus video. The subject was instructed that whenever they were aware that an expression is about to be expressed while watching videos, they should immediately suppress it and try to keep a neutral face. They were also required to keep their body and head still during the same time. The Intel® RealSense™ D415 camera has an important parameter, namely the minimum operating distance (minZ), less than which the scene depth information cannot be processed. To maximize the proportion of the face in the frame during video recording at a distance greater than minZ from the camera, and to avoid noise caused by bad exposure, we performed a head-position calibration before presenting each video. The subject took a brief break after watching a video. After watching each video, the subject was asked to give a subjective report on the video they just watched. The self-report method requires subjects to assess their emotional experiences on a rating scale, reflecting the true emotions within the subject and quantifying the emotions. In contrast, MEA investigates true emotions through facial expressions manifesting by facial muscle movements. However, external expressions and internal emotional experiences are not always perfectly consistent. With the benefit of the self-report, it is possible to make the external MEs correspond to the internal true emotions during the labeling process. (The self-reported questionnaire can be found in the supplementary file).

4.2.2 Annotation

In Part A, 1,300 videos belonging to 100 subjects are labeled by professional coders. The annotation includes Action Units (AUs) and the corresponding onset, apex, and offset frames. First, to improve the efficiency, we built the coding platform. Second, the coding process is introduced to ensure the fairness and transparency of annotation.

The first step of ME annotation process is AU coding frame by frame by coders. So, it is very challenging. One possibility to improve this situation is to involve as many ME or AU coders in manual coding as possible, regardless of geography or time limitation. Therefore, We built a multiuser online coding platform, i.e., ME Coding and Sharing System (MECSS) [3], bringing the possibility of building a large-scale ME database off-site.

The AU annotation is performed based on Facial Action Coding System (FACS) [65]. Two well-trained FACS coders

2. Main emotion means if 70% of the subjects or more chose a specific emotion word, and the average emotional intensity score is greater than 3.5.

3. http://mecss.psych.ac.cn/

TABLE 2
The overview of CAS(ME)^3

Part	Eliciting Paradigm	Number of ME	Number of MaE	Number of Subjects	Number of Videos per Subject	Total Length of Videos per Subject
A	2nd generation - Neutralization	943	3143	100	13	21.27 minutes
B	2nd generation - Neutralization	N/A	N/A	116	13	21.27 minutes
C	3rd generation - Mock crime	166	347	31	1	about 8 minutes
Total		1,109	3,490	247	Total video length	about 80.7 hours

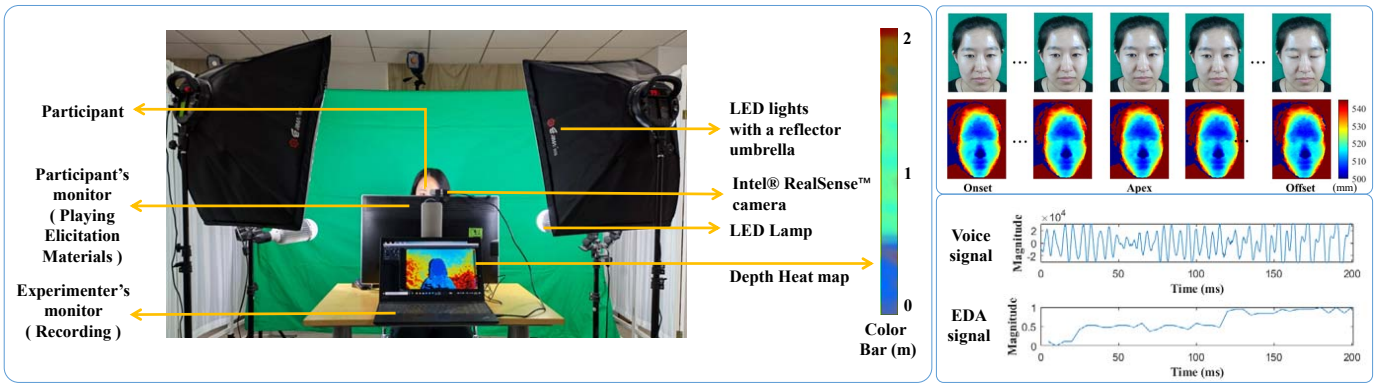


Fig. 8. The left block shows the recording environment for CAS(ME)^3 database. The upper right block illustrates the samples of RGB image and depth map (Subject spNO.216 in Part A: Surprise with AU R1+R2); and the lower right block displays the samples of voice and electrodermal activity (EDA) signals (Subject MC_11 in Part C: negative emotion with AU4).

coded the collected video samples frame by frame to determine the existence and duration of different AUs through MECSS. Once it is confirmed that an AU has occurred, its onset, apex, and offset also need to be determined.

According to the temporal characteristic of ME (<500ms), the FEs are first divided into two types: MaEs and MEs. Based on the obtained AU annotations, we need to determine the starting and ending moments (onset and offset) of their corresponding expressions to perform this classification. Suppose there are K time-overlapping AUs occurring in a time period, and for the kth of them, its start and end time is ($t_1^{AU_k}$,$t_2^{AU_k}$), where $k \in (1, .., K)$. Then the onset and offset frames of the FEs represented by these AU combinations are obtained by the minimal and maximal values of the AU moments, respectively, as shown in the following formulas.

$$\begin{aligned} \text{onset} &= \min(t_1^{AU_1}, ..., t_1^{AU_K}) \\ \text{offset} &= \max(t_2^{AU_1}, ..., t_2^{AU_K}) \end{aligned} \tag{1}$$

Since the frame rate of videos in CAS(ME)^3 is 30 fps, the number of frames for a 500ms video is 15. Therefore, MEs and MaEs are separated based on the temporal duration:

$$\text{FE type} = \begin{cases} \text{ME} & \text{if } f_{\text{offset}} - f_{\text{onset}} + 1 \leq 15 \\ \text{MaE} & \text{otherwise} \end{cases} \tag{2}$$

where f_t denotes the frame index of the moment t.

After identifying the onset, apex, and offset of ME, we further classify ME video clips in order to analyze different kinds of MEs with more specificity. To label the ME video clip with emotion, the coder compared three emotion types respectively based on the AU label, elicitation material, and subject's self-report of this video. The correspondence between AU and emotion was referenced to the annotation method in [26]. If at least two emotion types were consistent, then the emotion type of the current ME video clip was identified as the one that accounted for the majority. If none of the three were consistent, then the emotion type of the ME video clip was determined by the coder's judgment based on AU. Furthermore, some habitual behaviors should be eliminated, such as frown when blinking or sniffing.

To be consistent with the emotion classification of macro-expressions, we provide emotion labels based on six emotion classifications for ME samples. Fig. 9 lists the number of ME samples in each emotion class. The additional category of "Others" indicates MEs that have ambiguous emotional meanings or that are difficult to be classified into the six basic emotions.

However, there is more practical importance in classifying MEs according to four emotional categories (positive, negative, surprise, and other). Discovering negative emotions hidden under positive expressions or vice versa, such as covering the dagger with a smile, can help in lie recognition or emotional understanding of interpersonal interactions. Besides, this kind of classification is the current popular classification method, avoiding the problem of unbalanced samples distribution to a certain extent. Specifically, in the CAS(ME)^3, the criteria based on the above-mentioned four emotion classifications are as follows. Positive expression includes happy expressions, which are relatively easy to induce and have obvious characteristics. Negative expressions include disgust, sadness, fear, anger, etc. These MEs are relatively difficult to distinguish, but they are significantly different from positive MEs. Meanwhile, surprise has no direct relationship with positive or negative expressions, and expresses unexpected emotions, which can be interpreted according to the context. The category of "Others" has the same meaning as it in emotion classifi-

cation based on six basic emotions.

The number of ME samples for Part A and Part C are 943 and 166, respectively, as listed in Table 2. The coding reliability (R) is calculated as follows:

$$R = \frac{N(C_1 \cap C_2)}{N_{all}} \tag{3}$$

where the denominator and numerator represent the number of all MEs and the number of MEs coded consistently by the two coders, respectively. Regarding the four-emotions classification, R for Part A and Part C is 0.88 and 0.94, respectively. And for the seven-emotions classification, the coding reliability for Part A and Part C is 0.71 and 0.75, respectively.

The two coders did not participate in the ME sample video acquisition process and did not know the stimulus material corresponding to the labeled videos.

As supplementary annotation information, we also provide the corresponding stimulus's emotion and the feelings indicated in the self-report. This combination of emotional information from the subject and the eliciting environment will help researchers carry out contextual FE research and would be comprehensive support for interdisciplinary analysis of computer vision and psychology.

4.3 Part B: ME Data without Label

We recruited a total of 216 subjects to induce and collect their MEs under the same experimental configuration. Except for the 100 subjects in Part A, we found that the videos of the remaining 116 subjects showed relatively severe frame drop. We used the Intel® RealSense™ D415 camera to capture both depth and RGB information, experiencing frame drop is a known likely effect to be encountered [66]. Due to the serious frame drop during the recording process, accurate manual ME annotation cannot be performed, and hence these unlabeled videos comprise Part B. The male/female ratio in part B is 53/63. In sum, 1508 unlabeled long videos were obtained, amounting to 41 hours and 58 million frames.

One of the solutions to alleviate the ME SSS problem is unsupervised learning. As a form of unsupervised learning, self-supervised learning has become a hot topic. Self-supervised learning refers to a learning method that uses automatically generated labels to train the network explicitly [67]. The main problem of self-supervised learning is how to design pretext tasks, that is, the automatic generation of labels. It is a natural idea to use the additional depth information or more ME related frames to generate labels and construct self-supervised learning models for MEA.

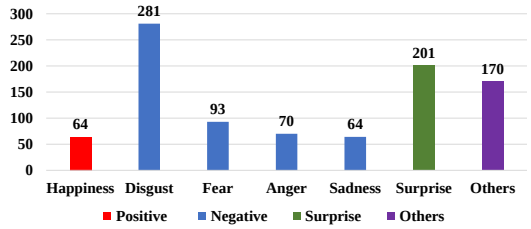


Fig. 9. Labeled ME samples distribution in Part A.

Even though the data in Part B has severe frame drop, this situation does not affect the construction of the self-supervised learning algorithm model on these data. First, the quality of the single-frame images in the recorded videos is not affected, and no information is dropped in the face region. Therefore, multiple pretext tasks for spatial facial feature learning can be designed based on part B. Second, such videos with dropped frames can be implemented by an unsupervised method similar to the one proposed in [68] for temporal interpolation, and the construction of a temporal feature extraction model is also achieved in this way. Furthermore, since part A and B have the same environment settings, a more optimal pre-trained model can be constructed for boosting the MEA performance of part A. In addition, to facilitate ME analysis based on unlabeled data, we provide the emotion type of the stimulus material corresponding to the unlabeled video, the number of which is listed in Tabel 3. Furthermore, we provide subjects' self-reports after watching the stimulus material, the number of emotions of which is also listed in Table 3. Both the elicitation protocol of micro-expression and the cultural background of the subjects can have an impact on the resulting emotional perception. Therefore, the number of stimuli and the number of subjects' self-reported emotion types do not necessarily correspond. In addition, the same eliciting stimulus has different effects on different subjects. For example, surprise is often associated with the perception of "unexpectedness ", so it can be elicited in many videos. Meanwhile, contempt is strongly related to the individual's personality, making it difficult to be elicited accurately. Thus, it is challenging to select video stimuli with these two kinds of emotions. These unlabeled videos and the labeled videos in Part A together provide a data platform for the development of self-supervised learning methods of MEA.

Real-life video samples have many uncontrollable factors, and it is not easy to measure whether the self-supervised algorithm is helpful for MEs in some specific scenarios. However, the platform for self-supervised learning based on all samples in Part A and B in CAS(ME)^3 extends the possibilities of applying MEA in real-life applications. Meantime, MEs are important cues for lie detection. Hence in Part C, we used the mock crime paradigm to induce ME samples with higher ecological validity.

4.4 Part C: ME Data with High Ecological Validity

According to the *Audience Effect for human smiling* [69], people are more likely to produce more expressions in socially interactive situations than in solitary ones. Psychological experimental paradigms conducted in interactive, high-stakes situations include mock crime, dictator games, prisoner's

TABLE 3
Quantitative report of the reference to the emotional variety in Part B. $\#_{stimuli}$ denotes the number of videos for different emotion types of stimuli; $\#_{self}$ denotes the number of emotions reported by subjects after watching the stimuli. H, D, F, A, Sa, Su and C denote Happiness, Disgust, Fear, Anger, Sadness, Surprise and Contempt, respectively.

Emotion	H	D	F	A	Sa	Su	C
$\#_{stimuli}$	232	232	464	232	348	NaN	NaN
$\#_{self}$	136	388	263	278	315	260	78

dilemma, etc. The mock-crime paradigm is considered the "gold standard" for lie detection in laboratory research [70].

The mock-crime experimental paradigm divided the subjects into a crime group and an innocent group. The experiment was organized into two phases: mock crime and interrogation. In the mock crime phase, subjects were asked to choose autonomously whether to enroll in the crime group or the innocent group. Subjects in the crime group would perform a mock theft crime. During the interrogation phase, the researcher would ask the subjects a series of questions directly related or unrelated to the crime information. The crime information was derived from the layout of the mock crime experiment room and the key clues, as shown in Fig. 10a. Subjects in both the criminal and innocent groups were given a choice to lie or not. A final judgment was made about the subjects' actual crime and lying [71].

We used the Concealed Information Test (CIT) [72] during the interrogation phase, interrogating subjects with both non-open-ended and open-ended questions. Compared to open-ended questions, non-open-ended questions were more effective in placing subjects in a high-stakes environment and producing more MEs. Under laboratory conditions, MEs were consistently correlated with deceptive behavior [16]. It was also found that during CIT, liars and non-liars differed in their cognitive activity when answering relevant questions [73]. Liars had different perceptions of criminogenic and criminally irrelevant information. In contrast, non-liars perceived both types of information equally. There are also differences in the responses to neural signals between the two [71]. Therefore, based on the subjects' CIT experimental performance, it is possible to distinguish between liars and non-liars effectively.

In addition, subjects were classified into the high- and low-stakes environments according to whether they chose to commit a crime or not, i.e., the crime group and innocent group. This classification is supported by significant differences in physiological data between the high-stakes and low-stakes environmental groups. The experimental results revealed the differences between MEs collected in these two environments. Specifically, 18 subjects chose the high-stakes environment, with a total of 113 MEs, i.e., 6.3 MEs per capita; meanwhile, 13 subjects chose the low-stakes environment, with 53 MEs, i.e., 4.1 MEs per capita. Thus, there is a trend that the number of MEs collected in the high-stakes group is higher than that of the low-stakes group.

As mentioned in subsection 4.3, multi-modal and multi-channel data could help improve the performance of self-supervised learning. The voice and physiological signals, including electrodermal activity (EDA) [74], heart rate/fingertip pulse (ECG) [75], respiration (RSP) [76], and pulse (PPG), were collected by the BIOPAC MP160 multi-channel physiological instrument and the video recorder, as shown in Fig 10b. And the essential parameters are listed in Table 4. In the experiment, we explicitly informed the subjects through the instructions that we were collecting their physiological signals through the wearable device and then determining whether or not they were lying through the "algorithm" (tricking the subjects, which we did not). So the device strapped to them would lead the subjects to believe that we would indeed use these techniques to detect whether they were lying. In this way, the subjects would produce emotional states (e.g., nervousness, anxiety) similar to real lying scenarios. This configuration improves the ecological validity of ME elicitation. Furthermore, These signals with multi-modality during interrogation phase contain abundant and useful features targeting spatio-temporal variations of MEs. For instance, by analyzing the voice signal, we can extract the corresponding sound quality features, rhythmic features, and spectral correlation features. These characteristics could reflect the person's inner emotions [77]. Besides, Reda et al. [78] combined features from ME videos and the pulse rate variability to recognize emotions.

For MEA, multimodal information can be used as supervisory information for each other. For example, if there are relatively distinct changes in one-dimensional signals such as speech and physiological data, or depth information, the moments of these changes can be used as a kind of annotation information for the corresponding RGB videos. Automated ME annotation could be achieved by designing, for example, contrastive-learning-based self-supervised learning model.

5 Benchmarks

In this section, we give benchmarks on MES and MER of CAS(ME)^3. The effect of depth on MEA is investigated in two different forms. First, we extend optical flow (OF) to depth flow to demonstrate how depth flow complements motion information and improves the performance of MES. Then, by comparing RGB and RGB-D feature learning, we confirm that depth information, as an additional modality, enhances feature extraction targeting to MEs and therefore contributes to MER.

5.1 Pre-processing

We used the facial landmarks for the identification of the face area and key regions belonging to ME. Their geometric

(a) Mock crime room setup. (b) Mock interrogation room setup

Fig. 10. The CIT questions during the interrogation are related to key clues in the red blocks and the room layout in Fig. 10a.

TABLE 4
Multi-modality signals, including physiological signals and voice signal.

Modality			
Physiological signal (BIOPAC MP160 multi-channel physiological instrument)	Channel	Type	Sampling rate
	1	PPG	200HZ
	2	RSP	
	3	ECG	
	4	EDA	
Voice Signal	Sampling rate		48000HZ
	# of sound channel		2
	Coding		16bit PCM

location information is not directly involved in the MES and recognition computation and has a minimal impact on the performance. Therefore, we directly choose the commonly used Dlib [79] to detect landmarks.

Since the MEA research is video-based, the face in all video frames needs to be cropped. For MES, long videos are divided into short clips by a sliding window, and the face cropping in each clip is based on the landmarks of the first frame of the current segment. For MER, the cropping of the whole ME video clip is based on the landmarks of the onset frame. The reason for face cropping based on the first frame instead of based on each frame is that ME is a short, subtle facial movement. If the cropping is performed frame by frame, continuous ME variation information might be lost due to subtle landmarks changes.

Since the depth map and RGB image are acquired simultaneously and have the same resolution and frame rate, the face region in the depth map is cropped by the same landmarks in the corresponding RGB image.

5.2 Benchmark for MES

The OF based method is an effective method for MES. However, the movement of the head can have a drastic effect on the spotting of subtle MEs. After extracting the OF features of the face region, it can be seen that this kind of head motion is manifested as most pixels have a similar OF. Therefore, the effect of such global head movement can be eliminated by subtracting a value of OF fixation from all pixel points. The efficiency of this idea was proven by Zhang et al. [80], and they won first place in the MES task of the ME Grand Challenge (MEGC2020). Besides, Liu et al. [81] demonstrated that face alignment in the OF domain could help improve the MES performance. Hence, in this section, we propose the *depth flow*, which extends the OF from the 2D plane to the 3D space by combing the depth information. Furthermore, the depth flow is applied to the MES framework proposed in [80] to verify the effectiveness of depth information for MES.

5.2.1 Depth tensor

Firstly, we need to transform RGB-D information captured by the Intel® RealSense™ D415 camera to a third-order depth tensor $\mathcal{D} \in \mathbb{R}^{W\times H\times D}$, where W and H represent the width and height of cropped face region and D means the scene depth of the face. The depth tensor $\mathcal{D}$ is constructed as follows:

$$\mathcal{D}(x,y,z) = \begin{cases} \mathbf{I}(x,y) & \text{if } z = \mathbf{D}(x,y) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where, $\mathbf{I}, \mathbf{D} \in \mathbb{R}^{W\times H}$ are the gray image and the corresponding depth information. The gray scale information is utilized since the OF is calculated on the gray image.

As the depth flow is computed between two frames I_1 and I_2, we need two same size tensors $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$. W and H of each frame are the same, D of each frame is different. Furthermore, as illustrated in Fig. 11, inevitably, the area preserved by the face crop may also include some background areas, i.e., the curtain. The distance between the camera and the curtain is about 1,500 mm. This results in that the maximal value of $\mathbf{D}$ might equals to 1,500. Besides, due to the recording environment (lighting conditions, target material, etc.) [82], and hardware effects and internal algorithm errors of the sensor [83], depth data obtained at some pixel points does not meet the confidence metric. Instead of an incorrect value, the Intel® RealSense™ D415 camera provides a zero value at these points [84]. Thus, $\mathbf{D}(x,y) \in (0, 1500)$. According to Eq. 4, we can infer that the number of zero elements in $\mathcal{D}$ is about 1500 times that of non-zero elements in $\mathcal{D}$. The depth tensor $\mathcal{D}$ is sparse. Furthermore, the ME itself is sparse in the face region. The superposition of the two sparsities leads to a decrease in ME discriminability.

In the facial depth tensor $\mathcal{D}$, non-zero elements almost lie in the range of z from 500 to 700. Here, we use the 3σ criterion to compact $\mathcal{D}$, i.e., supposing that the depth distribution can be considered as a concentration within a range of 3 standard deviations (3σ) above and below the depth mean value (m), which means in the range of ($m-3\sigma$, $m+3\sigma$).

$$\begin{aligned} m &= \frac{1}{W\times H\times 2}\sum_{k=1}^{2}\sum_{x,y=1}^{W,H}\mathbf{D}_{I_k}(x,y) \\ \sigma &= \sqrt{\frac{1}{W\times H\times 2}\times\sum_{k=1}^{2}\sum_{x,y=1}^{W,H}[\mathbf{D}_{I_k}(x,y)-m]^2} \end{aligned} \tag{5}$$

where, $\mathbf{D}_{I_1}, \mathbf{D}_{I_2} \in \mathbb{R}^{W\times H}$ are two matrices, of which the elements are the depth values of frames I_1 and I_2. The distribution of the extreme values is generally outside this range. Therefore, the extreme values can be removed by setting a threshold based on the mean and standard deviation. However, as shown in Fig. 11, the distance between the curtain and the face is greater than that between the face and the camera. Thus, some 0 depth values lie in 3σ range and are needed to be removed. We choose non-zero minimal depth values of the two frames used for comparison. That is, the scene depth D of facial depth tensors for I_1 and I_2 in depth flow computation is obtained as follows.

$$\begin{aligned} D = &\max(\lceil\mathbf{D}_{I_1}, m+3\sigma\rceil, \lceil\mathbf{D}_{I_2}, m+3\sigma\rceil) \\ &-\min(\lfloor\mathbf{D}_{I_1}, 0\rfloor, \lfloor\mathbf{D}_{I_2}, 0\rfloor)+1 \end{aligned} \tag{6}$$

where, $\max(\mathbf{A},\mathbf{B})$ denotes the maximal value in all elements of matrices $\mathbf{A}$ and $\mathbf{B}$, and $\min(\mathbf{A},\mathbf{B})$ denotes the minimal value in all elements of matrices $\mathbf{A}$ and $\mathbf{B}$. $\lceil\mathbf{A}, b\rceil$ denotes the maximal value in elements less than b of matrix $\mathbf{A}$ and $\lfloor\mathbf{A}, b\rfloor$ denotes the minimal value in elements greater than b of matrix $\mathbf{A}$. Compared with directly computing D with $\max(\mathbf{D}_{I_1}, \mathbf{D}_{I_2})$ and $\min(\mathbf{D}_{I_1}, \mathbf{D}_{I_2})$, the value of D in

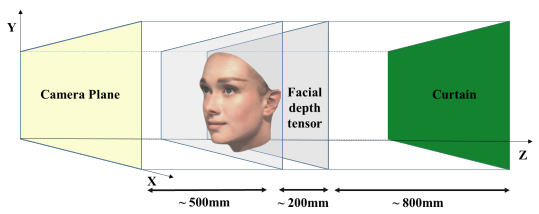


Fig. 11. Depth tensor illustration based on the ME data collection environment.

Eq. 6 is reduced from about 1,500 to 200. The operation of removing extreme values makes $\mathcal{D}$ become more compact.

After obtaining the scene depth D, the facial depth tensor could be built in a relative depth coordinate system. Supposing $\widetilde{\mathbf{D}}$ is the difference between the original depth information and the non-zero minimal depth values:

$$\widetilde{\mathbf{D}}_{I_k}(x,y) = \mathbf{D}_{I_k}(x,y) - \min(\lfloor \mathbf{D}_{I_1}, 0 \rfloor, \lfloor \mathbf{D}_{I_2}, 0 \rfloor) + 1 \quad (7)$$

where $k \in 1, 2$, representing the frame index for depth flow computation. Since the extreme values exist only for very few pixels, to simplify the calculation, we directly assign the relative depth values of these pixels to the minimum value in the depth direction. Hence, the depth information is updated as follow:

$$\mathbf{D}_{I_k}(x,y)^* = \begin{cases} \widetilde{\mathbf{D}}_{I_k}(x,y) & \text{if } 0 < \widetilde{\mathbf{D}}_{I_k}(x,y) \leq D \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

And the facial depth tensors $\mathcal{D}_1$ and $\mathcal{D}_2$ are constructed based on Eq. 4 with $\mathbf{D}_{I_k}(x,y)^*$:

$$\mathcal{D}_k(x,y,z) = \begin{cases} \mathbf{I}(x,y) & \text{if } z = \mathbf{D}_{I_k}(x,y)^* \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

5.2.2 Depth Flow

When we have two the same size facial depth tensors $\mathcal{D}_1$ and $\mathcal{D}_2$, we can compute the *depth flow* between them. In a facial depth tensor, for a specific moment t, the brightness of the voxel (x, y, z) is denoted as $I(x, y, z, t)$, where x, y and z are the coordinates in the horizontal, vertical and depth directions in the space, t is the time series coordinate of the tensor's corresponding frame in the ME video clip. Then the displacement of the voxel point between the two frames is embodied as: $I(\Delta x, \Delta y, \Delta z, \Delta t)$. Based on the conditions of constant brightness and continuous time, the equalization of the brightness of two frames is obtained as follows:

$$I(x,y,z,t) = I(x+\Delta x, y+\Delta y, z+\Delta z, t+\Delta t) \quad (10)$$

The right side of Eq. 10 could be transformed based on Taylor's theorem:

$$\begin{aligned} &I(x+\Delta x, y+\Delta y, z+\Delta z, t+\Delta t) \\ &= I(x,y,z,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial d}dz + \frac{\partial I}{\partial t}dt + \sigma \end{aligned} \quad (11)$$

where σ denotes the second-order infinitesimal term, which can be ignored. Replacing the right side of Eq. 10 with Eq. 11, we could obtain the following formula:

$$I(x,y,z,t) = I(x,y,z,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial z}dz + \frac{\partial I}{\partial t}dt + \sigma \quad (12)$$

which means:

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial z}dz + \frac{\partial I}{\partial t}dt = 0 \quad (13)$$

Then by dividing dt, Eq. 10 turns to:

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial z}\frac{dz}{dt} + \frac{\partial I}{\partial t} = 0 \quad (14)$$

By repesenting $\frac{dx}{dt}$, $\frac{dy}{dt}$, $\frac{dz}{dt}$, $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, $\frac{\partial I}{\partial z}$ and $\frac{\partial I}{\partial t}$ as u, v, w, I_u, I_v, I_w and I_t, we could get:

$$uI_u + vI_v + wI_w = -I_t \quad (15)$$

Here, (u, v, w) forms the depth flow $\boldsymbol{d_f}$ on voxel (x, y, z) for further motion analysis. In order to calculate $\boldsymbol{d_f}$, we introduce the classic Lucas-kanade algorithm [85]. Assuming that the depth flow for all the voxels in a small sliding cube around a voxel is the same, and W is the coefficient reflecting each voxel's weight, we transform Eq. 15 into:

$$\boldsymbol{W}^2\boldsymbol{A}\boldsymbol{d_f} = -\boldsymbol{W}^2\boldsymbol{\beta_t} \quad (16)$$

where:

$$\boldsymbol{A} = \begin{bmatrix} I_{u1} & I_{v1} & I_{w1} \\ & \vdots & \\ I_{um} & I_{vm} & I_{wm} \end{bmatrix}$$

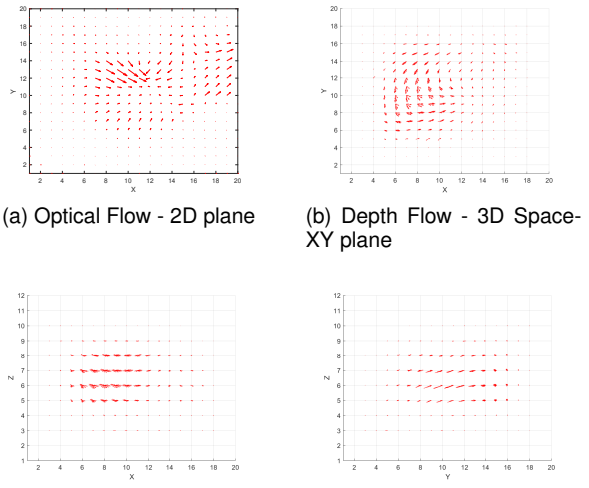
$$\boldsymbol{\beta_t} = \begin{bmatrix} I_{t1} \\ \vdots \\ I_{tm} \end{bmatrix}$$

and m is the number of voxels in the studied cube. Then by least squares method, the depth flow can be derived:

$$\boldsymbol{d_f} = (\boldsymbol{A}^T\boldsymbol{W}^2\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}^2(-\boldsymbol{\beta_t}) \quad (17)$$

After obtaining the depth flow vector $\boldsymbol{d_f}$ on each voxel (x, y, z), the depth flow tensor for the entire facial depth scene is constructed as a four-order tensor $\mathcal{F} = (\mathcal{U}, \mathcal{V}, \mathcal{W}) \in \mathbb{R}^{W \times H \times D \times 3}$, in which, the 3D tensor components $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ respectively represent the depth flow tensor on horizontal, vertical and depth direction, i.e., $\mathcal{U}(x, y, z) = u$, $\mathcal{V}(x, y, z) = v$, $\mathcal{W}(x, y, z) = w$.

Fig. 12 illustrates the comparison of OF and depth flow features. The sub-figures demonstrate that this new proposed feature considers the advantages of OF and depth image, and allows the algorithm to capture the motion features and the geometric deformation information sensitively. In this way, the extracted depth feature could be more representative for ME compared with the features without depth information.



(a) Optical Flow - 2D plane (b) Depth Flow - 3D Space-XY plane

(c) Depth Flow - 3D Space-XZ plane (d) Depth Flow - 3D Space-YZ plane

Fig. 12. Comparison between optical flow and depth flow.

5.2.3 MES with Depth Information

As the depth flow computation method is equivalent to extending the OF computation in the 2D plane to the depth scene, we chose the method proposed in [80] as mentioned above for MES with depth information.

The main idea in [80] is using the FD method. Firstly, the video is divided into short video clips by a sliding window. The OF on the face region for each frame is computed based on the following equalization.

$$\begin{aligned} I(x, y, f_1) &= I(x + \Delta x, y + \Delta y, f_1 + \Delta f) \\ &= I(x + \Delta x, y + \Delta y, f_i) \end{aligned} \tag{18}$$

where f_i denotes the ith frame in the short video clip, $I(x, y, f_i)$ represents the brightness of the pixels in f_i. The OF analysis is performed on the polar coordinate system. The effect of the overall head movement is removed by comparing it with the OF of the nose region. After extracting the spatial features from the regions of interest (ROIs), the final MES results are obtained by spatio-temporal feature fusion, multi-scale filtering, and thresholding processes.

For a fair comparison between OF and depth flow, the method procedure is identical. The difference is that we extract the depth flow (u, v, w) directly from the ROIs and the nose region to reduce the redundant computation. The angle calculation also differs from [80]. The depth flow is a 3D tensor, where the XY plane reflects the information of the grays image on the 2D plane. Since we would like to explore the effect of depth on the spotting performance, the angle θ is defined as being the angle between the vector $\boldsymbol{d_f}$ and the XY plane, i.e., the angle between $\boldsymbol{d_f}$ and its projection on XY plane ($\boldsymbol{d_{f_{XY}}}$, $(u, y, 0)$), the formula is shown below:

$$\begin{aligned} &\|\boldsymbol{d_f}\| = \rho = \sqrt{u^2 + v^2 + w^2} \\ &\|\boldsymbol{d_{f_{XY}}}\| = \sqrt{u^2 + v^2} \\ &\boldsymbol{d_f} \cdot \boldsymbol{d_{f_{XY}}} = u^2 + v^2 \\ &\theta = \text{acos}(\frac{\boldsymbol{d_f} \cdot \boldsymbol{d_{f_{XY}}}}{\|\boldsymbol{d_f}\| \times \|\boldsymbol{d_{f_{XY}}}\|}) \end{aligned} \tag{19}$$

Meantime, the magnitude of the depth flow (ρ) is also calculated in the above equation. After obtaining θ and ρ, we perform the algorithm with the same process to form the spatial-depth feature for the entire face region and then spot ME video clips in long videos.

5.2.4 Experimental result on MES

As introduced in Section 5.2.3, the method in [80] is utilized for result comparison. Therefore, we have two experimental settings: this basic method is performed on 2D videos, with the OF as the feature; the extended method is performed on videos with depth information, using our proposed depth flow as the feature.

We utilized the result evaluation method proposed by MEGC2020 in the MES task [86], in order to standardize the measurement criteria. As illustrated in Fig. 13, the addition of depth information allowed the spotting method to obtain more action information and improve the sensitivity of the system. Although many false positives (FPs) are detected, the number of true positives (TPs) for MEs is also increased. This improvement is also reflected in the Precision and F1 scores.

However, the results of MES are still not up to the requirements of practical applications. On the one hand, ME is very subtle and brief, and it is complicated for the algorithm to capture its features. On the other hand, traditional FD methods cannot discriminate MaEs, MEs, and other facial actions. In long videos, with the increase of interference, the algorithm's performance degrades compared with its performance in short videos. Depth information as an extra modal feature can help the system to improve the capacity of extracting ME features. Besides, in the future, supported by a large amount of ME sample data, the spotting method combined with deep learning may enhance the MES performance. The depth information reflects the scene depth and can better reflect the facial movement on the temporal domain, which can provide more possibilities for the construction of deep networks.

5.3 Benchmark for MER

Besides comparing depth flow in the facial depth space and OF on the 2D plane, we also explore the effect of depth information appending to RGB color information.

Since we only aim to demonstrate the additive effect of depth information on MER, we directly choose the pre-trained model: AlexNet [87] in Matlab to simplify network design workload. AlexNet has relatively shallow layers and requires fewer training parameters, so it can relatively alleviate overfitting to a certain extent. Many scholars also conduct research on network design applicable to the SSS problem for ME. However, we use AlexNet because it is the most common shallow layer network and is representative. We give a preliminary recognition result as a baseline to facilitate researchers to perform method comparisons.

As the apex frame normally contains the most representative ME feature, the apex RGB image and the corresponding depth image are used for feature learning. Before being imported into the network, the color image and the depth image are normalized respectively and then form the network's input through the channel-wise concatenation. The network structure is illustrated in Fig. 14.

Regarding the configuration, the experiment is performed by Matlab R2020b, with a single NVIDIA GeForce GTX 970 GPU. In order to maintain the specific design of

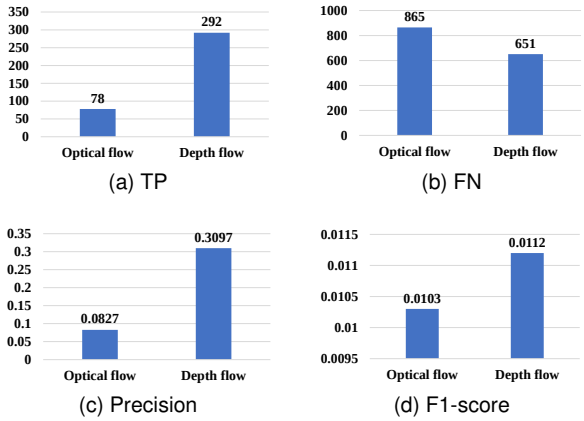


Fig. 13. Benchmark result for MES on CAS(ME)^3.

the network, the face region input on each channel is normalized to 227×227, i.e., the total input is with a dimension of $227 \times 227 \times 4$. Meanwhile, since we aim to study the effect of depth information on MER, in the comparison group, the input to the network consists of only the three channels of RGB color space ($227 \times 227 \times 3$). The initial learning rate is set to 0.5×10^{-4}, and the maximal number of epochs is 500.

Concerning the ME samples, ninety-five of the 100 subjects showed MEs during their recorded videos, for a total of 943 MEs. The leave one-subject-out cross-validation is used to validate the result. We performed experiments on two kinds of ME categories, i.e., ME classification based on the six basic emotions and the "Others" category and ME classification based on the four emotion classes (positive, negative, surprise, and others), represented by 7Emo and 4Emo in the latter, respectively.

For the result evaluation, we applied the metrics from the MER task in MEGC2019, i.e., unweighted average recall (UAR) and unweighted F1-score (UF1) [88], averaging the per-class recall and F1-score respectively. UAR and UF1 are adequate metrics in the case of unbalanced multiple classes, because they provide equal weight to the classes with smaller sample sizes by averaging. Thus, they are balanced judgments, reducing the likelihood that a method might be well-adapted to only some classes.

Fig. 15 lists the recognition result with or without depth information. As can be seen from the results, the addition of depth information improves the performance of the network in recognizing MEs. Furthermore, as illustrated in Fig. 16, we also analyze the recognition performance with depth information on each class using the confusion matrix. In conjunction with Fig. 9, it can be seen that the recognition performance of the categories with a large sample size is relatively better. In particular, negative MEs account for 54% of the total sample size. Consequently, the true positive rate (TPR) is much higher than the other categories. In contrast, positive MEs account for only 7% of the total, which is the smallest percentage, and therefore this category has the lowest TPR. The impact of sample size on recognition accuracy is higher than that of the features themselves since ME movements are very subtle. Therefore, it is difficult for the deep learning model to classify MEs with positive or happy emotions by the weak-amplitude AU12. In conclusion, the recognition performance is enhanced as the number of samples increases. Our database provides a relatively large number of ME samples, contributing to the improvement of MER. However, the sample imbalance problem of MEs seriously affects the performance of recognition for specific emotions. Therefore, how to improve the recognition performance for some specific kinds of ME samples with a relatively small amount deserves further exploration.

5.4 Database Comparison

We have performed state-of-arts (SOTA) methods to compare our CAS(ME)^3 database with other databases. Table 5 lists the MES performance of different SOTA methods on different databases. We selected the first-place method (SP-FD) [80] for the MEGC2020 spotting task and the top two methods (OF-FD and LSSNet) [43], [89] for the MEGC2021 spotting task. For the traditional FD methods, such as the SP-FD and OF-FD, better spotting results can be obtained for databases with relatively simple situations by suitable parameter settings and pre-processing. However, the generalization ability of this kind of method is weak. Since CAS(ME)^3 and CAS(ME)^2 have the same frame rate and similar resolution, we used the same parameter settings for MES. However, the results were unsatisfactory, and the first-place method of MEGC2021 (OF-FD) did not even detect TP. This is because the samples collected by our database are long videos, which contain many macro-expressions and head movements. In contrast, for the deep learning-based MES method, it can be seen that the performance is improved because the sample size of CAS(ME)^3 is larger than that of CAS(ME)^2. However, in general, the task of spotting MEs in long videos is still very challenging. Therefore, much subsequent research is expected to temporally localize and distinguish MEs from other facial movements.

In addition, regarding MER, we use three algorithms trained on SMIC, CASME II, and SAMM databases without fine-tuning for direct database evaluation while retaining the original parameters. The evaluation protocol is the same as Section 5.3, i.e., leave one-subject out cross-validation. Table 6 lists the MER performance of different SOTA methods

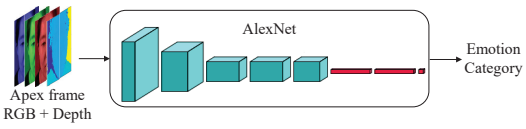
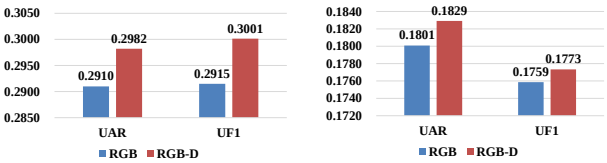


Fig. 14. MER network based on RGB-D information.



(a) 4Emo classification (b) 7Emo classification

Fig. 15. Benchmark result for MER on CAS(ME)^3. RGB and RGB-D denote the MER method without and with depth information, respectively.

	Positive	Negative	Surprise	Others
Positive	0.032	0.683	0.127	0.159
Negative	0.032	0.667	0.213	0.088
Surprise	0.055	0.592	0.224	0.129
Others	0.036	0.530	0.193	0.241

(a)

	Happiness	Disgust	Fear	Anger	Sadness	Surprise	Others
Happiness	0.016	0.254	0.048	0.143	0.000	0.381	0.159
Disgust	0.036	0.405	0.057	0.014	0.057	0.272	0.158
Fear	0.000	0.376	0.065	0.032	0.075	0.290	0.161
Anger	0.059	0.353	0.029	0.044	0.059	0.206	0.250
Sadness	0.048	0.210	0.097	0.065	0.097	0.355	0.129
Surprise	0.055	0.279	0.040	0.075	0.035	0.363	0.154
others	0.072	0.181	0.042	0.066	0.030	0.337	0.271

(b)

Fig. 16. Confusion matrix of MER based on AlexNet with depth information: (a) for 4Emo classification, (b) for 7Emo classification.

TABLE 5
MES performance (F1-score) comparison among different ME databases.

	CAS(ME)^3	CAS(ME)^2	SAMM Long Videos
SP-FD [80]	0.0103	0.0547	0.1331
OF-FD [89]	0	**0.1965**	**0.2162**
LSSNet [43]	**0.0653**	0.042	0.131

on different databases. Since our database contains more complex information about individuals and the distribution of sample types is different from the previously published database, the recognition results will be relatively unsatisfactory. In subsequent work, the performance of MER in large-scale complex data scenarios can be improved by adjusting the parameters or creating deep-learning models with better generalization.

5.5 Multimodal MER Analysis

We have performed the multimodality analysis combining physiological and voice data on Part C. EDA, as a sensitive and standard physiological indicator of emotional and sympathetic [92], is used for the result analysis. Since both the EDA signal and the speech signal are one-dimensional signals, we converted both signals into a speech spectrogram in grayscale form. Thus, we obtained three channels for RGB information, one channel each for depth information, EDA, and speech signals. We input the features to the fine-tuned AlexNet network for the MER task with different combinations of features. We obtained the recognition results by leave-one-subject-out validation, as shown in Table 7. It can be seen that the depth information can help the model to extract the ME movement information well and thus improve the MER performance. However, the recognition results of combining EDA or speech signals are not satisfactory. It may be because we did not perform better denoising and filtering of the 1D signal, and the speech spectrogram approach may not reflect the features of MEs well. The database provides a data platform for researchers to further optimize the processing of physiological and speech signals and explore the performance impact of these two modalities on ME analysis.

6 Conclusion and Perspective

6.1 Conclusion

MEs are very important nonverbal cues in emotion understanding. However, difficulties in elicitation, acquisition and annotation have caused the problems of SSS and the low ecological validity of MEs. In this paper, to address these two issues, we release a third-generation ME database by extending video samples to multi-modal data incorporating depth information and by combining the third generation of ME eliciting paradigm to obtain ME samples with high ecological validity. The amount of data in our database is comparable to the total amount of spontaneous ME database published, effectively avoiding the impact of database bias on MEA method validation. In addition, our paper contributes to MEA as follows. First, through an interdisciplinary exploration based on psychology, we verify the enhancement of depth information on human visual perception. Furthermore, the proposed depth flow allows the algorithm be more sensitive to the continuous changes of ME within the facial depth space and enhancing the MES performance. And we demonstrate the additive effect of depth information on RGB images by the improvement of the MER. Second, large scale of unlabeled data and the labeled data provide a platform for building self-supervised learning methods. And, we have explored multimodal self-supervised learning for ME accordingly by incorporating implicit learning in psychology. Third, mock crime is demonstrated to be feasible in eliciting high ecological validity MEs. Meanwhile, high ecological validity samples with physiological and voice signals provide a foundation for robust real-world MEA and emotion understanding.

TABLE 6
MER performance comparison among different ME databases.

		STSTNet [90]	RCN-A [47]	FR [91]
CAS(ME)^3	UF1	0.3795	0.3928	0.3493
	UAR	0.3792	0.3893	0.3413
SMIC	UF1	0.6801	0.6441	0.7011
	UAR	0.7013	0.6326	0.7083
CASME II	UF1	0.8382	0.8123	0.8915
	UAR	0.8686	0.8512	0.8873
SAMM	UF1	0.6588	0.6715	0.7372
	UAR	0.6810	0.7601	0.7155

TABLE 7
Multimodal analysis on MER for Part C. C, D, V and E denote RGB, Depth, Voice, and EDA, respectively.

	C	CD	CV	CE	CDV	CDE	CDVE
UAR	0.263	**0.296**	0.216	0.260	0.254	0.244	0.223
UF1	0.248	**0.296**	0.195	0.230	0.241	0.230	0.191

6.2 Perspective

Compared with expression analysis, MEA is a more complex task for both humans and computers. Dealing with complex tasks, the human could intuitively obtain abstract, unspeakable, and representational knowledge of internal structure by *implicit learning* [93]. The process of acquiring knowledge through implicit learning is similar to the process of self-supervised learning to auto-generate labels. There are three important features of implicit learning, namely unconscious process, more domains generalized and information presentation. They can be corresponded to these three features of self-supervised learning: unsupervised condition, stable parameter for downstream task, and multi-modality, respectively. Implicit learning is a very well-established study. By drawing on some of its theories, it may be possible to enhance further the performance and even the interpretability of self-supervised learning.

In the future, we will conduct self-supervised learning based MEA using unlabeled data in combination with depth information and lay special emphasis on the MES in long videos. In addition, research on MEA combining physiological and voice signals will also be explored. By combining traditional signal processing techniques and deep learning models, the performance of multi-modal MEA will be further improved. Finally, ME samples based on mock crime will be further analyzed to advance the ME application in complex real-world scenarios.

References

[1] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 792–800.

[2] J. Li, J. Zhao, C. Lang, Y. Li, Y. Wei, G. Guo, T. Sim, S. Yan, and J. Feng, "Multi-human parsing with a graph-based generative adversarial model," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1, pp. 1–21, 2021.

[3] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "Face. evoLVe: A high-performance face recognition library," *arXiv preprint arXiv:2107.08621*, 2021.

[4] Z. Wang, J. Zhao, C. Lu, F. Yang, H. Huang, Y. Guo *et al.*, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3443–3452.

[5] A. Kachur, E. Osin, D. Davydov, K. Shutilov, and A. Novokshonov, "Assessing the big five personality traits using real-life static facial images," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[6] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Facial color is an efficient mechanism to visually transmit emotion," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, pp. 3581–3586, 2018.

[7] B. A. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 1015–1023, 2014.

[8] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2018.

[9] C. Darwin, *The expression of the emotions in man and animals by Charles Darwin*. John Murray, 1872.

[10] W. E. Rinn, "The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions." *Psychological Bulletin*, vol. 95, no. 1, p. 52, 1984.

[11] J. Lee and M. R. Muzio, *Neuroanatomy, Extrapyramidal System*. StatPearls Publishing, Treasure Island (FL), 2020. [Online]. Available: http://europepmc.org/books/NBK554542

[12] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, p. 88–106, 1969.

[13] 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," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[14] P. Ekman, "Lie catching and microexpressions," *The philosophy of Deception*, p. 118–133, 2009.

[15] J. Endres and A. Laidlaw, "Micro-expression recognition training in medical students: a pilot study," *BMC Medical Education*, vol. 9, no. 1, p. 47, 2009.

[16] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological Bulletin*, vol. 129, no. 1, p. 74, 2003.

[17] R. A. de Wijk, V. Kooijman, R. H. Verhoeven, N. T. Holthuysen, and C. de Graaf, "Autonomic nervous system responses on and facial expressions to the sight, smell, and taste of liked and disliked foods," *Food Quality and Preference*, vol. 26, no. 2, pp. 196–203, 2012.

[18] P. Eckman, "Emotions revealed," *St. Martin's Griffin, New York*, 2003.

[19] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association. Sheraton New York, New York City*, 2009.

[20] M. Zhang, Q. Fu, Y.-H. Chen, and X. Fu, "Emotional context influences micro-expression recognition," *PloS ONE*, vol. 9, no. 4, p. e95018, 2014.

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[22] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[23] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[24] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[25] 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 *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[26] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS ONE*, vol. 9, no. 1, p. e86041, 2014.

[27] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME)^2: a database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, 2017.

[28] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–6.

[29] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, Jan 2018.

[30] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[31] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, 2020.

[32] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, and M. H. Yap, "Objective micro-facial movement detection using FACS-based regions and baseline evaluation," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 642–649.

[33] R. J. Sbordone and C. Long, "Ecological validity of neuropsychological testing," *Delray Beach Fl Gr Press/st*, 1996.

[34] B. H. Kantowitz, H. L. Roediger III, and D. G. Elmes, *Experimental psychology*. Cengage Learning, 2014.

[35] R. Whelan, "Effective analysis of reaction time data," *The Psychological record*, vol. 58, no. 3, pp. 475–482, 2008.

[36] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 51–56.

[37] S. Polikovsky, "Facial micro-expressions recognition using high speed camera and 3D-gradients descriptor," in *Conference on Imaging for Crime Detection and Prevention, 2009*, vol. 6, 2009.

[38] P. Husák, J. Čech, and J. Matas, "Spotting facial micro-expressions" in the wild," in *22nd Computer Vision Winter Workshop*, 2017.

[39] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[40] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE, 2015, p. 665–669.

[41] J. Li, C. Soladie, and R. Seguier, "Local temporal pattern and data augmentation for micro-expression spotting," *IEEE Transactions on Affective Computing*, 2020.

[42] S.-J. Wang, Y. He, J. Li, and X. Fu, "MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 3956–3969, 2021.

[43] W.-W. Yu, J. Jiang, and Y.-J. Li, "LSSNet: A two-stream convolutional neural network for spotting macro-and micro-expression in long videos," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4745–4749.

[44] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6034–6047, 2015.

[45] X. Li, H. Xiaopeng, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, 2017.

[46] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous

micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.
[47] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 8590–8605, 2020.
[48] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet: Dynamic imaging network for micro expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 1618–1627, 2019.
[49] M. Verma, M. S. K. Reddy, Y. R. Meedimale, M. Mandal, and S. K. Vipparthi, "AutoMER: Spatiotemporal neural architecture search for microexpression recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[50] Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 249–263, 2020.
[51] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018.
[52] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: a micro-expression recognition framework," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2936–2944.
[53] D. Zhang, G. Huang, Q. Zhang, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognition*, vol. 110, no. 11, p. 107562, 2020.
[54] P. G. Zimbardo and F. L. Ruch, "Psychology and life," 1975.
[55] F. Craik, "Aging and cognitive deficits," *Aging and Cognitive Processes*, 1982.
[56] T. K. Kim, "Understanding one-way anova using conceptual figures," *Korean Journal of Anesthesiology*, vol. 70, no. 1, 2017.
[57] J. G. Frederick, *Statistics for the behavioral sciences*. Wadsworth, 2013.
[58] C. Hong Liu, J. Ward, and A. W. Young, "Transfer between two- and three-dimensional representations of faces," *Visual Cognition*, vol. 13, no. 1, pp. 51–64, 2006.
[59] H. Liu, B. Laeng, and N. O. Czajkowski, "Does stereopsis improve face identification? a study using a virtual reality display with integrated eye-tracking and pupillometry," *Acta Psychologica*, vol. 210, p. 103142, 2020.
[60] E. Freud, A. K. Robinson, and M. Behrmann, "More than action: The dorsal pathway contributes to the perception of 3-d structure," *Journal of Cognitive Neuroscience*, pp. 1047–1058, 2018.
[61] H. Akhavein, *Depth-Cue Invariant Object Representations in the Visual Cortex*. McGill University (Canada), 2017.
[62] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927–1943, 2007.
[63] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.
[64] J. Foster, *The green screen handbook: real-world production techniques*. Routledge, 2014.
[65] P. Ekman and W. Friesen, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto: Consulting Psychologists*, 1978.
[66] https://dev.intelrealsense.com/docs/frame-management.
[67] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[68] Y. Yao, C. Liu, D. Luo, Y. Zhou, and Q. Ye, "Video playback rate perception for self-supervisedspatio-temporal representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[69] A. J. Fridlund, "Sociality of solitary smiling: Potentiation by an implicit audience." *Journal of Personality and Social Psychology*, vol. 60, no. 2, p. 229, 1991.
[70] G. Ben-Shakhar and E. Elaad, "The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review." *Journal of Applied Psychology*, vol. 88, no. 1, p. 131, 2003.
[71] K. Niioka, M. Uga, T. Nagata, T. Tokuda, I. Dan, and K. Ochi, "Cerebral hemodynamic response during concealment of information about a mock crime: Application of a general linear model with an adaptive hemodynamic response function," *Japanese Psychological Research*, vol. 60, no. 4, pp. 311–326, 2018.
[72] B. Verschuere, G. Ben-Shakhar, and E. Meijer, *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press, 2011.
[73] L. Sai, X. Zhou, X. P. Ding, G. Fu, and B. Sang, "Detecting concealed information using functional near-infrared spectroscopy," *Brain Topography*, vol. 27, no. 5, pp. 652–662, 2014.
[74] M. G. Coles, A. Gale, and P. Kline, "Personality and habituation of the orienting reaction: Tonic and response measures of electrodermal activity," *Psychophysiology*, vol. 8, no. 1, pp. 54–63, 1971.
[75] J. A. Matte, *Forensic psychophysiology using the polygraph: Scientific truth verification, lie detection*. JAM Publications, 1996.
[76] S. Vedam, P. Keall, V. Kini, H. Mostafavi, H. Shukla, and R. Mohan, "Acquiring a four-dimensional computed tomography dataset using an external respiratory signal," *Physics in Medicine & Biology*, vol. 48, no. 1, p. 45, 2002.
[77] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[78] R. Belaiche, R. M. Sabour, C. Migniot, Y. Benezeth, D. Ginhac, K. Nakamura, R. Gomez, and F. Yang, "Emotional state recognition with micro-expressions and pulse rate variability," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 26–35.
[79] https://pypi.org/project/dlib/.
[80] L.-w. Zhang, J. Li, S.-J. Wang, W.-j. Yan, X.-h. Duan, S.-c. Huang, and H. Xie, "Spatio-temporal fusion for macro- and micro-expression spotting in long video sequences," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020.
[81] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, p. 299–310, 2016.
[82] T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of noise in kinect depth images: A review," *IEEE Sensors Journal*, vol. 14, no. 6, pp. 1731–1740, 2014.
[83] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–10.
[84] A. Grunnet-Jepsen and D. Tong, "Depth post-processing for intel® realsense™ d400 depth cameras," *New Technologies Group, Intel Corporation*, vol. 3, 2018.
[85] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision." Vancouver, British Columbia, 1981.
[86] J. Li, S.-J. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "MEGC2020-the third facial micro-expression grand challenge," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE Computer Society, 2020, pp. 234–237.
[87] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
[88] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019– the second facial micro-expressions grand challenge," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.
[89] H. Yuhong, "Research on micro-expression spotting method based on optical flow features," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4803–4807.
[90] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.
[91] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition*, vol. 122, p. 108275, 2022.
[92] Critchley and D. H., "Book review: Electrodermal responses: What happens in the brain," *Neuroscientist*, vol. 8, no. 2, p. 132, 2002.
[93] A. S. Reber, "Implicit learning of artificial grammars," *Journal of Verbal Learning and Verbal Behavior*, vol. 6, no. 6, pp. 855–863, 1967.

Jingting LI is currently a postdoc at the Institute of Psychology, Chinese Academy of Sciences (CAS). She received the PhD degree in Signal, Image, Vision from CentraleSupélec in 2019. She served as the chair of the ACMMM'21 FME workshop and challenge, organized and hosted several China Society of Image and Graphics (CSIG) online ME workshop sessions. Her current research interests include image processing, computer vision and pattern recognition.

Zizhao Dong received the B.S.Ed degree in applied psychology major from China Women's University, China, in 2018. She is currently pursuing the M.Psy. degree in the Institute of Psychology, Chinese Academy of Sciences. Her current research interests include facial micro-expression , visual psychophysics and neurophysiology.

Shaoyuan Lu received the B.S. degree in Psychology major from Anhui Normal University, China, in 2020. She is currently pursuing the M.Psy. degree in the Institute of Psychology, Chinese Academy of Sciences. Her current research interests include facial micro-expression, visual psychophysics and neurophysiology.

Su-Jing Wang (M'12-SM'19) is an Associate Researcher, PhD supervisor at the Institute of Psychology, CAS. He received the Ph.D degree from the College of Computer Science and Technology of Jilin University in 2012. His current research interests include pattern recognition and machine learning. He won the first prize of the 8th Wu Wenjun Artificial Intelligence Science and Technology Award in 2018. He was selected as one of the top 2% of scientists in the world in 2020 for "Impact of the Year".

Wen-Jing Yan Wen-Jing Yan is an associate professor at Wenzhou Medical University. He received his PhD from the Institute of Psychology, Chinese Academy of Sciences. He is now conducting interdisciplinary research on affective computing, lie detection, and mental health. In addition to basic research, he is developing products for practical applications, such as intelligent lie detectors and adaptive mental health examination robots.

Yinhuan Ma received the B.S. degree in computer science and technology from Jiangsu University of Science and Technology, China, in 2019. He is currently pursuing the M.S. degree in computer science and technology at Jiangsu University of Science and Technology. His current research interests include computer vision, pattern recognition, and facial micro-expression analysis.

Ye Liu Ye Liu is an Associate Researcher, master supervisor at the Institute of Psychology, CAS. She received the Ph.D degree in psychology from the Institute of Psychology, CAS in 2005. Her current research interests include representation of semantic memory and affective computing. She is technical committee members of Human Computer Interaction of CCF, the Artificial Intelligence and Artificial Emotion of CAAI, and the Human Computer Interaction of CSIG.

Changbing Huang received his Ph.D degree in Biophysics and Neurobiology from the University of Science and Technology of China in 2006 and started postdoctoral training in Department of Psychology, University of Southern California from 2007. He joined the Institute of Psychology, CAS in 2011. He currently conducts researches on: (1) mechanisms and treatment of amblyopia, myopia, and aging, (2) psychophysical, imaging, EEG, and computational study of visual perception, perceptual learning, and visual memory.

Xiaolan Fu (M'13) received her Ph. D. degree in 1990 from Institute of Psychology, Chinese Academy of Sciences. Currently, she is a Senior Researcher at Cognitive Psychology. Her research interests include visual and computational cognition: (1) attention and perception, (2) learning and memory, and (3) affective computing. At present, she is the director of Institute of Psychology, Chinese Academy of Sciences and the director of department of psychology, University of the Chinese Academy of Sciences.