# FED-PsyAU: Privacy-Preserving Micro-Expression Recognition via Psychological AU Coordination and Dynamic Facial Motion Modeling

Jingting Li[†1,2], Yu Qian[†1,3], Lin Zhao[1,2], Su-Jing Wang[*1,2]
[1]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
[2]Department of Psychology, University of the Chinese Academy of Sciences, Beijing, 100049, China
[3]School of Computer, Jiangsu University of Science and Technology, Zhenjiang, 212100, China
[†]Equal contribution, [*]Corresponding author (wangsujing@psych.ac.cn)

## Abstract

*Micro-expressions (MEs) are brief, low-intensity, often localized facial expressions. They could reveal genuine emotions individuals may attempt to conceal, valuable in contexts like criminal interrogation and psychological counseling. However, ME recognition (MER) faces challenges, such as small sample sizes and subtle features, which hinder efficient modeling. Additionally, real-world applications encounter ME data privacy issues, leaving the task of enhancing recognition across settings under privacy constraints largely unexplored. To address these issues, we propose a FED-PsyAU research framework. We begin with a psychological study on the coordination of upper and lower facial action units (AUs) to provide structured prior knowledge of facial muscle dynamics. We then develop a DPK-GAT network that combines these psychological priors with statistical AU patterns, enabling hierarchical learning of facial motion features from regional to global levels, effectively enhancing MER performance. Additionally, our federated learning (FL) framework advances MER capabilities across multiple clients without data sharing, preserving privacy and alleviating the limited-sample issue for each client. Extensive experiments on commonly-used ME databases demonstrate the effectiveness of our approach. Our implementation is publicly available at https://github.com/MELABIPCAS/FED-PsyAU.git.*

## 1. Introduction

Micro-expressions (MEs) [9], resulting from conflicts between voluntary and involuntary expressions, reveal true emotions, making them valuable in non-contact emotion detection applications like criminal interrogation and psychological counseling. MEs are characterized by their brief duration, low intensity, localized occurrence, and occasional asymmetry [10]. These traits not only restrict large-scale data collection and annotation but also make it challenging for deep learning networks to effectively model their motion characteristics. Some methods focus on regions of interest (ROIs) [4, 21, 31, 49], while others utilize attention mechanisms [3, 14, 33] and transformer-based architectures [24, 50, 57] to capture salient features. Meantime, MEs are often complex manifestations involving the combination of multiple AUs. In contrast, individual action units (AUs) correspond to specific, anatomically-based facial muscle movements. Focusing on AUs allows us to model more fundamental and consistent facial appearance changes. Certain approaches use AU labels to construct adjacency matrices for graph convolutional networks (GCNs) [24, 39]. Yet, to the best of our knowledge, these methods often apply AU annotations directly without fully exploiting the anatomical structure and coordinated AU dynamics, limiting the network's understanding on ME.

To address these challenges, we leverage the theory that distinct neural pathways control upper and lower facial movements, resulting in two different motion patterns. Based on this, we conduct a psychological study to investigate upper/lower facial AU coordination, providing crucial prior knowledge on structural motion relationships underpinning MEs. Using both psychological priors and data-driven AU patterns, we design a Graph Attention Network (GAT) [47] based on dynamic prior knowledge, applied separately to the upper and lower regions. A final global GAT captures the topological relationships of muscle movements across the face. The resulting feature, combined with full-face optical flow (OF), is input into a subsequent dual-stream network, effectively integrating structural and dynamic information to enhance ME recognition (MER).

Besides, as mentioned earlier, although MEs have valuable applications in many privacy-sensitive contexts, the sensitivity of biometric data makes it impractical to aggre-

gate ME data across clients for training. Meanwhile, the sample size in a single client is often insufficient to train a robust model. Yet, the challenge of enhancing MER performance across diverse clients while preserving privacy remains largely unexplored. To address this, we adopt a federated learning (FL) framework, improving performance by merging and updating model parameters across clients without exchanging local data. This approach preserves data privacy and supports the MER practical deployment.

Overall, our work leverages psychological and data-driven priors to enhance the network's learning of structured ME motion patterns, particularly in limited-sample conditions. Additionally, we introduce FL to advance MER in real-world scenarios. The contributions of this paper are:

- We conduct a psychological experiment to derive AU coordination between the upper and lower facial regions, providing structural priors for ME and facial expression analysis in computer vision.
- We design a local-to-global GAT network that integrates psychological priors and data-driven AU patterns in the upper and lower facial regions, improving the network's ability to learn both local and global structural facial dynamics. This feature, combined with global OF feature in a dual-stream network, enhances MER performance.
- We employ FL in MER to improve data privacy and overcome small sample size challenges by aggregating features from distributed data-protected sources.

## 2. Related Works

### 2.1. Micro-expression Recognition

In recent years, MER has gained increasing attention from computer vision researchers. Traditional machine learning methods primarily rely on handcrafted feature extraction. For instance, Local Binary Pattern (LBP) [43], effectively utilizes local texture features, has inspired various improvements [18, 49]. Furthermore, OF [26] has been explored for capturing object motion. Liu et al. [38] developed the Main Directional Mean OF (MDMO) method by aligning faces in the OF domain, while Liong et al. [34] added optical strain features for MER. However, traditional approaches depend heavily on manually crafted features, limiting both accuracy and generalizability. Meanwhile, deep learning methods, with robust feature extraction capabilities, can automatically learn multi-level representations to capture subtle ME changes. The availability of several spontaneous ME databases, such as CASME II [51], SAMM [6], CAS(ME)$^3$ [27], DFME [59]and others [1, 30, 32, 44, 54], has accelerated deep learning advancements in MER. Inspired by Liong et al.'s work [35] on sequence redundancy, many studies [4, 13, 50] now use keyframes (onset and apex) as input, reducing model parameters while achieving excellent results.

Despite advancements in preprocessing and feature extraction, deep learning alone has yet to achieve high MER accuracy. As MEs stem from physiological muscle activity, incorporating prior knowledge offers meaningful context. FACS [10], which defines expressions through AUs, provides a basic theoretical framework for MER. Recent studies explore AU-ME relationships to enhance MER performance. Lei et al. [24] applied GCN [21] to model AU co-occurrences, especially around the eyebrow and mouth regions. Xie et al. [53] introduced an AU-assisted GCN, with modules for AU feature extraction and ME image generation. Wang et al. [48] used transformers to learn dynamic AU adjacency, boosting model generalization. These networks typically rely on AU labels from databases, without fully integrating structured facial muscle movements, AU interrelations, or their links to expressions. In this work, we refine structured ME feature extraction using a local-to-global GAT network, which integrates priors on facial AU coordination from psychological experiments with data-driven movement patterns.

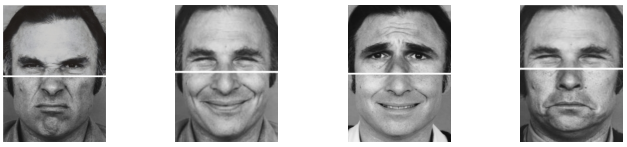
### 2.2. Federated Learning

To assess MER across databases, ME Grand Challenge (MEGC) 2018 [56] introduced the Holdout-database Evaluation and Composite Database Evaluation tasks, utilizing the CASME II and SAMM databases. MEGC 2019 [45] expanded the CDE task to include the SMIC databases. While research on MEs predominantly utilizes these public datasets, existing methodologies do not address the privacy concerns that arise in practical applications.

FL [41] could addresses ME data scarcity and privacy concerns through collaborative training. It includes two main phases: local training and global aggregation. Many existing FL works for expressions, while insightful, are often not open-source or use highly customized frameworks (e.g., non-random client data allocation [12], specialized parameter aggregation [11]) making standard methods more suitable for implementation and comparisons in our study. For instance, McMahan et al. [41] proposed FedAvg, a basic aggregation method that uses weighted averaging based on each client's data volume. Li et al. [28] introduced FedProx by adding a regularization term to the client loss function, aligning updates more closely with the global model. However, the one-global model-fits-all approach ignores client data diversity. To solve this, our approach extends FedProx by distributing a personalized global model to each client.

## 3. Proposed Psychological Study on AU Incoordination

This study employed a behavioral experiment to investigate individuals' cognitive responses to AU incoordination in the upper and lower facial regions.



(a) AU 4+9 (b) AU 6+12 (c) AU 4+12 (d) AU 6+15

Figure 1. Examples of AU composite images. (a) and (b) are coordinated composites. (c) and (d) are uncoordinated composites.

The study included 30 college participants (20 female, *Mean*= 22.22 years), a sample size considered adequate for statistical reliability under the Central Limit Theorem [23]. All procedures followed the Declaration of Helsinki and were approved by the institutional review board of the Institute of Psychology, Chinese Academy of Sciences.

This behavioral experiment employed a within-subjects design. The independent variable was the type of AU composite image (coordinated vs. uncoordinated AU composites), as shown in Fig. 1. The face images were segmented anatomically. The dividing line was consistently present in all stimuli, serving as a control variable and would not influence the experiment's outcomes. To control for individual facial differences in emotion recognition, the original images used for composites were sourced from the same person [10]. Seventy-two composite images were generated from upper and lower facial AUs across six basic emotions, with each emotion represented by six coordinated and six uncoordinated expressions.

The dependent variables are accuracy rates and scores. Specifically, participants made two judgments for each composite image: whether it was coordinated or uncoordinated, and the degree of coordination or incoordination.

Statistical analysis from this study showed that participants were significantly more accurate in recognizing coordinated AU composite images than uncoordinated ones (Please see Supplementary Materials (Suppl.) for more experimental details and statistical results). The study further highlighted specific AU combinations, such as AU4 and AU12, that participants quickly identified as uncoordinated, whereas combinations like AU6 and AU12 were rapidly recognized as coordinated. This suggests that these uncoordinated AU pairings may inherently convey emotional incoordination. Besides, the coordination patterns between specific AUs reveal systematic facial action structures. These patterns help the network distinguish both coordinated and uncoordinated facial movements, thus enhancing its ability to identify ME features more effectively.

# 4. Our Proposed FED-PsyAU Framework

We propose the FED-PsyAU framework (Fig. 2), which embeds a psychology-driven MER model within a federated aggregation module. The model features a hierarchical structure that learns from local ROIs to AU groups and finally to global facial regions, capturing dynamic topology while minimizing redundancy. It leverages a multi-scale InceptionNet to extract and fuse global OF with structural facial features. The federated module enables privacy-preserving collaborative training to boost MER performance on each client.

## 4.1. Localized ROI Modeling

Since MEs are manifested by muscle movements in key facial regions (e.g., eyes, mouth, nose) [10], we refine the standard 68 landmarks from Dlib [20] to precisely capture these changes. As shown in Fig. 3, we discard 10 less informative outer contour landmarks and add 7 new midpoints to better track movements in the cheek and eyebrow regions, resulting in a final set of 65 keypoints.

Then, onset-apex OF inputs ($128 \times 128$ pixels) are cropped into 65 ROIs, each measuring $5 \times 5$ pixels, centered on selected facial landmarks $\mathbf{p}_k = (x_k, y_k)$, where $k \in 1, 2, \ldots, 65$, denotes the landmark index. To enhance robustness to apex annotation, we also compute OF using a small window of frames around the apex. The ROI features input to the **L**ocal ROI **F**eature **E**xtractor (LFE), denoted as $\Theta_k$, contain the horizontal and vertical components of OF and optical strain. The LFE module consists of three convolutional layers, each followed by batch normalization and ReLU activation. All kernel sizes were $3 \times 3$. The output channels in the first two layers increase from 32 to 64, while the final layer outputs 3 channels. This design helps the LFE capture complex local patterns, with the output features referred to as $\Phi_k = \text{LFE}(\Theta_k)$.

Inspired by Vision Transformer [8], we use a **S**patial **S**tructure **E**ncoder (SSE), i.e., Transformer Encoder with three attention heads to capture diverse dependencies among ROIs, as MEs result from coordinated muscle movements across multiple areas. Precisely, we convert $\Phi_k$, into a token. To retain the positional information of the facial landmarks, a positional embedding $E_k$ is added based on the landmark index to each token, yielding $X_k = \Phi_k + E_k$. We add the output of the three head attention module to the input feature $X_k$ , and then normalize the result using LayerNorm. Finally, after passing through a feed-forward neural network and applying LayerNorm, the final output representation is obtained as $Z_k \in \mathbb{R}^{3\times5\times5}$.

## 4.2. Relationship Modeling among AU Features

Our proposed Relationship Modeling among **A**U **F**eatures (AFR) module consists of two components: an **A**U **F**eature **E**xtractor (AFE) to derive features from AU-specific facial regions, and a GAT-structure to model the topological relationships between these AUs.

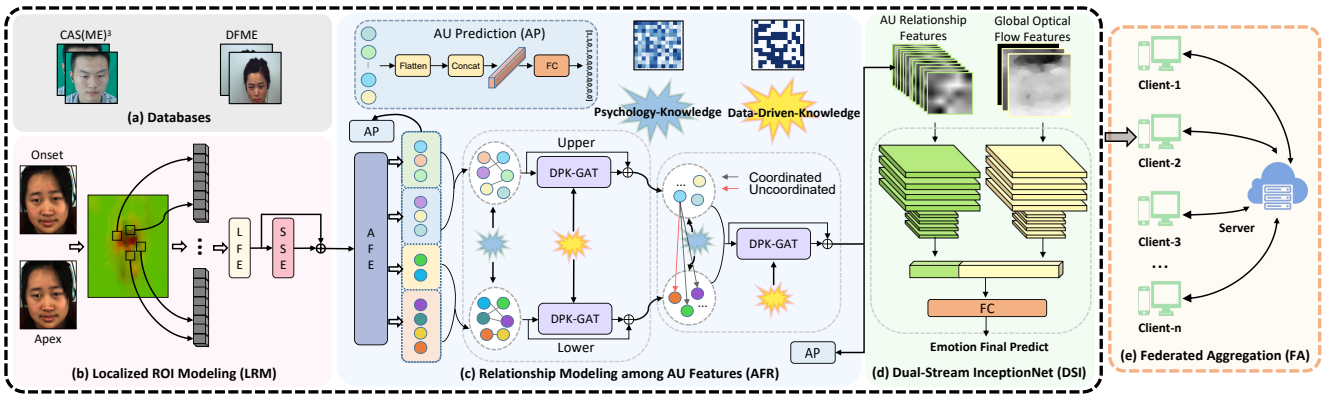



Figure 2. The pipeline of our proposed FED-PsyAU framework, including MER network (black block) and the Federated Aggregation (FA) module (orange block). In MER network, the LRM module consists of the local ROI feature extractor (LFE) and the spatial structure encoder (SSE), which preserves spatial structure information while efficiently extracting ROI features; the AFR module includes the AU feature extractor (AFE) and the dynamic prior knowledge-based graph attention network (DPK-GAT) for effective AU feature extraction and capturing AU dynamic relationships, the DSI module use a dual-stream structure to capture multi-scale facial muscle topology and motion information for MER. Then, the FA module integrates the MER network into the FL framework.

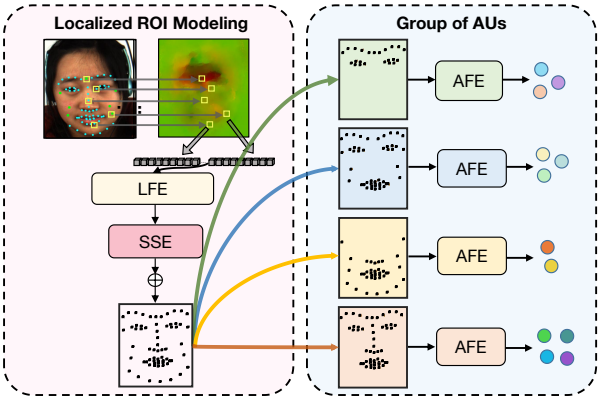



Figure 3. ROI grouping strategy for AU. OF features from each ROI are processed through LFE and SSE to yield high-dimensional representations. ROIs are then grouped by AU movement patterns to create inputs for AFEs, ultimately producing AU node features for the GAT network.

#### 4.2.1. AU Feature Extractor

MEs involve dynamic, coordinated interactions across multiple regions [10] and are characterized by subtle movements that differ from typical macro-expressions. Therefore, as shown in Fig. 3, when grouping ROIs for the AFE module, the AU selection and regionalization are based on: 1) AU frequency in MEs; 2) Grouping by primary/secondary muscle activation areas [7]. Notably, not all AUs employed in Section 3 are applied to MER. We specifically retain only the 12 AUs exhibiting the highest relevance to ME dynamics, as listed in Tab. 1. Detailed occurrences are provided in the Suppl.

The first step we construct in AFE is **G**roup **S**queeze and

Table 1. Selected AUs and Corresponding Regions (Primary and Secondary/ Coordinate). $g$, $N_g$ and $N_{g\_AU}$ represent ROI Group index, corresponding ROI number and AU amount.

| $g$ | AU | Primary | Secondary/ Coordinate | $N_g$ | $N_{g\_AU}$ |
|---|---|---|---|---|---|
| 1 | 1,2,4 | Eyebrows | Eyes | 23 | 3 |
| 2 | 5,6,7 | Eyes | Eyebrows, Cheeks, Mouth | 48 | 3 |
| 3 | 9,10 | Nose, Cheeks | Mouth, Chin | 38 | 2 |
| 4 | 12,14,15,17 | Mouth | Eyes, Eyebrows, Nose | 52 | 4 |

**E**xcitation (GSE). It computes the attention for each group of AUs (ROI group), which makes the model pay more attention to the critical ROIs. Specifically, for each ROI group input $R_g = \{Z_1, Z_2, \ldots, Z_{N_g}\}$, $g \in \{1, 2, 3, 4\}$, the GSE module performs global average pooling on each ROI and computes channel importance through a fully connected (FC) layer. The original features are then preserved through residual concatenation:

$$f_{\text{GSE}}(R_g) = \{W_k^{att} \odot Z_k + Z_k \mid Z_k \in R_g\} \quad (1)$$

where $W_k^{att}$ is the attention weight for each group of ROI, $\odot$ denotes element-wise multiplication along the channel dimension of $Z_k$. Next, $f_{\text{GSE}}(R_g)$ undergoes further feature extraction, reducing from $N_g$ composite features to $N_{g\_AU}$ representative AU features, yielding the output feature $F_{g\_AU}$. Precisely, the process comprises two convolutional layers: a $3 \times 3$ layer that doubles the input dimensions and a $1 \times 1$ layer that restores them to the original size. Both layers are followed by Batch Normalization and ReLU activation for improved stability and representation.

#### 4.2.2. AU Relationship Modeling

**Prior Knowledge:** To model AU dynamic relationships, we integrate two forms of prior knowledge: an "intrinsic" prior

from psychological principles and an "extrinsic" prior from data-driven patterns. The intrinsic prior, derived from our psychological findings on AU coordination (Sec. 3), provides structural guidance by defining the initial GAT adjacency matrix $A$. The extrinsic prior, based on statistical AU co-occurrence patterns in the data, offers data-driven refinement that theoretical models may overlook. This extrinsic prior forms a prior attention matrix $D$ to guide weight allocation among AU nodes. By fusing this foundational psychological structure with real-world dynamics, our model achieves superior adaptability and generalization (see Suppl. for matrix details).

**Facial Segmentation:** As presented in Sec. 3, the upper and lower facial regions, controlled by different neural pathways, exhibit distinct muscle movement patterns. AUs and their associated primary muscles in the upper and lower face can exhibit coordination (potential co-occurrence) or lack of coordination (infrequent simultaneous appearance). Leveraging these intra- and inter-regional interactions, we model AU relationships for the upper face, lower face, and whole face. Consistent with the experimental setup in Sec. 3, we segment the face into upper and lower parts based on the orbicularis oculi muscle. Specifically, following the primary occurrence regions of 12 listed AUs, the AUs in the first two groups from Tab. 1 are classified as upper face AUs, while those in the last two groups are classified as lower face AUs.

$$F_{AU}^{Upper} = F_{1\_AU} \cup F_{2\_AU}, F_{AU}^{Lower} = F_{3\_AU} \cup F_{4\_AU} \quad (2)$$

We input the AU features from the upper and lower face: $F_{AU}^{Upper}$, $F_{AU}^{Lower}$ into our proposed **D**ynamic **P**rior **K**nowledge GATs (DPK-GATs) in parallel (Details are presented in the next part of this subsection). These networks use psychology prior knowledge to build AU adjacency relationships and data-driven priors to guide attention learning, producing outputs $h^{Upper}$ and $h^{Lower}$, which are combined into $h^{Local}$ for the next stage.

In the whole-face DPK-GAT, we construct the network similarly but set adjacency values within the same region to zero, ignoring intra-regional AU relationships. By learning cross-regional topological structures and statistical patterns, we obtain global AU relationship features, denoted as $h^{Global}$, as illustrated in Fig. 4.

**DPK-GAT:** We model dynamic AU node relationships combing GAT and the prior knowledge. In particular, for the input feature belonging to $\{F_{AU}^{Upper}, F_{AU}^{Lower}, h^{Local}\}$, a linear mapping projects it to a higher-dimensional space as $f$. The similarity score $e_{ij}$ between AU nodes $i$ and $j$, is calculated using the attention weight vector $\vec{a}$ and concatenated node features $f_i$ and $f_j$:

$$e_{ij} = \begin{cases} \text{LeakyReLU}\left(\vec{a}^T[f_i||f_j]\right), & \text{if } A_{ij} > 0 \\ -\infty, & \text{if } A_{ij} = 0 \end{cases} \quad (3)$$

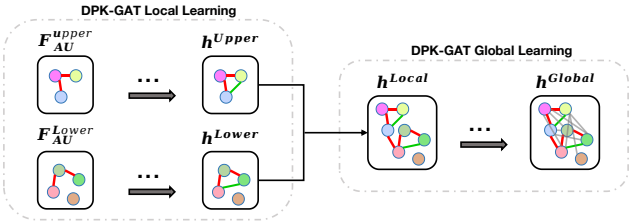



Figure 4. DPK-GAT from facial segmentations to global face. First, intra-regional AU node relationships are learned within the upper and lower facial regions. Second, cross-regional relationships are modeled to integrate interactions between these regions.

Here, only neighboring nodes defined by the psychology-derived adjacency matrix $A$ are considered.

The final attention $\alpha_{ij}$ is calculated by combining the data-driven prior attention $D_{ij}$ and the self-attention, obtained by the softmax operation applied to $e_{ij}$:

$$\alpha_{ij} = (1-\beta) \cdot \text{softmax}(e_{ij}) + \beta \cdot D_{ij} \quad (4)$$

where, $\beta$ is the dynamic prior attention weight, updated based on the training rounds. Along with a residual concatenation, the final node representation $h_i$ is aggregated as:

$$h_i = \text{ELU}(\textstyle\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \cdot f_j) + f_i \quad (5)$$

where $\mathcal{N}(i)$ denotes the neighboring nodes of $i$. In our implementation, we employ three attention heads and a two-layer structure, enabling the extraction of complex AU dependencies.

### 4.3. Dual-Stream-InceptionNet

Muscle movements vary widely across facial regions, making multi-scale information essential for the MER task. InceptionNet's core concept is leveraging parallel convolutional kernels to extract multi-scale features, enhancing the network's ability to capture fine details and overall structure [46]. We build on this by using a dual-stream InceptionNet as the ME feature extractor. One stream processes the AU relationship features $h^{Global}$, while the other handles global OF features $\Theta$. This dual-stream design allows the model to capture the topological relationships of facial muscles through AUs and holistic motion via OF. The output tensors from both streams are flattened, concatenated, and fed into a FC layer for ME prediction.

### 4.4. Federated Aggregation

Traditional centralized learning consolidates all data on a single server, enhancing model accuracy and generalization but neglecting privacy concerns in MER. FL could address this by training models locally and sharing only parameters, protecting data privacy and retaining each client's unique data characteristics.

FedProx [29] is a FL approach designed to address data heterogeneity among clients. It incorporates proximal terms during model aggregation to balance global and client-local learning, improving both performance and convergence speed. We extend FedProx by altering the server's processing and distribution strategy during aggregation, which we call **P**ersonalized **FedProx** (P-FedProx). Specifically, P-FedProx initializes each client's model $W_i^t$ in round $t$ by combining its previous model $\mathring{W}_i^{t-1}$ with models from other clients $\mathring{W}_j^{t-1}$.

$$W_i^t = \theta \cdot \mathring{W}_i^{t-1} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \omega_j \cdot \mathring{W}_j^{t-1}, \text{ where } \sum_{\substack{j=1 \\ j \neq i}}^{n} \omega_j = 1 - \theta \quad (6)$$

where weight $\omega_j$ is proportional to $j$th client's data size. For $i$th client, $\theta$ is set to 0.9 (optimal experimental result, see more detail in the Suppl.)

### 4.5. Loss Function

**AU Prediction Loss:** To enhance MER, our network models the topological relationships and statistical patterns of AUs, making AU prediction an auxiliary task. The prediction is utilized in two stages: one enhances the AU feature extraction performance of the AFE module, and the other improves the GAT network's ability to learn AU topological structures. AU recognition is formulated as a multi-label classification problem, with the Binary Cross-Entropy Loss for each AU label defined as:

$$L_{\text{A}} = \frac{\sum_{j=1}^{M} \left[ y_j \cdot \log(\sigma(\hat{y}_j)) + (1 - y_j) \cdot \log(1 - \sigma(\hat{y}_j)) \right]}{-M} \quad (7)$$

where $M$ is the total number of AUs, $y_j$ , $\sigma(\hat{y}_j)$ and $\hat{y}_j$ represent respectively the ground truth label (0 or 1), the logit output and prediction after applying the sigmoid function for the $j$-th AU. The loss function is consistent for both stages, i.e., $L_{\text{A}}^{\text{AFE}}$ and $L_{\text{A}}^{\text{GAT}}$.

**MER Loss:** The Cross-Entropy Loss function is utilized for emotion classification. Thus, our overall MER loss is composed as follows:

$$L_{\text{MER}} = \alpha_1 \cdot (-\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i)) + \alpha_2 \cdot L_{\text{A}}^{\text{AFE}} + \alpha_3 \cdot L_{\text{A}}^{\text{GAT}} \quad (8)$$

where, $N$ represents the number of classes, $y_i$ and $\hat{y}_i$ are the one-hot encoded ground truth label and the predicted probability for class $i$, $\alpha_1$, $\alpha_2$ and $\alpha_3$ are hyperparameters, set to 0.2, 0.8, and 0.8 to prioritize MER loss while balancing AU recognition (See parameter analysis in Suppl.).

**Federated Loss:** During federated training, the local client's loss function includes the MER loss $L_{\text{MER}}$ and a proximal term:

$$L_{\text{FL}} = L_{\text{MER}} + \frac{\alpha_4}{2} \cdot ||W - W_t||^2 \quad (9)$$

where $W_t$ denotes the global model in the server's $t$-th round of distribution, and $W$ represents the current model of the client. $\alpha_4$ is the hyperparameter for the weight of the proximal term.

## 5. Experiments

### 5.1. Configuration

**Databases:** We use two recently published large scale and challenging ME databases: CAS(ME)$^3$ and DFME. CAS(ME)$^3$-Part A offers 860 samples from 100 participants. DFME includes three subsets with (participants, samples): A (72,1,118), B (92,969), C (492,5,439). For Apex-first-frame scenario, we compute the OF between the apex and mid-frame towards the offset.

**Pre-training:** To enhance the model's ability to perceive facial motion patterns, we first pre-trained the AU module (LRM+AFR) on the DISFA [40] and CK [19] datasets. The effect of this pre-training is detailed in the Suppl.

**MER Experiment Settings:** The evaluation follows standard protocols on two datasets. On CAS(ME)$^3$, we perform 3-class classification (positive, negative, surprised) [45] with Leave-One-Subject-Out (LOSO) validation. On DFME, we conduct 7-class classification (Happiness, Surprise, Disgust, Sadness, Anger, Fear, and Contempt) using scale-adapted 10-fold cross-validation [59]. To address class imbalance, performance is measured by the UF1 and UAR. Experiments were implemented in PyTorch 1.13.0 and run on five NVIDIA GeForce RTX 4090 GPUs.

**Federated Experiment Settings:** The DFME and CAS(ME)$^3$ datasets are randomly split into 5 and 2 local clients, respectively, with each client assigned an equal number of subjects. However, variations in micro-expression (ME) counts per subject result in differing ME data distributions and quantities across clients, mirroring real-world data heterogeneity (See detailed data distributions in the Suppl.). Traditional LOSO and k-fold cross-validation can leak training samples in a federated setting, so we employ a random partition strategy with 70% of samples for training and 30% for testing, averaging results over 10 tests to minimize randomness. Furthermore, we assume an ideal FL scenario with no client dropouts and equal local updates per client. In FedAvg [41] and FedProx [28], the server assigns an identical global model to all clients each round; differently, FedProx adds a proximal term to reduce model bias in local training. Our P-FedProx enhances this by assigning tailored initial models, with 90% ($\theta$) weight on each client's own model and 10% on others'. Additionally, we compare against recent personalized FL methods: FedRep [5] uses shared representations with personalized heads, ELLP [61] employs parameter decoupling with Fisher Information weighting, and FedAS [55] applies parameter alignment and client synchronization. Unlike these approaches requiring architectural changes or complex procedures, P-FedProx achieves personalization through simple weighted initialization.

Table 2. SOTA methods comparison on DFME and CAS(ME)$^3$

| Category | Method | Year | DFME | | CAS(ME)$^3$ | |
|---|---|---|---|---|---|---|
| | | | UF1 | UAR | UF1 | UAR |
| *3D-CNN* | I3D [2] | 2017 | 0.2923 | 0.3058 | – | – |
| | R3D [17] | 2018 | 0.2164 | 0.2313 | – | – |
| *Hand-crafted* | LBP-TOP [49] | 2014 | 0.2336 | 0.2653 | – | – |
| | MDMO [37] | 2015 | 0.2489 | 0.2939 | – | – |
| *Deep Learning* | OFF-ApexNet [15] | 2019 | 0.2386 | 0.2806 | – | – |
| | AlexNet [22] | 2012 | – | – | 0.2570 | 0.2634 |
| | STSTNet [36] | 2019 | 0.2714 | 0.3108 | 0.3795 | 0.3792 |
| | RCN-A [52] | 2020 | 0.2751 | 0.3123 | 0.3928 | 0.3893 |
| | MERSiam [58] | 2021 | 0.3184 | 0.3532 | 0.3184 | 0.3532 |
| | FGRL [24] | 2021 | 0.0736 | 0.1429 | 0.3333 | 0.2636 |
| | FR [60] | 2022 | 0.3559 | 0.3814 | 0.3493 | 0.3413 |
| | MMNet [25] | 2022 | 0.2649 | 0.2776 | 0.3706 | 0.3646 |
| | BDCNN [4] | 2022 | 0.2975 | 0.3314 | 0.5050 | 0.5164 |
| | Micro-BERT [42] | 2023 | – | – | 0.5604 | 0.6125 |
| | HTNet [50] | 2024 | 0.3243 | 0.3368 | 0.5767 | 0.5415 |
| | HSTA [16] | 2024 | – | – | 0.593 | 0.618 |
| | **Ours** | - | **0.3853** | **0.3978** | **0.6221** | **0.6226** |

Table 3. Ablation study. L, G, S and D represent Local (upper and lower face), Global, Single stream and Dual Stream, respectively.

| Setting | LRM | | AFR | | InceptionNet | | UF1 ↑ | UAR ↑ |
|---|---|---|---|---|---|---|---|---|
| | LFE | SSE | L | G | S | D | | |
| I | | | | | ✓ | | 0.3249 | 0.3404 |
| II | ✓ | | | | | ✓ | 0.3362 | 0.3517 |
| III | ✓ | ✓ | | | | ✓ | 0.3480 | 0.3624 |
| IV | ✓ | ✓ | ✓ | | | ✓ | 0.3629 | 0.3723 |
| V | ✓ | ✓ | | ✓ | | ✓ | 0.3604 | 0.3658 |
| VI | ✓ | ✓ | ✓ | ✓ | | ✓ | **0.3853** | **0.3978** |
| **Setting** | **Without AU Group** | | | | **With AU Group** | | **UF1 ↑** | **UAR ↑** |
| VII | | ✓ | | | | | 0.3683 | 0.3793 |
| VIII | | | | | | ✓ | **0.3853** | **0.3978** |
| **Setting** | **Psychological prior** | | | | **Data-driven Prior** | | **UF1 ↑** | **UAR ↑** |
| IX | | | | | | | 0.3615 | 0.3701 |
| X | | ✓ | | | | | 0.3723 | 0.3832 |
| XI | | | | | | ✓ | 0.3705 | 0.3767 |
| XII | | ✓ | | | | ✓ | **0.3853** | **0.3978** |

## 5.2. Comparison to State-of-the-art Methods

Tab. 2 list the SOTA comparisons on DFME and CAS(ME)$^3$. The confusion matrices are shown in the Suppl.

Regarding DFME, classifying ME samples into seven categories poses significant challenges, resulting in relatively low UF1 and UAR scores across all methods. However, our proposed method outperforms all others in every metric. Specifically, Our method outperforms I3D by 9.30% in UF1 and 9.20% in UAR, and surpasses LBP-TOP by 15.17% in UF1 and 13.25% in UAR, highlighting its effectiveness in extracting critical and discriminative features from ME clips. Additionally, compared to recent deep learning methods like FR, our approach yields 2.94% and 1.64% higher UF1 and UAR, confirming the benefits of incorporating facial topology and structured muscle motion for enhanced performance.

Regarding CAS(ME)$^3$, the primary challenge lies in its sample complexity. Compared to baseline methods (STSTNet, RCN-A, FR), our proposed method achieves over approximately 20% higher UF1 and UAR scores, marking a significant advancement. Our method surpasses three recent SOTA approaches in MER, including HTNet, which employs a hierarchical Transformer for local feature learning via self-attention. Specifically, our approach outperforms HTNet by 4.54% in UF1 and 8.11% in UAR, leveraging psychological insights and statistical patterns to capture deeper facial muscle dynamics. Additionally, it exceeds Micro-BERT and HSTA by 6.17% and 2.91% in UF1, and 1.01% and 0.46% in UAR, respectively. By mining facial movement patterns to better understand topology, our model achieves improved metrics, indicating enhanced accuracy in classifying all emotion categories .

## 5.3. Ablation Study

Tab. 3 listed the ablation study results on DFME.

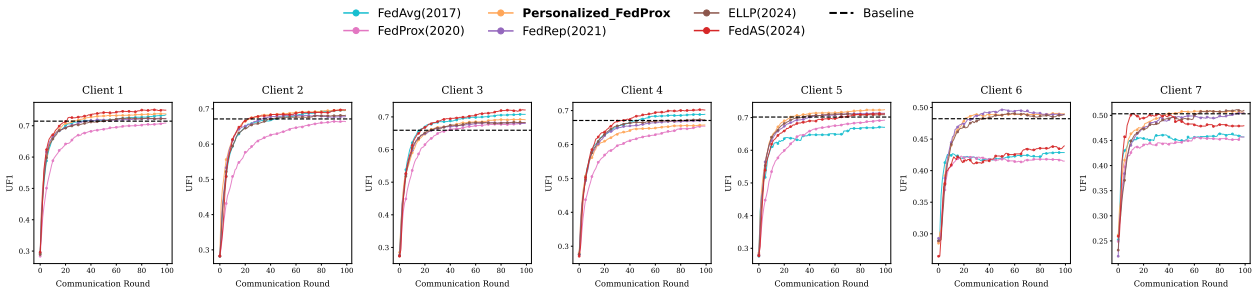
**Impact of AU Feature:** Setting I, which uses only full face OF as input, serves as our baseline for evaluating the role of AU features in MER. In Setting II, integrating AU features into the Dual Stream InceptionNet leads to a 1.13% improvement in both UF1 and UAR. This improvement demonstrates that incorporating AU features significantly enhances the model's ability to recognize MEs, and underscores the value of fusing multiple feature types to strengthen model performance.

**Impact of ROI Relationship Modeling:** The results in Setting III show improvements of 1.18% in UF1 and 1.07% in UAR compared to Setting II, indicating that the model effectively captures complex dependencies between ROIs and extracts more discriminative feature representations.

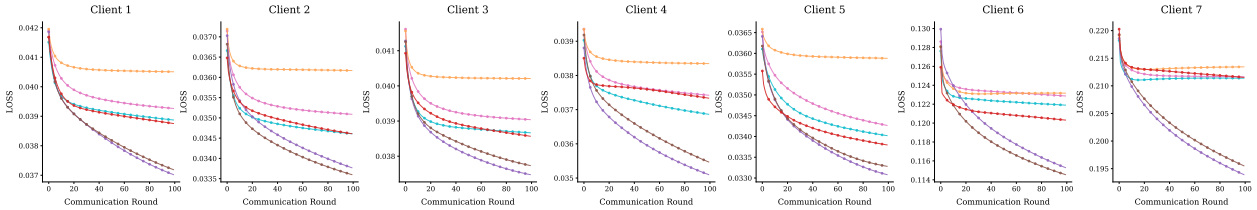
**Impact of AU grouping:** We compared AU feature extraction from the global face with our AU grouping strategy (Setting VII vs. VIII), the performance improvement (UF1: 1.70%, UAR: 1.85%), confirming that AU grouping enables the network to learn more representative AU features.

**Impact of Modeling the Topological Structure of ME Muscle Movements:** Compared to Setting III, Setting IV, including local muscle movement topology modeling, improves UF1 and UAR by 1.49% and 0.99%, respectively. Setting V, which adds global muscle movement topology modeling, shows gains of 1.24% in UF1 and 0.34% in UAR over Setting III. Combining both local and global topology modeling yields the best performance on DFME. These results demonstrate that incorporating topological modeling of muscle movements enhances MER.

**Impact of Prior Knowledge:** Tab. 3 presents the ablation experiments to evaluate the impact of using prior knowledge for constructing facial muscle topology. In particular, compared to Setting IX, Setting X, which uses only psychological knowledge, improves UF1 and UAR by 1.08% and




(a) UF1 comparison across different clients.




(b) Convergence comparison across different clients.

Figure 5. Compare the performance of different FL frameworks on different clients: UF1 and convergence. See UAR comparison in Suppl.

1.31%, respectively. Setting XI, incorporating only data-driven knowledge, shows gains of 0.9% in UF1 and 0.66% in UAR. When both psychological and data-driven knowledge are combined (Setting XII), the improvements reach 2.38% in UF1 and 2.77% in UAR, proving the effectiveness of the prior knowledge. Constructing a prior adjacency matrix allows the model to capture complex facial muscle topology, while data-driven knowledge directs attention to uncover intrinsic movements patterns of facial muscles.

### 5.4. Federated Experiment

Since FL aims to enhance MER performance on individual clients in scenarios with limited sample sizes and data privacy issue, and given that DFME and CAS(ME)$^3$ have large sample size, we partition the datasets as mentioned in Section 5.1 and treat the subdivisions as separate clients.

While our focus is on application rather than fundamental FL innovation, our experiments confirm the effectiveness of our P-FedProx method, which is tailored for ME data distributions. Benchmarked against several open-source SOTA methods, our approach excels at mitigating small-sample challenges while preserving data privacy. Specifically (Fig. 5), the results show P-FedProx outperforms both the single-client baseline and traditional FL methods (FedAvg, FedProx), with particularly significant gains on clients with more challenging data distributions (Clients 6-7). Compared to recent personalized FL methods, our approach surpasses FedRep and ELLP and achieves performance competitive with the state-of-the-art FedAS, yet with a significantly simpler implementation. These consistent gains demonstrate that our weighted model initialization effectively addresses the data heterogeneity that challenges traditional FL, thereby mitigating ME sample scarcity while ensuring data privacy.

Meantime, we note that all FL methods achieved good performance on the local clients from DFME, which may be due to the fact that these clients have a similar amount of data and slight differences in data distribution, presenting fewer challenges for federated learning.

Besides, we analyze the training loss progression of different federated algorithms over communication rounds. We found that P-FedProx demonstrates the smoothest convergence curves with minimal oscillations, particularly evident in Clients 1-5. On the other hand, P-FedProx converges significantly faster than other FL methods. This rapid early convergence suggests that personalized global model assignment provides clients with better initialization states.

## 6. Conclusion

MER is challenged by limited sample sizes, subtle features, and privacy concerns in real-world applications. To address these issues, we conduct a psychological study to provide structured insights into facial muscle dynamics. We then develop a FED-PsyAU framework that integrates these findings with data-driven patterns for hierarchical learning of dynamic features, enhancing MER performance. Additionally, the FL framework improves MER across clients while preserving data privacy through iterative model updates. The comprehensive experiments validate our approach's effectiveness. In the future, we will integrate additional AUs to build a more comprehensive ME topology information and further advance MER in real-world applications through FL and semi-supervised approaches.

## 7. Acknowledgment

This research was partially funded by 1) the National Natural Science Foundation of China (62476269, 62276252); 2) the Youth Innovation Promotion Association CAS.

## References

[1] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-Based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5826–5846, 2022. 2

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[3] Boyu Chen, Zhihao Zhang, Nian Liu, Yang Tan, Xinyu Liu, and Tong Chen. Spatiotemporal Convolutional Neural Network with Convolutional Block Attention Module for Micro-Expression Recognition. *Information*, 11(8):380, 2020. 1

[4] Bin Chen, Kun-Hong Liu, Yong Xu, Qing-Qiang Wu, and Jun-Feng Yao. Block Division Convolutional Network With Implicit Deep Features Augmentation for Micro-Expression Recognition. *IEEE Transactions on Multimedia*, 25:1345–1358, 2022. 1, 2, 7

[5] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021. 6

[6] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing*, 9(1):116–129, 2018. 2

[7] Zizhao Dong, Gang Wang, Shaoyuan Lu, Jingting Li, Wenjing Yan, and Su-Jing Wang. Spontaneous Facial Expressions and Micro-Expressions Coding: From brain to Face. *Frontiers in Psychology*, 12:784834, 2022. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. 3

[9] Paul Ekman and Wallace V Friesen. Nonverbal Leakage and Clues to Deception. *Psychiatry*, 32(1):88–106, 1969. 1

[10] Paul Ekman and Wallace V Friesen. Facial Action Coding System. *Environmental Psychology & Nonverbal Behavior*, 1978. 1, 2, 3, 4

[11] Ali et al. Privacy preserving personalization for video facial expression recognition using federated learning. *IMCI'22*, . 2

[12] Fan et al. Feeling without sharing: A federated video emotion recognition framework via privacy-agnostic hybrid aggregation. *ACM MM'23*, . 2

[13] Xinqi Fan, Xueli Chen, Mingjie Jiang, Ali Raza Shahid, and Hong Yan. SelfME: Self-Supervised Motion Learning for Micro-Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13843, 2023. 2

[14] Viswanatha Reddy Gajjala, Sai Prasanna Teja Reddy, Snehasis Mukherjee, and Shiv Ram Dubey. MERANet: facial micro-expression recognition using 3D residual attention network. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, New York, NY, USA, 2021. Association for Computing Machinery. 1

[15] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019. 7

[16] Haihong Hao, Shuo Wang, Huixia Ben, Yanbin Hao, Yansong Wang, and Weiwei Wang. Hierarchical Space-Time Attention for Micro-Expression recognition. *arXiv preprint arXiv:2405.03202*, 2024. 7

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 7

[18] Xiaohua Huang, Su-Jing Wang, Guoying Zhao, and Matti Piteikainen. Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1–9. IEEE, 2015. 2

[19] T. Kanade, J.F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53, 2000. 6

[20] Davis E King. Dlib-ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 3

[21] Thomas N Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*, 2016. 1, 2

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 7

[23] Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144, 2017. 3

[24] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1571–1580. IEEE, 2021. 1, 2, 7

[25] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao. Mmnet: Muscle motion-guided network for micro-expression recognition. *arXiv preprint arXiv:2201.05297*, 2022. 7

[26] Jingting Li, Moi Hoon Yap, Wen-Huang Cheng, John See, Xiaopeng Hong, Xiaobai Li, and Su-Jing 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*, pages 5700–5701. ACM, 2021. 2

[27] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan 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*, 45(3):2782–2800, 2022. 2

[28] Tian Li, Anit Kumar Sahu, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the Convergence of Federated Optimization in Heterogeneous Networks. *arXiv preprint arXiv:1812.06127*, 2018. 2, 6

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020. 6

[30] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti 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)*, pages 1–6, 2013. 2

[31] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. *IEEE Transactions on Affective Computing*, 9(4): 563–577, 2018. 1

[32] Xiaobai Li, Shiyang Cheng, Yante Li, Muzammil Behzad, Jie Shen, Stefanos Zafeiriou, Maja Pantic, and Guoying Zhao. 4DME: A Spontaneous 4D Micro-Expression Dataset With Multimodalities. *IEEE Transactions on Affective Computing*, 14(4):3031–3047, 2023. 2

[33] Yante Li, Xiaohua Huang, and Guoying Zhao. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231, 2021. 1

[34] Sze-Teng Liong, John See, Raphael C.-W. Phan, Anh Cat Le Ngo, Yee-Hui Oh, and KokSheik Wong. Subtle Expression Recognition Using Optical Strain Weighted Features. In *Computer Vision - ACCV 2014 Workshops*, pages 644–657. Springer International Publishing, 2015. Series Title: Lecture Notes in Computer Science. 2

[35] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C.-W. Phan. Less is More: Micro-expression Recognition from Video using Apex Frame. *Signal Processing: Image Communication*, 62:82–92, 2018. 2

[36] Sze-Teng Liong, Y. S. Gan, John See, Huai-Qian Khor, and Yen-Chang 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)*, pages 1–5, 2019. 7

[37] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. *IEEE Transactions on Affective Computing*, 7 (4):299–310, 2015. 7

[38] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7(4):299–310, 2016. 2

[39] Ling Lo, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks. In *2020 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 79–84. IEEE, 2020. 1

[40] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 6

[41] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. ISSN: 2640-3498. 2, 6

[42] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-BERT: BERT-Based Facial Micro-Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1482–1492, 2023. 7

[43] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 2

[44] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. CAS(ME)$^2$: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing*, 9(4):424–436, 2018. 2

[45] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. MEGC 2019 – The Second Facial Micro-Expressions Grand Challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5, 2019. 2, 6

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*, pages 1–9, 2015. 5

[47] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. *arXiv preprint arXiv:1710.10903*, 2017. 1

[48] Lei Wang, Pinyi Huang, Wangyang Cai, and Xiyao Liu. Micro-expression recognition by fusing action unit detection and spatio-temporal features. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5595–5599. IEEE, 2024. 2

[49] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces.

*IEEE Transactions on Image Processing*, 24(12):6034–6047, 2015. 1, 2, 7

[50] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024. 1, 2, 7

[51] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan 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)*, pages 1–7. IEEE, 2013. 2

[52] Zhaoqiang Xia, Wei Peng, Huai-Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29: 8590–8605, 2020. 7

[53] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. AU-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2871–2880. ACM, 2020. 2

[54] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 9(1):e86041, 2014-01-27. 2

[55] Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11986–11995, 2024. 6

[56] Moi Hoon Yap, John See, Xiaopeng Hong, and Su-Jing Wang. Facial Micro-Expressions Grand Challenge 2018 Summary. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 675–678, 2018. 2

[57] Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22086–22095, 2023. 1

[58] Sirui Zhao, Hanqing Tao, Yangsong Zhang, Tong Xu, Kun Zhang, Zhongkai Hao, and Enhong Chen. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. *Neurocomputing*, 448:276–289, 2021. 7

[59] Sirui Zhao, Huaying Tang, Xinglong Mao, Shifeng Liu, Yiming Zhang, Hao Wang, Tong Xu, and Enhong Chen. DFME: A New Benchmark for Dynamic Facial Micro-Expression Recognition. *IEEE Transactions on Affective Computing*, 15(3):1371–1386, 2024. 2, 6

[60] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature Refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. 7

[61] Xu Zhou, Jie Li, Gongjin Lan, Rongrong Ni, Angelo Cangelosi, Jiaxin Wang, and Xiaofeng Liu. Efficient lower layers parameter decoupling personalized federated learning method of facial expression recognition for home care robots. *Information Fusion*, 106:102261, 2024. 6