Micro-Expression Key Frame Inference

Su-Jing Wang, *Senior Member, IEEE,* Yu-Han Miao, Jingting Li, *Member, IEEE,* Ling Zhou, Zizhao Dong, Mengyi Sun and Xiaolan Fu, *Member, IEEE*

Abstract—Micro-expressions (MEs) are brief, involuntary facial movements critical for detecting lies, drawing growing interest in psychology and computer science. However, annotating ME can burden human coders with excessive time commitment and overwhelming information that compromises coding reliability and efficiency. Such difficulties in data annotation also led to the small sample size problem and hindered the development of ME analysis. Specifically, our psychological research highlights the complexities involved in human annotation of key frames. To facilitate the annotating process of ME, we proposed the Micro-Expression Key Frame Inference (ME-KFI) problem, aiming to identify MEs' temporal locations from a single frame, reducing manual annotation effort. We propose a Micro-Expression Contrastive Identification Annotation (MECIA) method as a solution to ME-KFI, including three modules: a contrastive module, an identification module, and an annotation module, corresponding to the three steps of manual annotation. The network's outputs infer the key frame of ME clips. MECIA demonstrates superior performance over random baselines on SAMM and CAS(ME)^2 databases and maintains comparable recognition accuracy with ground-truth clips.

Index Terms—Micro-Expression, Key Frame Inference, Micro-Expression Annotation.

1 Introduction

MICRO-expression (ME) is an involuntary, momentary, and subtle facial expression with a brief duration of less than 500ms [1]. It reflects one's genuine emotions that people are trying to conceal. In contrast to ordinary facial expressions, ME is consciously suppressed but unconsciously leaked. By exploring the momentary movements of facial muscle tonus, researchers have found that ME provides valuable diagnostic information for affective appraisal that reveal suppressed emotion and internal conceptual conflict. These findings have offered glimpses into critical areas of human behavior that was not possible using traditional observation assessments. More importantly, these results have shined lights into the possible direction of applying ME in practical settings, such as national security [2], medical care [3], studies on political psychology [4], and educational psychology [5].

To improve human's ability to detect ME, Ekman et al. developed a tool called Micro-Expression Training Tool (METT) [6]. However, even after METT training, human ME detection accuracy was still below the random level [7]. Consequently, efficient automatic ME analysis is necessary for further practical applications of ME. ME analysis consists of ME spotting and ME recognition, respectively. ME spotting is to detect the presence of ME clips in facial videos, and, if so, to temporally locate the onset and offset frames of ME clips. Meantime, ME recognition refers to classifying a given ME clip into an emotional category.

Research on ME analysis has been developing since the beginning of the century. Along with the rise of deep-learning methods in the past two years, increased efforts have been made to upscale the ME analysis performance. However, the performance of ME analysis has not been greatly improved due to the limitation of its small sample size problem [8]. Specifically, ME samples are the basis of the automatic ME analysis research. Current published spontaneous ME databases include CASME [9], CASME II [10], CAS(ME)^2 [11], CAS(ME)^3 [12], SMIC [13], 4DME [14], SAMM [15], MMEW [16] and DFME [17]. The total sample size in these databases is relatively small, only around 10,000 samples in total, limiting the development of deep learning in ME analysis.

Despite the urgent need for large-scale ME databases, increasing the sample size in a ME database faces tremendous difficulties, especially on ME manual annotation. The limitation can be attributed to the intensive training required for obtaining the qualification and the extremely time-consuming nature of the annotation process. Taking the Facial Action Coding System (FACS) as an example, proficiency in such an anatomically based coding system requires more than l00 hours of training [18]. In addition to that, the ME temporal annotation is also time-consuming but necessary.

The research on the visual neuroscience of human perception demonstrated that the low-dimension space by time manifold representation of facial expression leads to reliable categorization of each emotion [19]. In addition, psychological studies have suggested that temporal information can facilitate individuals' ability to perceive emotions by subtle expression [20]. Furthermore, typical confusions of fear and surprise and of disgust and anger caused by shared Action Unit activations can be resolved by identifying the

This paper is supported in part by grants from the National Natural Science Foundation of China (62276252, 62476269, 62106256), and in part, by a grant from the Youth Innovation Promotion Association CAS.

S.J Wang, J.T Li, Z.Z Dong and M.Y Sun 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
Y.H Miao is with the School of Computer Science, Jiangsu University Of Science and Technology, Jiangsu, 212028, China.
L. Zhou is with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China.
X.L Fu is with the School of Psychology, Shanghai Jiao Tong University, Shanghai 200030, China.

sequencing of Action Unit components, i.e., temporal information of facial expressions. Besides, for computer vision, the static spatial information of single frames in ME videos is not distinctly different from a neutral face due to the very low intensity of ME. In contrast, the video contains distinguishing transition information of facial actions, allowing algorithms to learn ME temporal features. Hence, this inter-frame dynamic information is vital for ME analysis.

However, temporal annotation can burden human coders with excessive time commitment and overwhelmed information that compromises annotation reliability and efficiency. In particular, when annotating the temporal location of ME clips, it takes only a few seconds to confirm if there is an action on the face. However, the coder spends much more time looking for the onset and the offset frames. Specifically, the change at these critical locations may be slight compared with the neutral face, thus it could take a few minutes to compare and confirm which frame is the transition frame.

Therefore, advancing the traditional annotation techniques with computer vision technique to identify the changing frames (both onset and offset frames) within long videos would improve the efficiency and reliability of ME manual annotation. For the situation, contributions of the paper are as follows.

- We conduct a psychological study on manual key frame annotation, the experiment results reveals the challenges faced by individuals in annotating key frames.
- To the best of our knowledge, it is the first time to propose the ME Key Frame Inference (ME-KFI) problem in ME analysis area, as illustrated in Fig. 3. ME-KFI targets to infer the temporal location of ME based on the single manually annotated frame in ME clips, alleviating the difficulty in ME manual annotation.
- We proposed a ME Contrastive Identification Annotation (MECIA) method as a ME-KFI solution to infer the key frame of ME clips. Inspired by the three steps of manual annotation, we designed three modules in the network, i.e., Identification module, Contrastive module and Annotation module, improving the key frame accuracy for ME clips.
- Furthermore, the experimental results and ablation study demonstrate the effectiveness of our proposed MECIA method. Besides, the method is proven capable of significantly reducing the labor cost of ME Manual Annotation by over 70%.

2 Related Works

2.1 Micro-Expression Recognition

The task of ME recognition involves classifying ME video clips into specific emotional categories. Methods used for ME recognition can be divided into handcrafted feature extraction methods and deep learning methods. With the low computational complexity, LBP-TOP [21] families with SVM are the most popular handcrafted feature extraction methods for ME recognition.

While handcrafted feature extraction mostly relies on a manually designed extractor, which needs specialized knowledge and a tedious parameter adjustment process, ME recognition methods have gradually evolved from handcrafted feature extraction methods [22]–[29] to deep learning-based recognition methods [30]–[38]. For instance, Li et al. [39] proposed a novel ME AU detection method using high-order spatial and channel-wise statistics to enhance robustness and capture subtle facial changes. During the research process, it is shown that shallow networks with motion features obtained from the onset and apex frames can efficiently reduce over-fitting and improve the recognition performance in ME recognition. For example, Zhou et al. [40] applied a shallow network named Dual-Inception Network with the feed of TV-L1 optical flow, which worked well in ME recognition without any data augmentation. Besides, based on Dual-Inception Network, they proposed Feature Refinement (FeatRef) [41] with expression-specific feature learning and fusion for ME recognition that successfully obtains salient and discriminative features for ME recognition. Furthermore, Liong et al. [42] proposed STST-Net, which is a two-layer neural network that is capable of learning the features from three optical flow features. More recently, RCN-A [31] is employed for ME recognition. With the integration of RCN without increasing any learnable parameters, RCN-A can enhance the representation ability in various perspectives. Furthermore, Nguyen et al. proposed Micron-BERT (u-BERT) [43], a method for ME recognition using Diagonal Micro-Attention and a Patch of Interest module, achieving state-of-the-art performance.

2.2 Micro-Expression Spotting

ME spotting involves detecting the occurrence of MEs in a video and identifying their temporal windows. Since 2014, many research teams [10], [22], [44], [45] have used the feature difference method to spot spontaneous ME in videos. The feature differences between the frames are calculated to spot the most significant movement through a threshold in the video. However, the ability of these methods 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 machine learning / deep learning are gradually becoming the mainstream of ME spotting since the model could learn the features specific to ME. For instance, [46] and [47] extracted handcraft features and then used the SVM classifier to spot ME frames. Liu et al. [48] proposed a duration and mode-aware framework using ensemble learning with multi-scale sliding windows for ME spotting. Esmaeili et al. [49] introduced intelligent cubic-LBP with CNN and PACF for improved apex detection in micro-movements. Besides these, there are some ME spotting methods using deep learning techniques, learning ME features directly through the model, such as MESNet [50], LSSNet [51], action unit-based network [52], LGSNet [53]. However, 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 Distinction from Micro-Expression Spotting and Recognition

Noteworthy, the ME-KFI and ME spotting have different meanings. ME spotting involves determining the ME pres-

ence in entirely unknown videos. The significance of this process is predominantly manifested in the practical applications of MEs. However, its evaluation metrics typically only require that the detected interval overlaps with the ground-truth interval (e.g., Intersection over Union (IOU) ¿ 0.5). This relatively loose criterion means that detection results are not precise enough to be released as accurately annotated samples in a database, since the exact key-frame positions remain uncertain.

ME recognition, on the other hand, assumes that a ME has already been precisely identified and focuses on classifying its emotional category.

Our proposed ME-KFI task differs from both approaches. We assume that a ME is known to exist in the video and that at least one of its frames has been identified. Our method aims to infer the onset and offset frames of the ME. This is neither about discovering MEs in unlabeled videos (spotting) nor classifying a known ME's emotion (recognition). Instead, ME-KFI is a precursor task to ME spotting and recognition, primarily designed for ME annotation. It infers the key frames of an ME clip based on given annotated frames, utilizing a weakly supervised learning process. By simplifying the manual annotation process, ME-KFI has the potential to replace the labor-intensive task of manually annotating the onset and offset frames of MEs, significantly saving time and reducing labor costs.

The ME-KFI problem, along with the corresponding ME-CIA solution, serves as a trade-off between manual annotation and automation. Its implementation lays a solid foundation for efficiently constructing large-scale, high-quality ME databases. Such a foundation is critical for advancing data-driven deep learning in intelligent ME analysis, and it provides support for the development of ME spotting methods as well. As far as we know, this is the first attempt at ME localization, avoiding tedious annotations and expensive costs.

3 Micro-Expression Manual Annotation

In this section, we describe the manual annotation process for two key reasons. First, many researchers in micro-expression analysis are not well-versed in the procedures used to annotate micro-expression data, particularly those employed by FACS-certified coders who follow objective and standardized protocols. Providing this background offers informative context and clarity. Second, our aim is to address the significant challenges human coders face in identifying the onset and offset frames. By explaining the typical annotation process, we highlight the necessity of our approach and demonstrate how it alleviates the time-consuming and complex nature of key-frame labeling.

3.1 Manual Annotation Process

The process of manually annotating MEs can be roughly divided into three steps, namely the *global* observation, the *local* observation, and the *frame-by-frame* observation.

In the *global* observation, the coder should watch the video clip of ME occurring to ascertain the approximate time frame for generating the facial movement. Next, In the *local* observation, the coder must pinpoint the precise location of the facial movement. In particular, the movement patterns of the forehead, eyebrows, or other areas. Finally, the coder should concentrate on the specific area and compare frame by frame before and after, looking for the frame where the facial movement has just begun to change from a neutral expression, i.e. the onset; the frame where the facial movement has just returned to a neutral expression, i.e. the offset; and the frame where the facial movement has reached its highest amplitude, i.e. the apex. Therefore, this step is the *frame-by-frame* observation.

To investigate the time cost of manual annotation, we selected 8 ME clips from the SAMM database to manually annotate their onset, apex, and offset. The 8 ME clips correspond to 8 ME categories in the database. Each subject would have to appear only once in the 8 ME clips. Two coders participate in the annotation. In addition, we removed data having a margin of error in the annotation of more than 20 frames compared to the SAMM database, resulting in 41 data listed in Table 1. The time spent on the onset and offset of the annotation accounted for 74.85% of the total annotation time.

The paper proposes the ME-KFI problem, which aims to infer the onset and offset of ME clip based on the single annotation ME frame. This proposition seeks to reduce the time cost associated with ME annotation.

3.2 Comparison of Manual Annotating Onsets of Micro-expression and Macro-expression

MEs and macro-expressions(MaEs) represent more subtle and more overt emotional expressions, respectively. Based on this, we hypothesized that compared to MaEs, the onset frames of MEs would be more challenging for human annotators to accurately identify. To test this hypothesis, we designed an experiment comparing the accuracy and reaction times of human annotators in identifying the onset frames of MEs versus MaEs. The experiment utilized a single-factor within-subjects design with two levels. The results showed that there were no significant differences in accuracy or reaction times between identifying the onset frames of MaEs and MEs for the human annotators. Furthermore, manually annotating the key frames of micro-expressions is extremely challenging. In contrast, our proposed ME-KFI approach (Section 4) infers the remaining onset and offset frames automatically. This improves labeling efficiency and enables the creation of large-scale, accurately annotated datasets.

3.2.1 Subjects

Ten graduate students (6 males and 4 females; Mean (M) = 24.3 years, Standard Deviation (SD) = 1.33 years) were recruited as subjects for this experiment and were compensated for their involvement. All subjects volunteered for the experiment, possessing normal or corrected-to-normal vision, with no known psychiatric disorders. Importantly, all research involving human subjects mentioned in this article adhered to the Declaration of Helsinki and received approval from the Ethics Review Committee of the Institute of Psychology, Chinese Academy of Sciences.

3.2.2 Stimuli

In the CAS(ME)^2 database, 60 expression samples were selected, comprising 30 MEs and 30 MaEs. Each ME cor-

TABLE 1
Manual annotation time. The - represents removed data having a margin of error.

	Coder1								Coder2								Mean	%over total
Sample	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8		
Onset (s)	44	85	55	49	77	60	37	49	187	154	66	31	39	20	55	-	67.20	35.44%
Apex (s)	19	60	27	-	27	11	17	28	77	63	116	-	48	-	93	34	47.69	25.15%
Offset (s)	33	65	56	100	-	-	-	48	132	73	165	57	73	29	-	66	74.75	39.42%

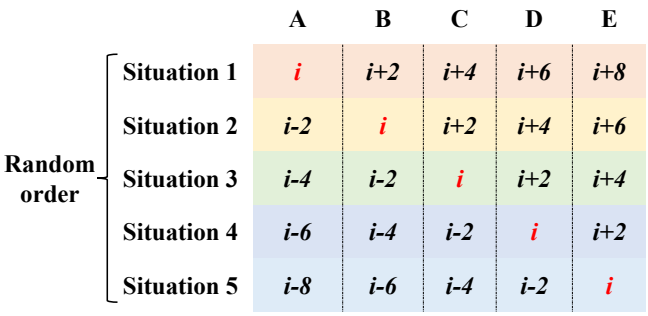


Fig. 1. Five situations of frame sets for onset frame identification. i denotes the frame index of onset frame.

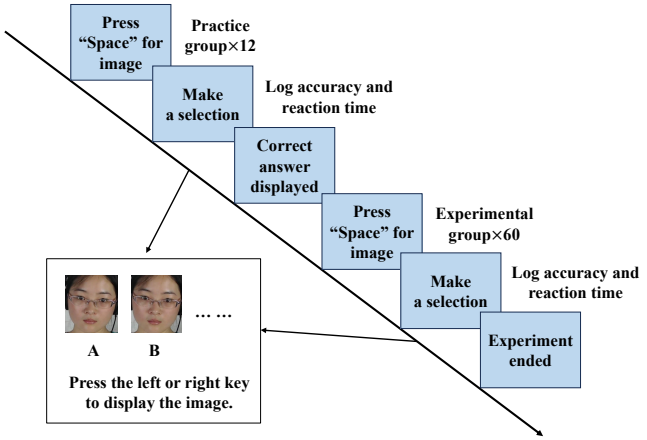


Fig. 2. Experiment procedure of psychological study on the onset manual annotation.

responds to a MaE of the same face and the same emotional type. There are six types of emotions covered: disgust, happiness, anger, surprise, fear, and sadness. To mitigate the impact of background information, the images were cropped to focus predominantly on the facial region. As shown in Fig. 1, the onset frame is indexed at i. For our experiment, we selected a subset of frames, taking every second frame from $i-8$ to $i+8$, with i serving as the midpoint. The rationale for selecting every second frame, skipping one between each, stems from the difficulty for participants untrained in annotating to accurately identify the correct onset frame when presented with a sequence of consecutive frames around onset. This challenge could potentially impede the execution of the experiment. Then, each expression displayed to the subject consists of five frames, sequenced according to the progression of the expression. To simulate the real-life process of observing facial movement patterns, subjects could navigate back and forth through these images using the left and right arrow keys. Each image was labeled with a letter option below it, i.e., A, B, C, D, E. The appearance of each situation was ensured to be random. These expression images, serving as stimulus materials, were displayed on a laptop screen placed approximately 0.5 meters away from the subjects.

3.2.3 Procedure

Subjects observed facial expression images in a quiet laboratory setting, being instructed to select the onset frame of each expression (i.e., the instant when the expression starts) as quickly and accurately as possible. The experiment procedure is illustrated in Fig. 2. The accuracy of their choices and their reaction times were recorded the moment they pressed the corresponding keys. Before starting the main experiment, subjects were required to complete a practice trial to familiarize themselves with the procedure. During the practice phase, the correct answer is displayed after subjects make their selection. This enables subjects to more intuitively grasp what the onset frame actually looked like. The 12 facial expression images used for practice were also selected from the CAS(ME)^2 database and were not repeated in the main experiment. Thus, data from the practice were not included in the results analysis.

3.2.4 Result

A paired samples t-test was employed to analyze the experimental results, as shown in Table 2.

Subjects' accuracy and reaction time means and standard deviations were calculated, and the differences in accuracy and reaction times between ME and MaE conditions were also compared. Results showed that subjects' mean accuracy for identifying MEs (M = 0.21, SD = 0.40) was slightly lower than that for MaEs (M = 0.23, SD = 0.42), and the mean reaction time for identifying MEs (M = 14.57 s, SD = 12.00 s) was also marginally slower than for MaEs (M = 14.23 s, SD = 12.42 s). However, statistically, there were no significant differences in subjects' accuracy or reaction times ($t(299) = 0.57$, $p>0.5$; $t(299) = -0.40$, $p>0.5$). Regardless, the accuracy significantly below random levels and the average reaction time exceeding 10 s both highlight the difficulty of accurately annotation the onset frame.

Moreover, we emphasize that while coding consistency is a crucial factor in data annotation, in our experiment, we used data with high coding consistency as ground truth, eliminating the need to re-calculate inter-annotator reliability. Additionally, since individual differences in annotation are not directly related to our experimental objective, i.e., highlighting the challenge in annotating key frames, we did not include such calculations.

TABLE 2
The result for recognizing the accuracy and reaction time of the onset frame of MEs and MaEs.

	Accuracy				Reaction time			
Expression type	M	SD	T	P	M	SD	T	P
ME	0.21	0.40	0.57	>0.5	14.57	12.00	-0.40	>0.5
MaE	0.23	0.42			14.23	12.42		

3.2.5 Discussion

In this study, we explored the ability of human annotators to identify the onset frames of MaEs and MEs, hoping to uncover a significant difference between the two. Contrary to our hypothesis, there was no significant difference in the accuracy and reaction time when recognizing the onset frames of both types of expressions. We believe that for annotators, identifying the onset frames of both MaEs and MEs may present a similar level of challenge. This challenge did not manifest in substantial differences in accuracy or reaction times, which may be due to participants reaching their recognition ceiling when confronted with more difficult tasks.

The common feedback on task difficulty might reflect that participants reached a limit in their ability to recognize expression onset frames within the task context. This situation prompts us to further reflect on the experimental design and to explore how to adjust task difficulty and experimental settings effectively to measure human recognition capabilities without limiting performance due to excessive task difficulty.

4 Micro-Expression Key Frame Inference

4.1 Formal problem

We give the formal definition for the ME-KFI problem. The given long video $\mathbf{V} = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$ with T frames contains one or several ME clips. x_t represents the image at t-th frame with a corresponding annotation label $y_t^a \in \{-1, 0, 1\}$. $y_t^a = 1$ means that x_t is labeled as the ME frame. In this case, x_t has another label $\mathbf{y}_t^m \in \mathbb{R}^4$ called ME label. It is a one-hot vector to indicate which ME categories x_t belongs to. $y_t^a = 0$ means that x_t is labeled as the non-ME frame. $y_t^a = -1$ means that x_t is an unlabeled frame. And it further needs to be labeled as a ME or non-ME frame by the proposed algorithm.

Initially, there is one and only one frame that is manually labeled as the ME frame, i.e. $y_t^a = 1$, in each ME clip in the long video $\mathbf{V}$. And others are unlabeled frames. We need to infer to all unlabeled frames that is the ME frame or the non-ME frame. The proposed algorithm aims to make that there are a few -1 elements as possible in $\mathbf{y}^a$. If x_t is in a certain ME clip, then the corresponding label $y_t^a = 1$, otherwise $y_t^a = 0$.

The ME-KFI problem is a weak-supervised learning problem. Extremely few frames are labeled in the long video. Manually labeling each frame is time-consuming and labor-intensive. So, to save labor, we propose a strategy which only needs label a frame for each ME clip. Then we infer the key frame of MEs, i.e., the onset, apex, and offset of each ME clip, by the proposed Algorithm 1, where $\mathrm{I_m}(\mathbf{y})$ denotes the index of the maximum element in the vector $\mathbf{y}$.

In the algorithm, some frames which are not any frame in the ME clips are annotated as non-ME frames. Then a MECIA-Net is trained by the labeled frames. And the trained MECIA-Net has an output $\hat{y}^a$ to predict whether each unlabeled frame is a ME frame. If the unlabeled frame is predicted as a ME frame, MECIA-Net has another output $\hat{\mathbf{y}}^m$ to predict the ME category further. Finally, according to outputs of the MECIA-Net, the algorithm infers the onset, apex, and offset of ME clips.

Algorithm 1: Micro-Expression Contrastive Identification Annotation (MECIA)

Input: a long video $\mathbf{V} = \{x_1, x_2, \ldots, x_T\}$ and two label sets: annotation label set $\mathbf{y}^a \in \mathbb{R}^T$ and ME label set $\mathbf{Y}^m \in \mathbb{R}^{4\times T}$

Initialization: randomly pick out several frames x_t are not any frame in the ME clips and assign the corresponding annotation label as 0, i.e., $y_t^a \leftarrow 0$;
while $-1 \notin \mathbf{y}^a$ **do**
  training a MECIA-Net by the labeled frames, and calculating the $\hat{y}_t^a$ of each frame x_t by the trained MECIA-Net;
  if $\hat{y}_t^a > 0.5$ **then**
    /* if x_t is predicted to a ME frame, then calculating its ME category. */
    calculating the $\hat{\mathbf{y}}_t^i$ by the trained MECIA-Net;
  end
  /* Frame Inference (see Section 4.4) */
  find anchor frames $\{x_a\}$;
  for *each anchor frames* x_a **do**
    inferring frames toward the left with the interval r by calling Algorithm 2;
    similarly, inferring frames toward the right;
  end
  Update y_t^a and $\mathbf{y}_t^m$ where $t = 1, 2, \ldots, T$
end
for *each inferred ME clip* $\{x_i, x_{i+1}, \ldots, x_j\}$ $(x < j)$ **do**
  $m = \mathrm{I_m}([\hat{\mathrm{y}}_{\mathrm{i}}^{\mathrm{a}}, \hat{\mathrm{y}}_{\mathrm{i+1}}^{\mathrm{a}}, \ldots, \hat{\mathrm{y}}_{\mathrm{j}}^{\mathrm{a}}]^{\mathrm{T}})$;
  x_m is the apex of the inferred ME clip;
end

4.2 Initialization

It is known that MEs are very fast with a brief duration. Thus, given a long video $\mathbf{V}$, only a few frames belong to ME clips (positive samples), while the vast majority of frames are non-ME frames (negative samples), which means that the size of ME frames and non-ME frames are unbalanced in the long videos. To guarantee that the initial annotation ME/non-ME quantity is more balance to train a model with

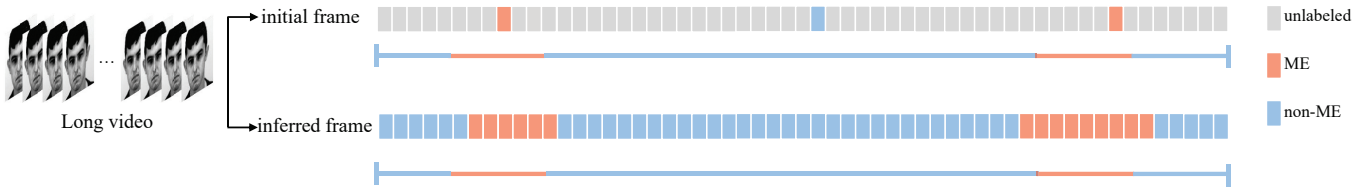


Fig. 3. ME-KFI problem: There is a video containing one or several ME clips. Each ME clip has one or labeled frame. ME-KFI needs to infer key frame of onset and offset for each ME clip. GT means ground truth.

better feature learning ability, we need to ensure that the number of initial annotation of ME frames and that of non-ME frames are more approximate. Thus, in our approach, we randomly selected some frames x_i not located in the episode of MEs, and initially annotate them as non-ME frames with the label of 0, i.e., $y_i^a \leftarrow 0$.

4.3 Micro-Expression Contrastive Identification Annotation Network (MECIA-Net)

We used labeled frames, which annotation labels are 1 or 0, to train a deep network named ME Contrastive Identification Annotation Network (MECIA-Net). When an unlabeled frame x_t is fed to the trained MECIA-Net, outputs of the network are $\hat{y}^a \in \mathbb{R}^1$ and $\hat{\mathbf{y}}^i \in \mathbb{R}^4$. $\hat{y}^a$ denotes the probability that x_t belongs to ME clips. $\hat{\mathbf{y}}^i$ is a one-hot vector and denotes the ME categories to which x_t belongs.

The architecture of MECIA-Net is shown in Fig. 4. In the network, AlexNet [54] or ResNet18 [55] is used as the backbone of MECIA-Net to extract features of each frame. Specifically, while more advanced backbones have emerged, classical architectures such as AlexNet or ResNet-18 remain widely used for feature extraction due to their simplicity and well-understood behavior [38], [56], [57]. Since our focus is on the proposed ME-KFI method rather than introducing a novel feature extraction network, using established backbones allows us to concentrate on the core idea without unnecessary complexity. The extracted features are fed into one Fully-Connected layer with the ReLU activation function to generate a 1024-dimensional feature $\mathbf{f} \in \mathbb{R}^{1024}$ as the input of three modules of MECIA-Net. The three modules are the annotation module, the identification module, and the contrastive module. They correspond to the global observation, the local observation, and the frame-by-frame observation of the manual annotation in Section 3.1, respectively. While all three modules share the same input features, they serve distinct purposes and are designed to simulate the three key stages of manual annotation, making their roles meaningful and specialized. This structured approach ensures that the framework aligns with real-world annotation practices while effectively performing key-frame inference.

4.3.1 The annotation module

The annotation module roughly corresponds to the *global* observation. In the *global* observation, the coder glances at the entire video and quickly locates the clip where there is facial movement. This process is the coder roughly classifying frames as ME or non-ME frames.

In the annotation module, we use one Fully-Connected layer with the sigmoid activation function to transform $\mathbf{f}_t$ into a scale $\hat{y}_t^a \in [0, 1]$. $\hat{y}_t^a$ represents a probability that the frame x_t belongs to a ME frame. The higher $\hat{y}_t^a$, the higher probability that x_t belongs to a ME frame. The lower $\hat{y}_t^a$, the higher probability that x_t belongs to a non-ME frame.

Suppose Ω_m is a set of indexes of ME frames, and Ω_n is a set of indexes of non-ME frames. Then the set of indexes of all labeled frames can be described as $\Omega_l = \Omega_m \cup \Omega_n$. The loss of the annotation module is calculated by

$$\mathcal{L}^a = -\frac{1}{|\Omega_m|}\sum_{t\in\Omega_m}\log \hat{y}_t^a - \frac{1}{|\Omega_n|}\sum_{t\in\Omega_n}\log\left(1-\hat{y}_t^a\right) \quad (1)$$

where $|\Omega|$ denotes the cardinality of Ω.

4.3.2 The Identification module

The identification module roughly corresponds to the *local* observation. In the *local* observation, the coder focus on the areas of facial movement. According to the facial movement areas, the coder may refer to the movement belongs to which ME category.

In the identification module, ME frames will further be identified as which ME category. One Fully-Connected layer with the softmax activation function is used to transform $\mathbf{f}_t$ into a 4-dimensional vector $\hat{\mathbf{y}}_t^i$. $\hat{\mathbf{y}}_t^i$ represents a probability distribution over a discrete variable with 4 ME categories. The loss of the identification module is calculated by

$$\mathcal{L}^i = -\frac{1}{|\Omega_m|}\sum_{t\in\Omega_m}\mathbf{y}_t^i \log \hat{\mathbf{y}}_t^i \quad (2)$$

4.3.3 The contrastive module

The contrastive module roughly corresponds to the *frame-by-frame* observation. In the *frame-by-frame* observation, the coder will repeatedly watch several consecutive frames and contrastive subtle differences to determine the onset (or offset). This module is a supervised contrastive learning module inspired by [58].

In contrastive module, to better contrast subtle differences between frames, we normalize the $\mathbf{f}$ by

$$\mathbf{f}^c = \frac{\mathbf{f}}{\|\mathbf{f}\|_2} \quad (3)$$

where $\|\mathbf{f}\|_2$ denotes the 2-norm of $\mathbf{f}$.

In order to measure similarity between two normalized feature vectors $\mathbf{f}_i^c$ and $\mathbf{f}_j^c$, we introduce a similarity function denoted as $\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)$. In this paper, we use the inner (dot)

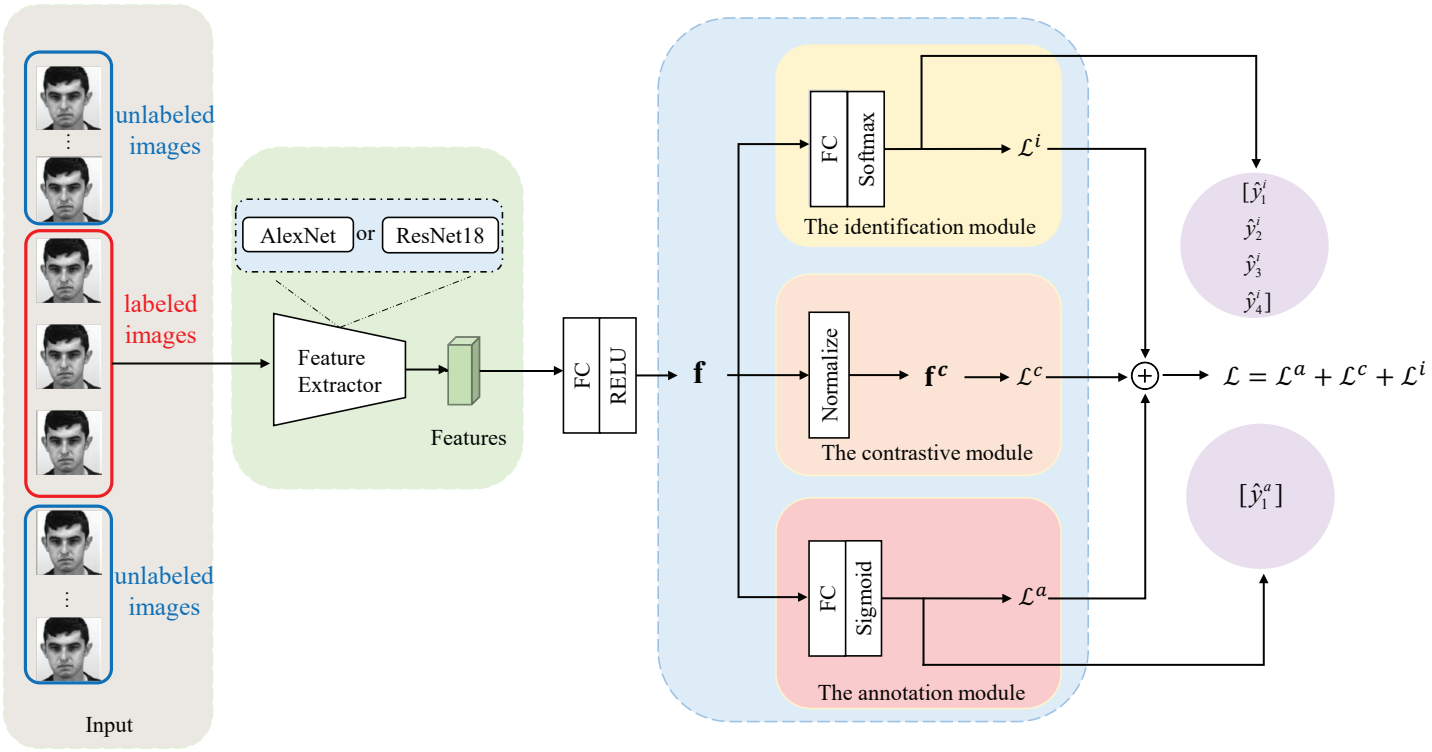


Fig. 4. The architecture of Micro-Expression Contrastive Identification Annotation Network (MECIA-Net)

product of two vectors $\mathbf{f}_i^c$ and $\mathbf{f}_j^c$ as the similarity function, namely

$$\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c) = \mathbf{f}_i^c \cdot \mathbf{f}_j^c \tag{4}$$

If x_i and x_j both are ME (or non-ME) frames, i.e, $i, j \in \Omega_m$ (or Ω_n), we desire the value of $\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)$ is as large as possible. Meanwhile, we also desire the similarity between any two normalized features is as small as possible.

For any ME frame, we have

$$\mathcal{L}_m^c = \sum_{i \in \Omega_m} \frac{-1}{|\Omega_m(i)|} \log \sum_{j \in \Omega_m(i)} \frac{\exp(\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)/\tau)}{\sum_{l \in \Omega_l(i)} \exp(\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_l^c)/\tau)} \tag{5}$$

where $\Omega(i)$ is a set of all elements in Ω except i. And $\tau \in \mathbb{R}^+$ is a scalar temperature parameter.

For any non-ME frame, similarly, we have

$$\mathcal{L}_n^c = \sum_{i \in \Omega_n} \frac{-1}{|\Omega_n(i)|} \log \sum_{j \in \Omega_n(i)} \frac{\exp(\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)/\tau)}{\sum_{l \in \Omega_l(i)} \exp(\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_l^c)/\tau)} \tag{6}$$

The contrastive loss function can be expressed as follows:

$$\mathcal{L}^c = \mathcal{L}_m^c + \mathcal{L}_n^c \tag{7}$$

Fig. 5 shows the relationship between the similarity $\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)$ and the loss $\mathcal{L}^c$. Suppose $\mathbf{f}_i^c$ is fixed. To a certain extent, the similarity $\mathrm{s}(\mathbf{f}_i^c, \mathbf{f}_j^c)$ can be regarded as the distance between the positions of the two frames x_i and x_j in the video. When x_j is close to x_i, x_j is called the *easy* ME. When x_j is far to x_i, x_j is called the *easy* non-ME. When x_j is close to the onset (or offset), x_j is called the *hard* ME/non-ME. The loss gradient contributions from *hard* ME/non-ME are large while those for *easy* ME/non-ME are small [58]. So, the loss function punishes the *hard* ME/non-ME more severely.

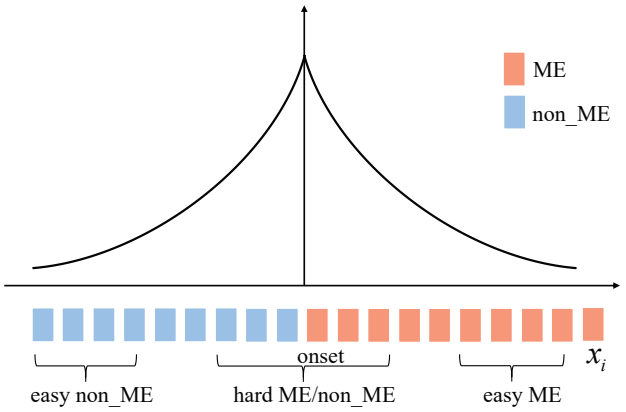


Fig. 5. The contribution of the *hard* ME/non-ME to the loss of the contrastive module.

The total loss function of MECIA-Net is

$$\mathcal{L} = \mathcal{L}^a + \mathcal{L}^c + \mathcal{L}^i \tag{8}$$

Since we have no strong priors to guide the distribution of weights among the three modules, we adopted a straightforward solution by assigning them equal importance. Moreover, this setup was validated by experienced coders, who recognized that the three modules—identifying whether the expression is a micro-expression, recognizing its emotion category, and comparing adjacent frames to determine onset/offset frames—are all equally critical for accurate annotation.

4.4 Frame Inference

In the previous subsection, we use labeled frames to train a MECIA-Net, which has two outputs $\hat{y}_t^a$ and $\hat{\mathbf{y}}_t^i$ for each unlabeled frame x_t. For convenience, we refer to $\hat{y}_t^a$ and $\hat{\mathbf{y}}_t^i$ collectively as $\hat{y}_t$. The two outputs are used to determine whether x_t can be labeled as a new ME or non-ME frame. Here, we give some definitions.

Definition 1. $\hat{y}_i$ and $\hat{y}_j$ are outputs of MECIA-Net for frames x_i and x_j, respectively. If $\hat{y}_i \doteq \hat{y}_j$, we say that the relationship between x_i and x_j is the *homogeneity* on $\hat{y}$, denoted as $x_i \overset{\hat{y}}{\simeq} x_j$.

Homogeneity represents that the frames we infer exhibit consistent scoring patterns across different modules. For the annotation module, specifically, if $\hat{y}_i^a$ and $\hat{y}_j^a$ are both greater than 0.5 or both smaller than 0.5, $\hat{y}_i^a \doteq \hat{y}_j^a$ holds, i.e., we regard these two frames as homogeneous under the annotation module. This ensures that adjacent frames align in their assessment of whether a ME is present. For the identification module, if $\mathrm{I_m}(\hat{\mathbf{y}}_\mathrm{i}^\mathrm{m}) = \mathrm{I_m}(\hat{\mathbf{y}}_\mathrm{j}^\mathrm{m})$, $\hat{\mathbf{y}}_i^m \doteq \hat{\mathbf{y}}_j^m$ holds, where $\mathrm{I_m}(\mathbf{y})$ denotes the index of the maximum elements in the vector $\mathbf{y}$. In other words, if both frames share the same index for their maximum identification score, these two frames are considered homogeneous under the identification module. This ensures that adjacent frames consistently identify the same ME category.

ME intervals must maintain homogeneity in both the annotation and identification modules, since we need consistent evidence both for the existence of a ME and its specific category. In contrast, intervals determined to be non-ME do not require consistency in the identification module (as there is no ME category assigned), and hence only need to maintain homogeneity under the annotation module.

Definition 2. $x_i \overset{\hat{y}}{\Leftrightarrow} x_j$ denotes that the relationship between x_i and x_j is the *connectivity* on $\hat{y}$. If $x_i \overset{\hat{y}}{\Leftrightarrow} x_j$ $(i < j)$ holds, both of the following conditions need to be met:

1) $x_i \overset{\hat{y}}{\simeq} x_j$
2) $x_k \overset{\hat{y}}{\Leftrightarrow} x_j$, where $k \in (i+1, j)$.

when $i > j$, $k \in (j, i-1)$ in the second condition.

Definition 3. The frame x_t is called the *anchor frame*, only if x_t is the labeled frame and at least one of x_{t+1} and x_{t-1} is unlabeled frame.

We will find all anchor frames and then infer frames from each anchor frame. Based on the anchor frame x_a, we infer frames toward the left with the interval r. If x_{a-i} $(i = 1, 2, \ldots, r)$ is an unlabeled frame and that x_{a-i} and x_a are connectivity on $\hat{y}$, x_{a-i} is annotated as the same label (s) with the anchor frame x_a. The inference is described in algorithm 2. We also use Fig. 6 to show the process. The frame inference toward the right is similar.

Algorithm 2: Frame inference

Input: the index of the anchor frame a and the extended interval r
Data: the long video $\mathbf{V} = \{x_1, x_2, \ldots, x_T\}$

```
i ← 0;
while i < r do
    i ← i + 1;
    if x_{a−i} ⇔^ŷ x_a and y^a_{a−i} = −1 then
        y_{a−i} ← y_a
    end
end
```

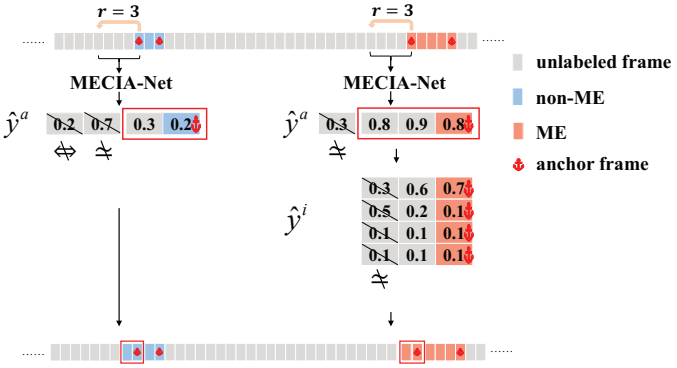


Fig. 6. An example of the frame inference. Blue frames (non-ME) require only the annotation module for expansion, i.e., we only use $\hat{y}^a$ to determine whether the inferred frame can be labeled. For example, the leftmost frame (score 0.2) is excluded due to a lack of connectivity with its adjacent frame (score 0.7), which is also excluded for failing homogeneity with its anchor frame. Red frames, representing MEs, must satisfy both annotation and identification modules. For instance, the leftmost frame (score 0.3) fails homogeneity with its anchor frame (score 0.8). Additionally, the sequence [0.3, 0.5, 0.1, 0.1] is excluded due to a mismatch in category indices, as determined by the identification module.

5 Experimental Results

5.1 Database

There are two ME databases containing long videos: CAS(ME)^2 [11] and SAMM Long Videos (SAMM) [59]. Since the frame rate of CAS(ME)^2 is 30 FPS (frame per second), the number of available frames for a ME clip is much smaller than that in SAMM whose frame rate is 200 FPS. Therefore, to better explore the ME-KFI problem, we perform the experiments mainly on SAMM database. SAMM consists of 147 long videos, including 159 MEs in the long videos. CAS(ME)^2 consists of 98 long videos, including 57 MEs in the long videos. The index of onset, apex, and offset of ME clips are all manually annotated in the two ME databases.

5.2 Evaluation Metrics

The ME-KFI problem is evaluated from two different level localizations: clip-level localization and frame-level localization. For the clip-level localization, the Intersection over Union (IoU) is used for evaluating the ME-KFI problem. IoU measures the overlap ratio of the inferred interval (denoted as I_{inf}) and the GT interval (denoted as I_{GT}). Its formula is shown as follows:

$$\mathrm{IoU} = \frac{I_{inf} \bigcap I_{GT}}{I_{inf} \bigcup I_{GT}} \tag{9}$$

For the frame-level localization, we also use F1 scores to evaluate the ME-KFI problem. The formula is shown as

follows:

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

where TP is the true positive and indicates the case that the inferred ME frame lies in the GT interval (e.g., the frame c in Fig. 7b). FP is the false positive and indicates the case that the inferred ME frame lies outside the GT interval (e.g., the frame a in Fig. 7b). FN is the false negative and indicates the case that the inferred non-ME frame lies in the GT interval (e.g., the frame d in Fig. 7b).

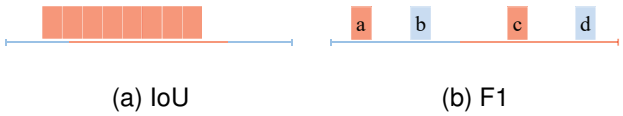


Fig. 7. Two different evaluation metrics

The clip-level localization only involves ME frames. On the other hand, the frame-level localization involves ME and non-ME frames.

These two evaluations are designed to assess the method's performance more comprehensively at two levels, global and local, respectively. Specifically, first, our evaluation draws on the current evaluation criteria in ME analysis, especially the metric used in the ME spotting task of three consecutive ME Grand Challenges (MEGC2019, 2020, and 2021) [60]–[62] is based on clip-level evaluation. Second, since manual annotation is not always accurate at the frame level, the clip-level evaluation is relatively robust. On the other hand, the frame-level evaluation is more detailed and makes it possible to validate our final target, determining the onset and offset frames of ME clips.

5.3 Experimental Setup

In a long video, there are only a few ME clips that have very short duration. ME-KFI only needs to infer the onset and offset of ME clips. To infer each frame in the long video is not only time-consuming but also unnecessary. The average length of ME clips in the SAMM database is 75.3 frames. For convenience, several sample clips with fixed 150 frames length are edited from long videos in the SAMM database. Each sample clip includes only one ME clip. We make the ME clip as far as possible in the middle of the sample clip when editing[1]. If two ME clips are close to each other in a long video, one ME clip is located to the left of one sample clip, and another ME clip is located to the right of another sample clip (see Fig. 8). 159 sample clips are selected from the SAMM database.

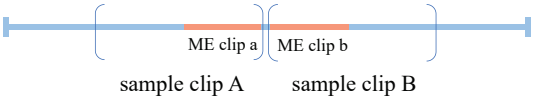


Fig. 8. An example of two ME clips closed to each other

1. For the sake of reproducibility of the experiment, we use the frames at two endpoints of sample clips as initial labeled non-ME frames.

For the CAS(ME)^2 database, the average length of ME clips is 13.6 frames. So, the length of sample clips is fixed as 27 frames. There are three ME clips close together in a long video. The middle one[2] is removed. So, 56 sample clips are selected from the CAS(ME)^2 database.

We use original gray images and resize them to 224×224 as the input of MECIA-Net. We set the learning rate to 0.0001 and the batch to 128, using the adam optimizer. In step 3 of Algorithm 1, MECIA-Net needs to be trained multiple times. The weights of MECIA-Net in the previous time are used to initialize the parameters of MECIA-Net. The model parameters are not stable at the first training, so we set the maximum training epoch to 2000 and the rest to 200. To avoid wasting time during training, the current training is also stopped when the loss value is less than 0.01. Empirically set the temperature parameter mentioned in Eq. (5) and Eq. (6) as 0.07.

Since there are few training data at the beginning of frame inference and the model is unstable, the inference interval r in Section 5.2 should be set smaller to prevent wrong inference from affecting the final result. As the number of inferences increases, the model becomes more and more stable, and the inference interval can be increased to save time. In SAMM, we set r to 2 for the first to fifth inference, 3 for the sixth to tenth, and 5 for the rest. In CAS(ME)^2, we set r to 1 in the whole process of inference because each ME clip has only 27. Each frame inference process calculates IoU. When the value of IoU does not change twice, or the condition until step 5 of Algorithm 1 is satisfied, the algorithm will end.

5.4 Results

5.4.1 Analysis of the manually annotated single frame

First, we analyze the effect of the manually annotated single frame in ME clips on the proposed algorithm. Theoretically, if the annotated frame is precisely the apex of the ME clip, the proposed algorithm will get desirable performance. However, the precise annotating apex is also very time-consuming. So we conjecture whether the algorithm can also obtain relatively desirable performance when the annotated single frame is within the interval where the apex is located.

Each frame in a ME clip is regarded as a vertex in the directed graph. The adjacent two frames x_t and x_{t+1} are connected by a directed edge from x_t to x_{t+1}. We use the k-means [63] algorithm to implement a clustering operation on frames of each ME clip in the SAMM database, and get 3 clusters shown in Fig. 9. If all frames (or vertexes) in a cluster are connected, we call the cluster a connected cluster shown in Fig. 9a. There are 138 ME clips. The three clusters obtained by the k-means on each ME clip are all connected. The middle clusters of other 16 ME clips are connected clusters, while the onset and offset clusters are not connected clusters (see Fig. 9b)[3]. In most cases, the apex falls into the middle cluster. It is desirable that the manually annotated single frame is in the interval spanned by frames in the middle cluster. However, there are still a few cases where the apex falls into the start or end clusters. For convenience, we

2. The ME clips is 's25_disgust2_3'.

3. In the SAMM database, five ME clips do not fit either Fig. 9a or Fig. 9b.

consider the interval spanned by frames in the cluster where the apex is located as the desired interval and call it the *apex interval*.

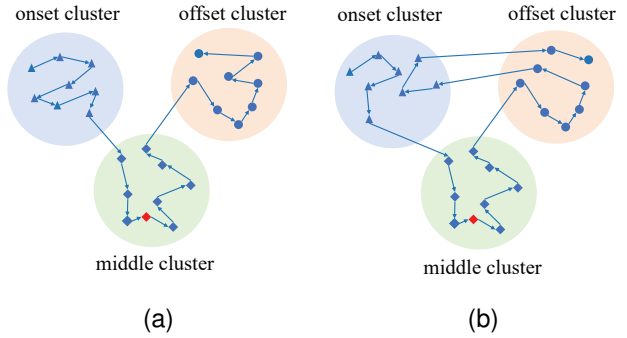


Fig. 9. Schematic of results for apex interval

The position of the manually annotated frame is located at $0\%, 5\%, \ldots, 100\%$ of a ME clip. The IoU is shown in Fig. 10. When the position of the manually annotated frame locates in the interval spanned by frames in the middle cluster, the proposed method gets good performance. And we also can see that the performance of the proposed method with ResNet18 is better than that with AlexNet. This is because ResNet18 introduces residual learning and deep convolutional computation compared to AlexNet, which can effectively reduce the overfitting problem. Moreover, although our dataset is small, the training samples and test samples are adjacent in time, and the feature difference is not very large, so ResNet18 with deeper network layers will not cause the problem of overfitting. On the contrary, it can extract better features.

To demonstrate the accuracy of *apex interval*, we let the coder perform experiments on SAMM. We selected 24 ME clips from the SAMM database, with three videos for each of the eight emotion types in the database. Each subject's face would have to appear only once in the ME clips we chose to remove the influence of facial familiarity on coders. For these ME clips, a coder annotated a single frame, i.e., the frame with the largest change in facial muscle movem ent. The single frames of 19 ME clips were annotated in the calculated *apex interval*.

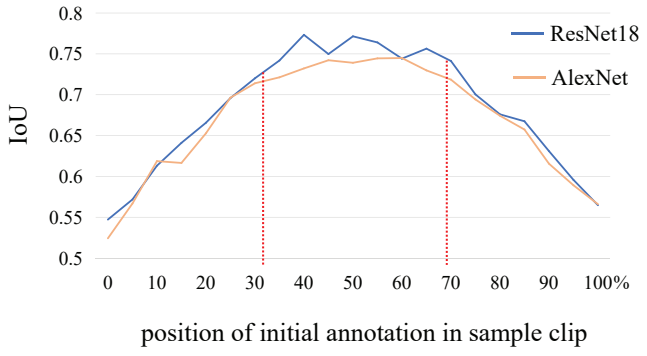


Fig. 10. The effect of the manually annotated single frame in ME clips on the proposed algorithm.

5.4.2 Ablation experiments

We evaluate the performance of MECIA-Net through ablation experiments. In experiments, ResNet18 is used as the backbone. The manually annotated single frame positions are the apex, the left, and the right of the *apex interval*, respectively. Table 3 lists IoUs and F1 scores of various module combinations on two ME databases. The relatively low frame rate of CAS(ME)^2 leads to the fluctuational performance of MECIA-Net. So, The IoUs and F1 scores on CAS(ME)^2 are the averages on ten folds.

Both for the apex and left of the apex interval, performances of two-module combinations are better than that of the annotation module. However, the performances of the three-module combination get the best performance. The experimental results are not desirable for the right of the apex interval. Specifically, MEs usually exhibit representative features in the interval from onset to apex. For instance, Li et al. learned the local temporal pattern from onset to apex to spot ME movements [47]. However, features extracted from apex to offset do not represent MEs well. This is because the latter half of the MEs have variability in movement. For example, they may remain unchanged as the apex state, disappear quickly, or fade slowly.

5.4.3 Results Analysis

We randomly selected 11 locations of the manually annotated single frame in the apex interval for each ME clip in SAMM. So, we experimented with 11 folds and got 11 IoUs. Among them, the maximum, the median, and the minimum are 74.50%, 72.96%, and 71.83%, respectively. The same experiment is conducted on CAS(ME)^2. The maximum, the median, and the minimum are 68.72%, 65.57%, and 57.08%, respectively.

Since the proposed MECIA method is an unprecedented study in the field of intelligent ME analysis, there is no SOTA approach for comparison. Therefore, in order to perform a basic performance evaluation, we compare with a random level of inference. However, it is not appropriate to simply divide the frames in a clip to be inferred into ME and non-ME frames by half. This is because MECIA gradually infers ME and non-ME frames based on known ME frames. Therefore, we assume random frame inference (RFI) based on the above premises. Precisely, the results of RFI are classified into four cases, as shown in Fig11. In particular, as presented earlier, the video length to be inferred in the SAMM database is 150 frames, denoted as $2l$. The average length of the ME clips is 75 frames, which can be denoted as l.

1) The inferred ME clip (I^R_{inf}) is within the GT ME clip, i.e., $I^R_{inf} \subset I_{GT}$. The RFI, in this case, is defined as half of the frames in the GT ME clip are inferred as MEs, and the other half are non-MEs. The IoU of RFI in this case (IoU$_{R1}$), is calculated as follows:

$$\text{IoU}_{R1} == \frac{I^R_{inf}}{I_{GT}} = \frac{\frac{l}{2}}{l} = 0.5 \tag{11}$$

2) The GT ME clip is within the inferred ME clip, i.e., $I_{GT} \subset I^R_{inf}$. The RFI, in this case, is defined as half

TABLE 3
The IoUs and F1 scores on different module combinations. Capital letters A, I, and C indicate the annotation, identification, and contrastive modules. The **bold** indicates the best performance on different module combinations. The underline indicates the best performance among the apex, left, and right.

	SAMM						CAS(ME)^2					
	IoU (%)			F1 (%)			IoU (%)			F1 (%)		
	apex	left	right	apex	left	right	apex	left	right	apex	left	right
A	69.74	68.24	**71.87**	80.11	79.42	**82.75**	60.29	57.56	58.12	73.31	71.85	72.49
A+I	74.31	70.93	70.50	84.63	82.11	81.59	62.59	57.12	**62.05**	75.63	71.94	**75.47**
A+C	73.56	69.15	71.85	83.87	80.62	82.71	61.34	61.98	59.74	73.68	73.77	72.36
A+I+C	**75.09**	**72.17**	70.27	**85.12**	**82.93**	80.53	**66.67**	**66.23**	61.02	**78.45**	**77.71**	73.34

of the frames in each of the GT non ME clips are inferred as MEs, and the other half are non-MEs.

$$\text{IoU}_{R2} == \frac{I_{GT}}{I_{inf}^R} = \frac{l}{l + \frac{l}{2}} = 0.67 \tag{12}$$

3) The inferred ME clip and the GT ME clip overlap, but do not belong to each other. The RFI, in this case, is defined as follows: (1) in the GT ME clip, half of the frames are inferred as MEs, and the other half are non-MEs; (2) in GT non-ME clip with the inferred ME, half of the frames are inferred as MEs, and the other half are non-MEs.

$$\text{IoU}_{R3} == \frac{I_{inf}^R \bigcap I_{GT}}{I_{inf}^R \bigcup I_{GT}} = \frac{\frac{l}{2}}{l + \frac{l/2}{2}} = 0.4 \tag{13}$$

In particular, 10.7% of the videos had ME clips at the beginning or the end. Random inferences for these extreme cases could be included in case 1) and case 3). As listed in Table 4, our proposed MECIA outperforms the RFI in all four cases.

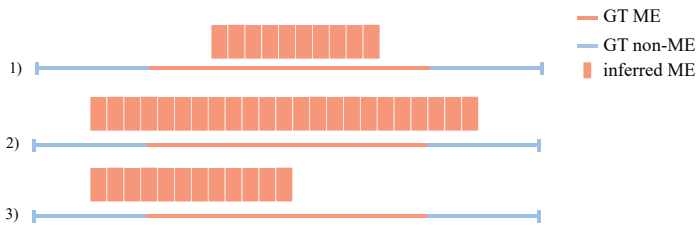


Fig. 11. Three cases for RFI.

TABLE 4
IoU comparison between MECIA and RFI.

IoU (%)	RFI	SAMM	CAS(ME)^2
case1	50.00	74.36	65.16
case2	66.67	68.90	74.76
case3	40.00	74.96	63.37

The inferred ME clips corresponded to the median are partly visualized in Fig. 12. The figure visualizes GT of ME clips and the inferred ME clips. Due to space limitations, we only visualize partly representative video clips[4]. From the figure, we can see that one or both ends of inferred ME clips are out of corresponding GT ME clips in most cases.

4. All ME clips are visualized in the appendix.

We use t-SNE to visualize the 1024-dimensional features $\mathbf{f}$ by giving each data point a location in a two-dimensional map [64]. t-SNE is good at creating a single map that reveals structure at many different scales. This is particularly important for high-dimensional data that lie on several different but related, low-dimensional manifolds, such as continuous frames in ME clip [65]. Fig. 13 shows the t-SNE maps of four sample clips from two databases. The figure shows that data points of ME and non-ME frames can be well discriminated. Compared with data points in CAS(ME)^2, data points in SAMM are more compacted. This also explains why the performance in SAMM is better than that in CAS(ME)^2.

5.4.4 Micro-Expression Recognition Experiments

To verify the effectiveness of the proposed MECIA, we conduct the ME recognition experiments on GT and inferred clips separately. Four ME recognition methods are implemented, including one handcrafted feature extraction methods LBP-TOP [21], and three deep learning methods, i.e., FeatRef [41], RCN-A [66] and STSTNet [42]. Specifically, the input for extracting LBP-TOP is the sequence of ME clips. The input of those three deep learning methods is the optical flow features extracted from onset and apex frames, where the apex frames are achieved by the proposed MECIA.

Before conducting ME recognition experiments, we evaluate MECIA's performance on inferring the apex frame. The metric is defined as follows.

$$differ = \frac{|p_{GT} - p_{inf}|}{l_{GT}} \tag{14}$$

where p_{GT}, p_{inf}, and l_{GT} represents the position of the GT, the inferred apex frame and the length of the GT ME clip, respectively.

As shown in Fig. 14, we enumerate $\hat{y}^a$ of four inferred ME clips and compare the inferred apex frame with the GT apex frame. After calculating, the difference between the inferred apex frame and the GT apex frame is 23.28%.

Considering the data imbalance, the F1 score should be weighted by the number of samples in the corresponding categories before averaging to improve the performance evaluation for ME recognition methods [67]. Thus, we use accuracy (Acc) and weighted F1 (WF1) as the evaluation metrics for ME recognition.

Following [30], [41], we merge eight categories in SAMM into four categories. Specifically, the happy ME is classified into "Positive" category. The anger, sadness, fear, contempt, and disgust ME is classified into the "Negative" category. The surprise and other ME is remained as "Surprise" and

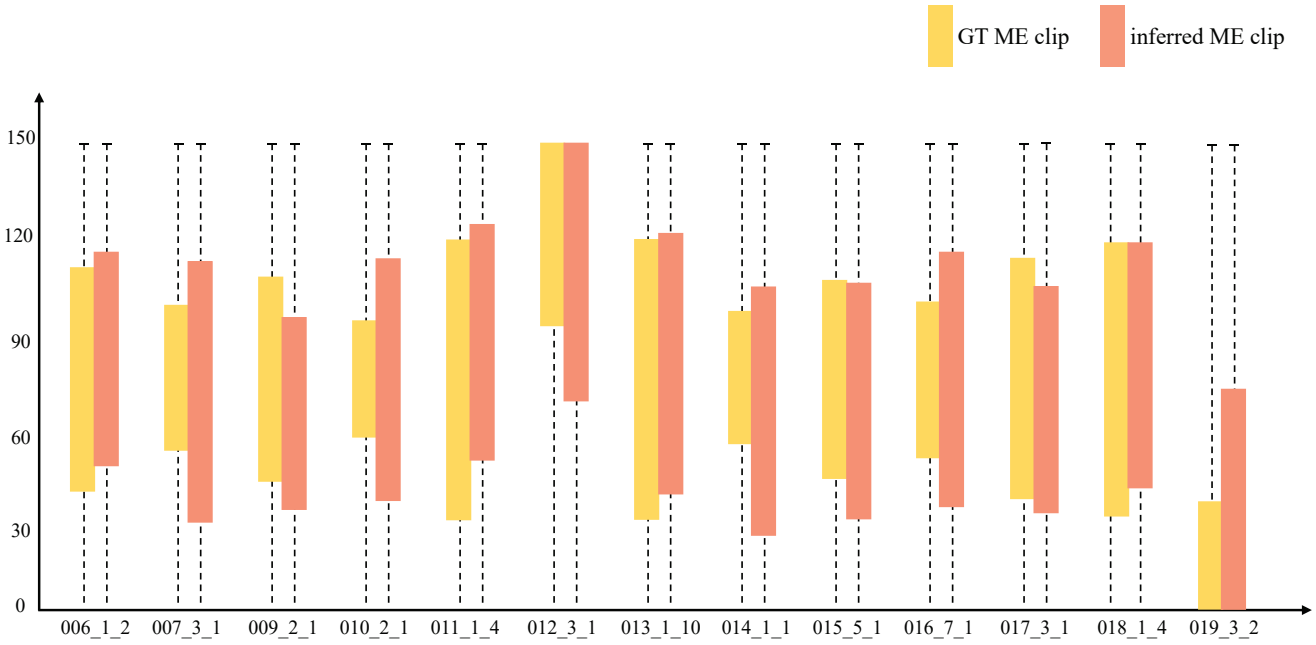


Fig. 12. The Visualization of GT and the inferred clips. The horizontal and vertical coordinates are the name of clips and the position of the frame in the sample clip, respectively.

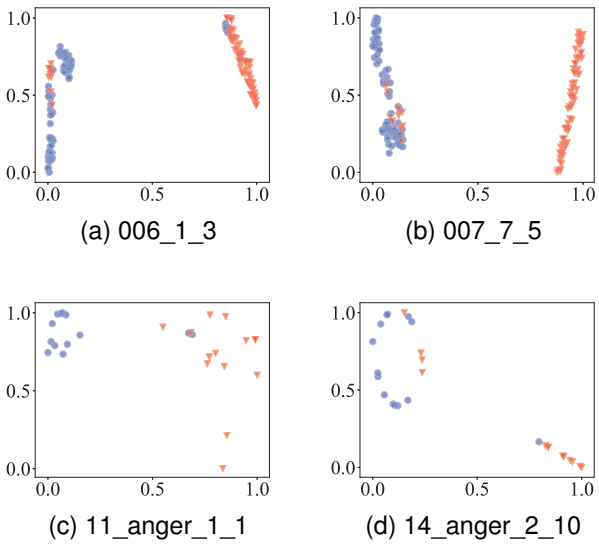


Fig. 13. The t-SNE two-dimensional maps of four sample clips. (a) and (b) are from SAMM. (c) and (d) are from CAS(ME)^2. Titles of sub-figures is the names of ME clips. Orange represents ME and blue represents non-ME.

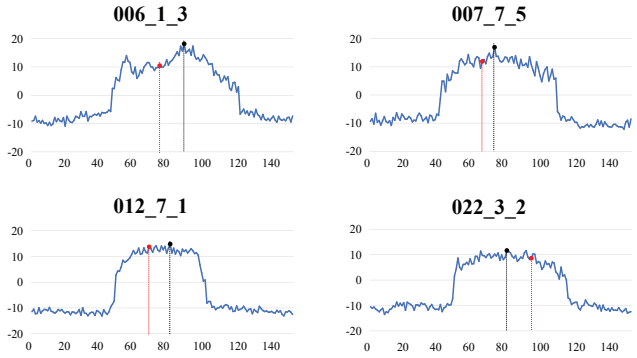
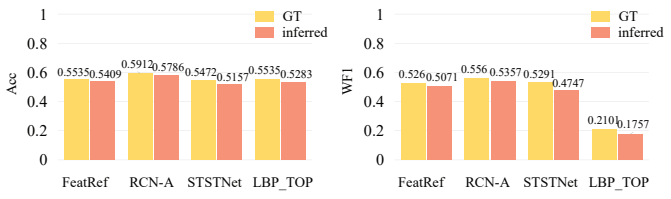


Fig. 14. Comparison of position between inferred apex frame and GT apex frame. The x-axis represents the index of frames in the video clip, the y-axis represents $\hat{y}_1^a$, which is the score of the annotation model after Sigmoid. The red dot represents GT apex frame, and the black dot represents the inferred apex frame.

"Other" categories, respectively. For LBP-TOP, SVM with default parameters is used as the classifier. We also employ the temporal interpolation model (TIM) to normalize the frame number of all the ME clips to 16, and the uniform pattern is applied. The neighboring radius R and the number of the neighboring points P are set as 3 and 8, respectively. For the other three deep learning methods, we extract TV-L1 optical flow from the onset and apex frames and resize them to 28×28 pixels before feeding to the networks. All the ME recognition experiments are evaluated under the Leave-One-Subject-Out (LOSO) protocol.

As shown in Fig. 15, the performance on the four methods of the inferred ME clips are comparable with those of GT. Though there is a slight gap between the performance of our inferred ME clips and GT in the four methods, the inferred ME clips extracted by the proposed method MECIA is more labor-saving as it only needs to annotate one frame manually for each clip, while in GT, all the frames between onset and offset are annotated manually.

Although the results of ME recognition experiments on our inferred clip were worse than those on GT, the difference between the two results is likely to be insignificant, so we



(a) Acc score (b) WF1 score

Fig. 15. Comparison of ME recognition result between GT ME clips and inferred ME clips

analyzed the significance of the difference. We conducted a paired-samples T-test, and the results are presented in Table 5. We can see that the p-values are more than 0.05 (i.e., $p > 0.05$), which indicates no significant differences between the ME recognition result of GT and our inferred ME clips. The term "significant difference" refers to assessing the data difference [68]. In particular, when comparing two data groups, the p-values for the experimental results are less than 0.05, indicating that the data is a "significant difference."

6 Conclusion

The small samples size problem limits the development of ME analysis. Difficulty in data labeling is one of the reasons for the problem. Our psychological experiment demonstrate the difficulty of key frame annotation for human beings. Hence, the paper proposes the ME-KFI problem and gives a simple solution MECIA. The IoU of MECIA is better than that of the random level of inference. We also use the inferred clip and GT to conduct ME recognition experiments based on deep learning and handcrafted feature-based methods. For the deep learning method, the accuracy of GT is a little higher than that of the inferred clip. However, there are no significant differences between them. Moreover, we have communicated with experienced ME coders. They believed our proposed ME-KFI problem and corresponding solution could reduce the 70% labor cost of ME Manual Annotation.

Currently, our approach has only been tested on controlled datasets. In the future, we plan to evaluate it on in-the-wild samples to see how well it generalizes. We also hope to use this method in real data annotation scenarios, reducing the workload for coders. Additionally, we have introduced a new concept for micro-expression key-frame inference. Going forward, we can explore more advanced feature extraction techniques to improve the accuracy of our frame inference modules. These enhancements may further boost performance and help build larger, more reliable micro-expression datasets.

References

[1] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.
[2] P. Ekman, "Lie catching and microexpressions," *The Philosophy of Deception*, p. 118–133, 2009.
[3] 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.
[4] P. A. Stewart, B. M. Waller, and J. N. Schubert, "Presidential speechmaking style: Emotional response to micro-expressions of facial affect," *Motivation and Emotion*, vol. 33, no. 2, p. 125, 2009.
[5] M.-H. Chiu, H. L. Liaw, Y.-R. Yu, and C.-C. Chou, "Facial micro-expression states as an indicator for conceptual change in students' understanding of air pressure and boiling points," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 469–480, 2019.
[6] P. Eckman, "Emotions revealed," *St. Martin's Griffin, New York*, 2003.
[7] 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.
[8] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "A comprehensive survey on facial micro-expression: approaches and databases," *Multimedia Tools and Applications*, vol. 81, no. 28, pp. 40 089–40 134, 2022.
[9] 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 2013)*. IEEE, 2013, pp. 1–7.
[10] 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.
[11] 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*, vol. 9, no. 4, pp. 424–436, 2017.
[12] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[13] 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 2013)*. IEEE, 2013, pp. 1–6.
[14] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic, and G. Zhao, "4dme: A spontaneous 4d micro-expression dataset with multimodalities," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3031–3047, 2023.
[15] 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.
[16] 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.
[17] S. Zhao, H. Tang, X. Mao, S. Liu, Y. Zhang, H. Wang, T. Xu, and E. Chen, "Dfme: A new benchmark for dynamic facial micro-expression recognition," *IEEE Transactions on Affective Computing*, pp. 1–16, 2023.
[18] P. Ekman, "Methods for measuring facial action," *Handbook of Methods in Nonverbal Behavior Research*, pp. 45–90, 1982.
[19] I. Delis, C. Chen, R. E. Jack, O. G. Garrod, S. Panzeri, and P. G. Schyns, "Space-by-time manifold representation of dynamic facial expressions for emotion categorization," *Journal of Vision*, vol. 16, no. 8, pp. 14–14, 2016.
[20] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, 2005.
[21] T. Pfister, X. Li, and G. Zhao, "Recognising spontaneous facial micro-expressions," in *Proceedings of International Conference on Computer Vision*, 2011, pp. 1449–1456.
[22] 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*, vol. 9, no. 4, pp. 563–577, 2017.
[23] A. C. Le Ngo, J. See, and R. C.-W. Phan, "Sparsity in dynamics of spontaneous subtle emotions: analysis and application," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 396–411, 2017.

TABLE 5
Difference test between ME recognition results obtained by GT and inferred ME clips. T-test, freedom, and p-value are commonly used measures in statistic analysis [69].

inferred v.s. GT	Difference in pairs					t-test	freedom	p-value (two sides)
	Mean	Standard Deviation	Standard Error	95% Confidence Interval				
				Lower Limit	Upper Limit			
FeatRef	-0.019	1.250	0.099	-0.215	0.177	-0.190	158	0.849
RCN-A	-0.006	1.133	0.090	-0.184	0.171	-0.070	158	0.944
STSTNet	-0.031	1.285	0.102	-0.233	0.170	-0.309	158	0.758
LBP_TOP	0.130	0.490	0.039	-0.064	0.089	0.324	158	0.747

[24] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 32–47, 2019.

[25] 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.

[26] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, "Sparse MDMO: Learning a discriminative feature for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 254–261, 2018.

[27] 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.

[28] S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 394–406, 2017.

[29] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 147–158, 2022.

[30] X. Xia, Zhaoqiang 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.

[31] 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.

[32] 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.

[33] 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*, pp. 1–13, 2021.

[34] 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.

[35] 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.

[36] 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.

[37] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1037–1043, 2022.

[38] L. Zhang, X. Hong, O. Arandjelović, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1973–1985, 2022.

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

[40] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.

[41] 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.

[42] 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," IEEE, pp. 1–5, 2019.

[43] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, and K. Luu, "Micron-BERT: Bert-based facial micro-expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1482–1492.

[44] 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.

[45] 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.

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

[47] J. Li, C. Soladie, and R. Seguier, "Local temporal pattern and data augmentation for micro-expression spotting," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[48] J. Liu, X. Li, J. Zhang, G. Zhai, Y. Su, Y. Zhang, and B. Wang, "Duration-aware and mode-aware micro-expression spotting for long video sequences," *Signal Processing: Image Communication*, vol. 129, p. 117192, 2024.

[49] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "Spotting micro-movements in image sequence by introducing intelligent cubic-lbp," *IET Image Processing*, vol. 16, no. 14, pp. 3814–3830, 2022.

[50] 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.

[51] 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.

[52] B. Yang, J. Wu, Z. Zhou, M. Komiya, K. Kishimoto, J. Xu, K. Nonaka, T. Horiuchi, S. Komorita, G. Hattori *et al.*, "Facial action unit-based deep learning framework for spotting macro-and micro-expressions in long video sequences," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4794–4798.

[53] W.-W. Yu, J. Jiang, K.-F. Yang, H.-M. Yan, and Y.-J. Li, "Lgsnet: A two-stream network for micro- and macro-expression spotting with background modeling," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 223–240, 2024.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[56] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang, and Y. Liu, "Lightweight ViT model for micro-expression recognition enhanced by transfer learning," *Frontiers in Neurorobotics*, vol. 16, p. 922761, 2022.

[57] M. Verma, S. K. Vipparthi, and G. Singh, "Non-linearities improve originet based on active imaging for micro expression recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.
[58] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
[59] C. H. Yap, C. Kendrick, and M. H. Yap, "SAMM long videos: A spontaneous facial micro-and macro-expressions dataset," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 771–776.
[60] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC2019– 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.
[61] 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.
[62] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li, and S.-J. Wang, "FME'21: 1st workshop on facial micro-expression: Advanced techniques for facial expressions generation and spotting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5700–5701.
[63] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
[64] G. H. Laurens van der Maaten, "Visualizing data using t-SNE," *Journal of Machnine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[65] S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, and X. Fu, "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine," *Neural processing letters*, vol. 39, no. 1, pp. 25–43, 2014.
[66] 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.
[67] M. H. Yap, J. See, X. Hong, and S. Wang, "Facial micro-expressions grand challenge 2018 summary," in *Proceedings of International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 675–678.
[68] J. R. Crawford, D. C. Howell, and P. H. Garthwaite, "Payne and jones revisited: estimating the abnormality of test score differences using a modified paired samples t test." *Journal of Clinical & Experimental Neuropsychology*, vol. 20, no. 6, pp. 898–905, 1998.
[69] J. G. Frederick, *Statistics for the behavioral sciences*. Wadsworth, 2013.

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".

Yu-Han Miao graduated from Jiangsu University of Science and Technology with a master's degree in Computer Science. His research interests include deep learning, computer vision, and micro-expression analysis.

Jingting Li is currently an associate researcher at the Institute of Psychology, Chinese Academy of Sciences (CAS). She received the PhD degree in Signal, Image, Vision from Centrale-Supélec in 2019. She served as the chair of the ACMMM'21, 22, 23 and 24 FME workshop and MEGC Grand 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.

Ling Zhou is currently a assistant professor of Faculty of Information Technology, Macau University of Science and Technology, Macau, China. She received the PhD degree in Jiangsu University in 2021. Her current research interests include deep learning, computer vision, pattern recognition, and facial micro-expression 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.

Mengyi Sun received the B.S. degree in Management Accounting from Beijing Wuzi University, China, in 2023. She is currently pursuing the M.Psy. degree in the Institute of Psychology, Chinese Academy of Sciences. Her current research interests include facial ME, visual psychophysics and neurophysiology.

Xiaolan Fu 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 dean of School of Psychology, Shanghai Jiao Tong University.