A Multi-Modal Multi-Expert Framework for Pain Assessment in Postoperative Children

Zequan Liang, Hao Luo, Xi Chen, Zhipeng Zhong, Cheng Fan, Xingrong Song, Bilian Li, and Jianming Lv, *Member, IEEE*

***Abstract*—Automatic pain assessment in postoperative children is crucial for monitoring their health and preventing potential complications. However, the automatic pain assessment still faces the following challenges. Firstly, the individual variation of painful expressions in children enhances the difficulty of mapping diverse features of expressions to pain scores accurately. Secondly, the imbalanced label distribution caused by abundant non-painful samples usually makes the model more likely to predict an unexpectedly lower pain score. To address the above challenges, we propose a novel multi-modal multi-expert framework, namely *MMF*, for postoperative pain assessment in children. Specifically, the samples are clustered in each modality to train multiple expert models, each focusing on a smaller feature subspace for easier regression of pain scores. Meanwhile, some expert models are allocated to rare painful samples to relieve the side effects caused by the imbalanced distribution of labels. Moreover, a confidence-based integration of multi-modal features from multiple experts is made to achieve a more accurate final prediction. Experimental results show that *MMF* exhibits superior accuracy of pain assessment on the multi-modal pain database collected from postoperative children by us. In particular, *MMF* can achieve the mean absolute error (MAE) of 1.03 and the Pearson correlation coefficient (PCC) of 0.88.**

***Index Terms*—Confidence, multi-expert, multi-modal, pain assessment, postoperative children.**

Received 26 October 2023; revised 11 April 2025; accepted 23 April 2025. Date of publication 13 May 2025; date of current version 3 December 2025. This work was supported in part by the Science and Technology Program of Guangzhou, China under Grant 2023B03J1388, in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2024A1515012287 and Grant 2023A1515220013, in part by the National Key Research and Development Program of China under Grant 2023YFA1011601, and in part by the Major Clinical Research Program of Scientific Research Capability Enhancement Plan of Guangzhou Medical University under Grant GMUCR2024-02019. Recommended for acceptance by F. Schwenker. *(Zequan Liang and Hao Luo are co-first authors.) (Corresponding authors: Bilian Li; Jianming Lv.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Guangzhou Women and Children's Medical Center, under the Institutional Review Board (IRB) under Application No. 044A01, and performed in line with the Chinese Clinical Trial Registry, bearing the ID ChiCTR2200058967, with B. Li as the Principal Investigator.

Zequan Liang and Jianming Lv are with the Department of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: a542282173@gmail.com; jmlv@scut.edu.cn).

Hao Luo, Zhipeng Zhong, Cheng Fan, Xingrong Song, and Bilian Li are with the Department of Anaesthesiology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510006, China (e-mail: lh18607479594@stu.gzhmu.edu.cn; 2022210325@stu.gzhmu.edu.cn; fancheng@stu.gzhmu.edu.cn; sxjess@126.com; bilian.li@gzhmu.edu.cn).

Xi Chen is with the Department of Electrical and Electronic Engineering, University College London, WC1E 6BT London, U.K. (e-mail: x.chen.20@ucl.ac.uk.).

Digital Object Identifier 10.1109/TAFFC.2025.3567307

I. Introduction

POSTOPERATIVE pain commonly occurs for underlying illnesses or improper surgical procedures. In the field of clinical medicine, pain assessment is critical for monitoring patient health and preventing potential complications [1]. This is particularly important for children who may not be able to verbally inform their pain levels [2]. The pain scales are commonly used tools for pain assessment in children. In particular, the Visual Analog Scale (VAS) [3] and the Wong-Baker FACES Pain Rating Scale [4] require children to self-report their pain levels, which rely on children's memory and are not in real-time. The Face, Legs, Activity, Cry, Consolability (FLACC) scale [5], relies on the observations by healthcare professionals to assess pain levels in real-time. However, this approach is subject to observer bias and requires significant human resources. To improve the objectivity and efficiency of pain assessment, the use of artificial intelligence for automatic pain assessment has become increasingly important.

The field of automatic pain assessment is developing with the construction of several pain databases, such as the UNBC-McMaster Shoulder Pain Database [6] and the BioVid Heat Pain Database [7]. The former only contains facial expressions and the latter includes multi-modal information, including facial expressions and biological metrics, which record the reactions during stimulation. However, these databases solely focus on adults and lack audio information. As for children, vocal expression is also essential information when they cry unconsciously due to their lower pain tolerance compared to adults. The level of pain is relevant to the loudness and duration of their cries [8]. The Infant Cry Signals Database (IIIT-S ICSD) [9] records audio cry samples from 33 infants for pain analysis. Recently, some researchers integrated more signals such as biological metrics [10] and body behavior [11] to support multi-modal pain assessment. The fusion of multiple modalities usually leads to higher accuracy in pain assessment [12]. More and more pain databases include different modalities, as detailed in the survey by Werner et al. [13].

Despite the rapid development of automatic pain assessment, it still faces some critical challenges: 1) **Individual variation of painful expressions** makes it hard to learn the correct mapping between the diverse visual features and pain scores [14], especially for the models trained on a small labeled database. Fig. 1 shows an example of the facial feature space, where the features with the same pain label (indicated with the same color) are scattered in different regions, representing different

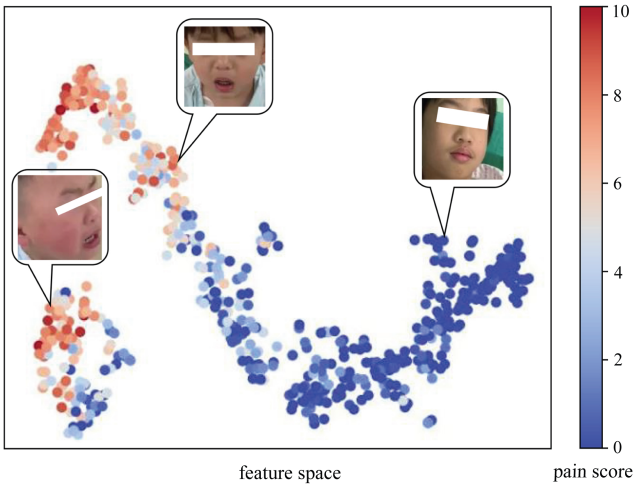


Fig. 1. T-SNE plots of the feature vectors for facial expressions in the *MPP* Database. Each point represents the facial feature vector of a sample, and the color of a point indicates the pain score of the sample.

facial expressions. 2) **Imbalanced distribution of pain scores** is inevitable in the construction of a clinical database. Real clinical scenarios usually have a limited number of patients experiencing severe pain [15]. The model trained on the whole database can be easily affected by abundant non-painful samples, leading to the high possibility of predicting lower pain scores. 3) **Shortage of multi-modal pain databases for postoperative children.** Many pain databases either focus solely on adults [6], [7] or contain only uni-modal information for children [9], [16]. Although the YouTube database [17] includes videos of 142 infants in pain during injections, it has shaky scenes and lacks consistent focus on infants' faces.

To address the above challenges, we propose a novel multi-modal multi-expert framework, namely *MMF*, for automatic postoperative pain assessment in children. *MMF* deploys multiple expert models to learn diverse multi-modal features and measure the confidence of an expert based on the distance between the test sample to its prototype learned in the training set. Each expert focuses on a relatively smaller cluster in the feature subspace and a confidence-based integration of all experts is made to generate the final decision. Meanwhile, some experts are allocated to rare painful samples, so as to avoid over-fitting to the abundant non-painful ones. Furthermore, a multi-modal pain database for children was collected and used for comprehensive performance analysis, which showed the effectiveness and efficiency of *MMF*.

In summary, the contributions of this paper are as follows:

- We introduce a novel multi-modal multi-expert Framework, namely *MMF*, for accurate pain assessment in postoperative children. Each expert is trained to make predictions accurately in a smaller feature subspace, so as to relieve the negative influence of the individual variation of painful expressions.
- We propose a prototype based method to measure the confidence of each expert and integrate the multi-level information, including the middle features and prediction results, from multiple experts in different modalities for the final prediction of pain scores.
- We constructed a multi-modal pain database for postoperative children, which contains the facial and vocal expressions from 701 individual video samples, to validate the effectiveness of our framework.

II. Related Work

A. Multi-Modal Pain Assessment

Multi-modal fusion is an essential aspect of artificial intelligence, where information processed by computers is combined to generate more accurate predictions through computer models [18]. Some studies about automatic pain assessment have employed multi-modal fusion techniques to improve the accuracy of pain-level classification. For example, Thiam et al. [19] used a random forest to perform both feature-level and decision-level fusion on the Sense Emotion Database [20] for multiple modalities. They extracted head descriptors and Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [21] calculators from the video. Classical statistical methods were also used to extract feature values of biological metrics on a certain time window. The fusion experiments showed that multi-modal fusion improved performance compared to each modality individually. Zamzmi et al. [22] trained models separately for facial expressions, body movements, and physiological signals, extracting features similar to the previous method. They then used a simple majority voting method to combine multi-modal predictions into a pain indicator, following the decision-level fusion method. Kächele et al. [23] extracted multi-modal features based on the BioVid Database and proposed an adaptive confidence algorithm to weigh the contributions of each modality based on their reliability.

However, due to the limited amount of data available, the above classification models were trained on a small number of labeled samples (usually not more than 200 individuals) [6], [7]. The individual variation of painful expression usually makes it difficult for these models to generalize. Furthermore, these models that are designed for healthy adults who endure temporary stimulation are not sufficiently qualified to be used for multi-modal pain assessment in postoperative children.

B. Multi-Expert Framework

The multi-expert framework was first introduced as the Mixture-of-Experts (MoE) concept by Michael et al. [24] and has been widely applied in the field of artificial intelligence for over 30 years. This framework primarily employs multiple distinct models to learn from different subsets of training samples, with a gating network determining which data should be trained by which model.

Recently, a lot of efforts have been made to expand the concept of the multi-expert framework. For instance, Shazeer et al. [25] proposed the Sparsely-Gated MoE, which activates only a few experts from a large pool to improve computational efficiency. They utilized the Top-K sampling strategy and softmax calculation to limit the number of experts that data passes through while

TABLE I
ANNOTATIONS FOR SYMBOLS USED IN OUR METHODOLOGY

Symbol	Annotation
N	the number of samples, $i \in [1, N]$
T	the length of a sequence, $t \in [1, T]$
m	the type of modality from face or voice, $m \in \{F, V\}$
y_i	the ground-truth pain score of the i^{th} sample
X_i^m	the frame sequence extracted from the i^{th} sample
E_i^m	the feature sequence extracted from X_i^m
f_i^m	the temporal feature extracted from E_i^m by $BiLSTM$
K_m	the number of clusters in the modality m
c_k^m	the center of the k^{th} cluster in the modality m
C_k^m	the set of the feature vectors in the k^{th} cluster
$f_i'^m$	the feature vector of the i^{th} sample in the prototype space
$c_k'^m$	the center of the k^{th} cluster in the prototype space
D	the dimension of feature space
P	the dimension of prototype space
ρ_k^m	the prototype of the k^{th} cluster
$w_{i,k}^m$	the confidence for the k^{th} expert
$y_{i,k}'^m$	the pain score of the i^{th} sample predicted by the k^{th} expert
$v_{i,k}^m$	the feature extracted by the k^{th} expert in the modality m
s_i^m	the fusion pain score of the i^{th} sample in the modality m
r_i^m	the fusion feature of the i^{th} sample in the modality m
R_i	the final fusion feature of the i^{th} sample
S_i	the final prediction of the pain score for the i^{th} sample

balancing the gating weights to ensure that each expert is valuable. Dmitry Lepikhin et al. [26] proposed GShard that extended the MoE concept to Transformers and employed a Top-2 gating network following the sparsing method. Basil et al. [27] later proposed the Language-Image Mixture of Experts (LIMoE), which uses a sparse mixture of experts to simultaneously process both images and text for multi-modal retrieval tasks. Most of the above multi-expert methods are designed for single-modal data classification and operate on the entire database. In this paper, we propose a multi-modal multi-expert framework for pain assessment.

III. TASK DEFINITION

In this paper, we propose a method for assessing pain levels by regressing a pain score from video data that captures both facial and vocal information of postoperative children. The i^{th} video in the database of N videos is preprocessed to obtain an image frame sequence $X_i^F = \{x_{i,0}^F, x_{i,1}^F, \ldots, x_{i,T}^F\}$ and a vocal frame sequence $X_i^V = \{x_{i,0}^V, x_{i,1}^V, \ldots, x_{i,T}^V\}$, where T represents the sequence length. The ground-truth pain score of the i^{th} video is denoted by y_i, which corresponds to the average FLACC score assessed by medical professionals. The pain assessment model generates a prediction score S_i that approximates y_i as closely as possible using both modalities. A list of annotations for symbols used in our methodology is provided in Table I.

IV. METHODS

A. Framework Overview

As shown in Fig. 2, *MMF* mainly consists of three steps. Firstly, each input video is preprocessed into a facial frame sequence X_i^F and a vocal frame sequence X_i^V, and the features of these sequences are extracted. Secondly, the sequences are clustered based on the similarity of their features and multiple experts are trained on these clusters in each modality. Thirdly, prototype-based confidence is adopted to weigh the experts in the fusion of the final decision, where the prototype indicates the typical features of the samples learned by the expert. Each step will be detailed in the following sections.

B. Feature Extraction

Each input file contains a T seconds video of postoperative children, and it is divided into T image segments and audio segments according to the order of frames.

From each image segment, a single frame is randomly selected. As a result, a sequence of T frames is achieved for each video. Each frame contains an RGB image of a child lying in bed, expressing comfort or discomfort through facial expressions. A trained stacked hourglass network [28] is utilized to detect and extract the child's face from the frame. In the instances where other medical persons are present in the video, only the largest face is captured and tracked using the Intersection over Union (IOU) compared with the previous frame. This results in a sequence of facial frames $X_i^F = \{x_{i,t}^F\}$ for the target patient as shown in Fig. 2, where $i \in [1, N]$ indicates the i^{th} video out of N, and $t \in [1, T]$ represents the video timestamp out of T. A ResNet18 model [29] pre-trained on the CK+ face database [30] is employed to extract the features from each facial frame, so as to achieve a sequence $E_i^F = \{e_{i,t}^F\}$ of facial features for each video.

Meanwhile, for each audio segment, the Mel Frequency Cepstrum Coefficients (MFCC) [31] are extracted and transformed into a frequency-time spectrogram using the librosa [32] python package. This results in a sequence of spectrograms $X_i^V = \{x_{i,t}^V\}$ as shown in Fig. 2. A ResNet-18 model is then utilized to extract the features from each 2D spectrogram, resulting in a sequence $E_i^V = \{e_{i,t}^V\}$ of vocal features for each video.

To extract the temporal features of human expressions, we employ the Bidirectional Long Short-Term Memory (BiLSTM) [33] on the above sequence $E_i^m = \{e_{i,t}^m\}$ where $m \in \{F, V\}$ as follows:

$$f_i^m = BiLSTM\left(E_i^m\right) \tag{1}$$

The temporal features f_i^m ($m \in \{F, V\}$) are then mapped to the pain score by the linear projection:

$$y_i'^m = \left(W^m\right)^T f_i^m + b^m \tag{2}$$

where $W^m, f_i^m \in \mathbb{R}^D$ and $y_i'^m, b^m \in \mathbb{R}$. The feature extraction model is optimized by minimizing the MSE (Mean Squared Error) loss as follows:

$$\mathcal{L}_{mse}^m = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i'^m\right)^2 \tag{3}$$

where y_i represents the ground-truth pain score of the i^{th} video, and $y_i'^m$ represents the predicted pain score of the feature extraction model.

C. Training of Multiple Experts

To address the challenge of individual variation of painful expressions, multiple expert models are trained on the set of

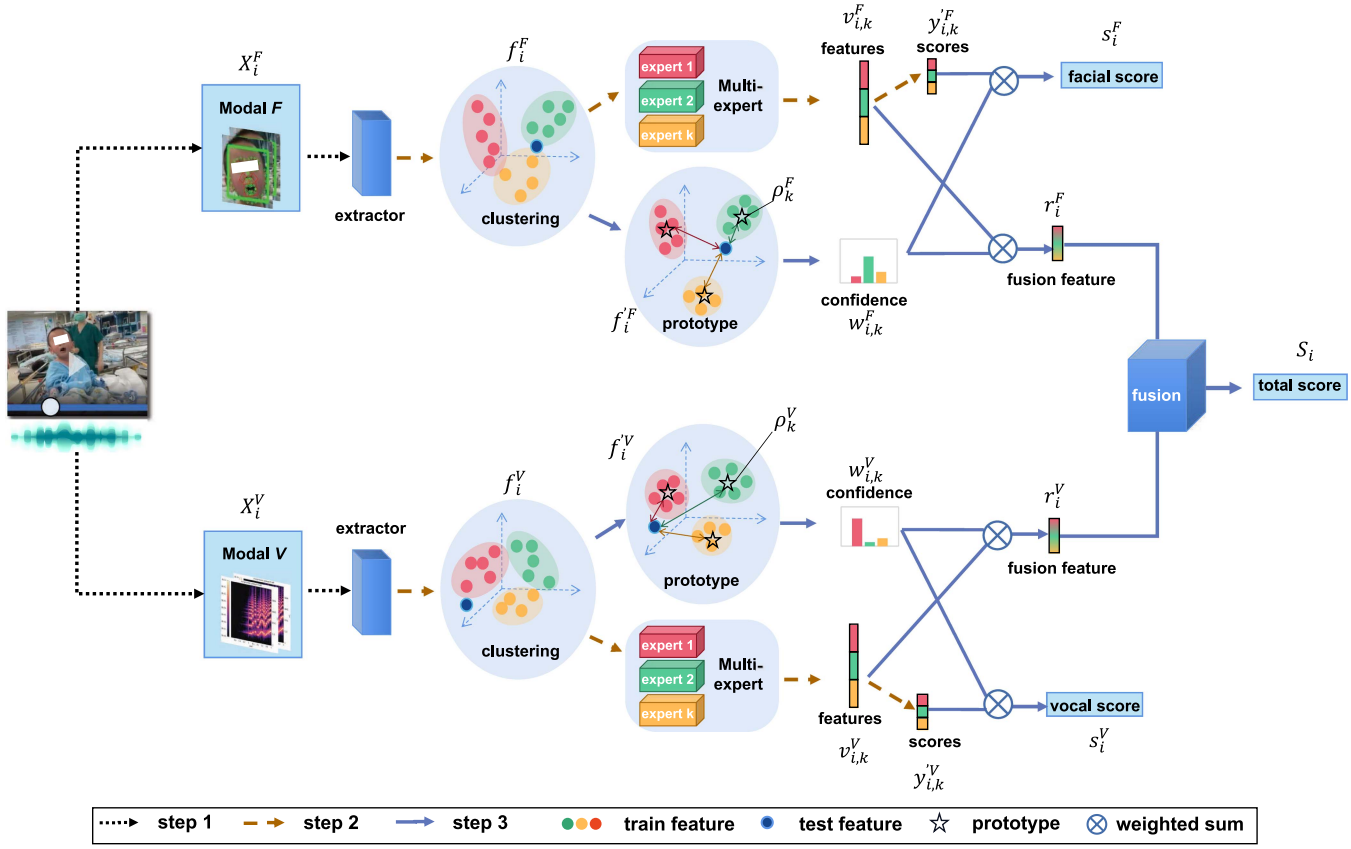


Fig. 2. The *MMF* mainly consists of three steps: 1) Preprocess information and extract features from different data modalities; 2) Cluster samples based on their features, and train the multi-experts on these sample clusters; 3) Generate a prototype feature of each sample cluster and calculate the expert's confidence, which is used for the final multi-modal multi-expert fusion.

feature vectors. Each expert is trained and verified in a smaller cluster of samples sharing similar features, and tends to learn more consistent knowledge to make easier regression of pain scores. The details of the clustering algorithm and the training procedure of multiple experts will be illustrated in the following subsections.

1) Clustering in the Feature Space: In the case of regression tasks, the extracted features $\{f_i^m\}$ typically exhibit a continuous distribution with a band-like shape as shown in Fig. 1, making it difficult to partition them into distinct clusters. We propose a drag-push clustering method to address this issue.

A hyperparameter K_m is set to represent the desired number of clusters for modality m. Firstly, K_m vectors $\{c_k^m | 1 \leq k \leq K_m\}$ are randomly selected to serve as the centers of the clusters, where $c_k^m \in \mathbb{R}^D$. For each feature vector f_i^m, the Euclidean distance to each center c_k^m is calculated to determine its cluster assignment.

Specifically, f_i^m is assigned to the k'^{th} cluster, where

$$k' = \underset{k \in [1, K_m]}{\arg\min} \left(\| f_i^m - c_k^m \|_2 \right) \tag{4}$$

It is also denoted as $f_i^m \in C_{k'}^m$, where $C_{k'}^m$ indicates the k'^{th} cluster in the feature space.

Secondly, we optimize the centers of the clusters by dragging each center c_k^m towards the feature vectors in the same cluster C_k^m and pushing it away from the ones in the other clusters. This is achieved by minimizing the sum of all Euclidean distances between c_k^m and $f_i^m \in C_k^m$, while maximizing the sum of all Euclidean distances between c_k^m and $f_i^m \notin C_k^m$. The centers $\{c_k^m\}$ are optimized by minimizing the following loss function:

$$\begin{aligned} \mathcal{L}_{clt}^m = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K_m} [& -Y_{i,k}^m \log \left(G \left(f_i^m, c_k^m \right) \right) \\ & - \left(1 - Y_{i,k}^m \right) \log \left(1 - G \left(f_i^m, c_k^m \right) \right)] \end{aligned} \tag{5}$$

where

$$G(x, y) = \exp \left(- \| x - y \|_2 \right) \tag{6}$$

$$Y_{i,k}^m = \begin{cases} 1 & \text{if } f_i^m \in C_k^m \\ 0 & \text{if } f_i^m \notin C_k^m \end{cases} \tag{7}$$

Here $G(x, y)$ indicates the Gaussian similarity, which measures the distance between feature vectors and projects the range of values into the interval (0,1]. By minimizing $\mathcal{L}_{clt}^m$, the trainable centers $\{c_k^m\}$ shift within the feature space $\{f_i^m\}$, resulting in the generation of clusters based on (4). The corresponding samples are also grouped into several clusters for the following training of the multi-expert model.

2) Multi-Expert Training: For each cluster C_k^m, one expert model is trained based on the samples of the cluster to predict

pain scores. Following the similar pipeline of (1∼3), the $k^{th}(1 \leq k \leq K_m)$ expert model applies the BiLSTM model to extract the temporal features and maps them into the pain score by the linear projection as follows:

$$v_{i,k}^m = BiLSTM(E_i^m) \tag{8}$$

$$y_{i,k}^{\prime m} = (W_k^m)^T v_{i,k}^m + b_k^m \tag{9}$$

where $v_{i,k}^m, W_k^m \in \mathbb{R}^D$, $y_{i,k}^{\prime m}, b_k^m \in \mathbb{R}$. The model is optimized by minimizing the following loss function:

$$\mathcal{L}_{mse}^{\prime m} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_{i,k}^{\prime m}\right)^2 \tag{10}$$

D. Multi-Modal Multi-Expert Fusion

In the testing phase of a multi-expert framework, it is necessary to determine which experts are best suited to evaluate an unseen test sample. We define the confidence of each expert by measuring the similarity between the test sample and the typical sample of its training set, namely the prototype as defined in [34]. While the test sample is more similar to the training samples of the expert model, it is more possible to make the correct decision.

1) Gaussian Prototype Extraction: The simplest way to define the prototype of the $k^{th}(1 \leq k \leq K_m)$ expert model is to choose the center c_k^m of the corresponding cluster, which includes the training sample of the model. However, as observed in Fig. 1, the feature space typically exhibits a non-Gaussian manifold structure, which makes it difficult to measure the similarity between the prototype and other test samples. Inspired by our previous work [34], we propose a Gaussian prototype extraction procedure, which transforms the feature space into a prototype space to achieve a more compact Gaussian distribution and measure the similarity with the prototype based on the Gaussian similarity.

Firstly, we apply a linear projection to transform the original feature space into a prototype space. In particular, for any original feature vector f_i^m, the new vector $f_i^{\prime m}$ with new dimension P in the prototype space is achieved as follows:

$$f_i^{\prime m} = W_p^m f_i^m + b_p^m \tag{11}$$

where $W_p^m \in \mathbb{R}^{P\times D}$ and $f_i^{\prime m}, b_p^m \in \mathbb{R}^P$. In a similar way, each cluster center $c_k^m \in \mathbb{R}^D$ is also projected into $c_k^{\prime m} \in \mathbb{R}^P$. Both $\{f_i^{\prime m}\}$ and $\{c_k^{\prime m}\}$ are further optimized by minimizing the following loss function:

$$\mathcal{L}_{pro}^m = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K_m}[-Y_{i,k}^m \log\left(G\left(f_i^{\prime m}, c_k^{\prime m}\right)\right) - \left(1 - Y_{i,k}^m\right)\log\left(1 - G\left(f_i^{\prime m}, c_k^{\prime m}\right)\right)] \tag{12}$$

The optimization based on (12) may drag the feature vectors belonging to the same cluster towards the center, and push them away from other clusters, which leads to a more compact Gaussian distribution of the prototype space. The prototype of the k^{th} cluster is defined as $\rho_k^m = (\mu_k^m, \sigma_k^m)$, which contains the mean value μ_k^m and variance σ_k^m of the vectors in the clusters:

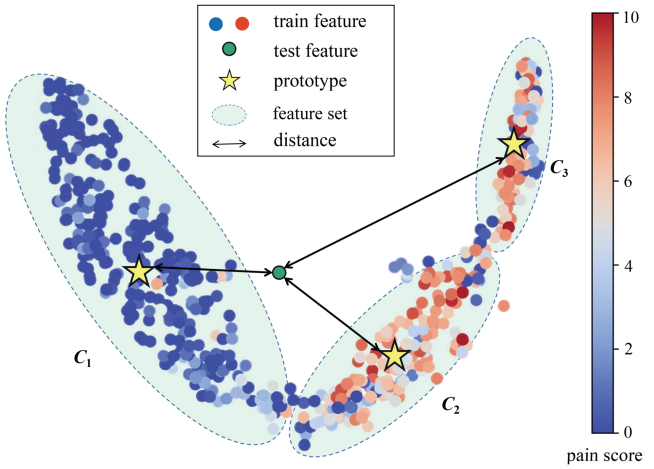


Fig. 3. T-SNE plots of feature vectors and prototypes in the prototype space. Each point represents the facial feature vector of a sample, and the color of a point indicates the pain score of the sample. For a test feature, the Gaussian distances from different prototypes are calculated.

$$\mu_k^m = \frac{1}{|C_k^m|}\sum_{i=1}^{N} Y_{i,k}^m f_i^{\prime m} \tag{13}$$

$$\sigma_k^m = \frac{1}{|C_k^m|}\sum_{i=1}^{N} Y_{i,k}^m (f_i^{\prime m} - \mu_k^m) \odot (f_i^{\prime m} - \mu_k^m) \tag{14}$$

where $|C_k^m|$ denotes the cardinality of the set C_k^m, and $\odot$ indicates the element-by-element Hadamard product.

While μ_k^m effectively represents a typical sample in a cluster, it neglects the distribution of the cluster. As an instance shown in Fig. 3, although distances from the test feature to the centers of the clusters C_1 and C_2 are equal, the distribution of C_1 is wider than that of C_2. As a result, the test feature should be closer to C_1. This challenge can be addressed by incorporating σ_k^m, which represents the variance of a cluster, into the measurement of the similarity with the prototypes.

2) Confidence-Based Fusion: After achieving the prototype $\rho_k^m = (\mu_k^m, \sigma_k^m)$ of the k^{th} expert model, the confidence of the model to judge a test sample $f_i^{\prime m}$ can be defined based on the Gaussian similarity:

$$w_{i,k}^m = \frac{\exp\left(-\left\|\frac{f_i^{\prime m} - \mu_k^m}{\sqrt{\sigma_k^m}}\right\|_2\right)}{\sum_{k'=1}^{K_m}\exp\left(-\left\|\frac{f_i^{\prime m} - \mu_{k'}^m}{\sqrt{\sigma_{k'}^m}}\right\|_2\right)} \tag{15}$$

which is computed based on the Gaussian similarity between $f_i^{\prime m}$ and the center μ_k^m of a prototype, normalized by σ_k^m. That is, the higher similarity between $f_i^{\prime m}$ and the prototype of the training set, the higher possibility that the k^{th} expert can make correct prediction.

As shown in Fig. 2, for each modality m, the pain score s_i^m is predicted by weighted summing the results from multiple

experts based on their confidence:

$$s_i^m = \sum_{k=1}^{K_m} w_{i,k}^m y_{i,k}^{\prime m} \tag{16}$$

Meanwhile, based on the confidence measurement from (15), the feature vectors $v_{i,k}^m$ (8) of all experts in modality m are combined to form a fusion feature r_i^m:

$$r_i^m = \sum_{k=1}^{K_m} w_{i,k}^m v_{i,k}^m \tag{17}$$

Subsequently, the fusion features r_i^m of all modalities are concatenated to fuse the information from all modalities:

$$R_i = concat\left\{r_i^m \mid m \in \{F, V\}\right\} \tag{18}$$

Finally, the final predicted score of *MMF* is obtained as follows:

$$S_i = W_d^T R_i + b_d \tag{19}$$

where $W_d \in \mathbb{R}^{2D}$ and $b_d \in \mathbb{R}$. They are both trained by minimizing the MSE loss:

$$\mathcal{L}_{mse}'' = \frac{1}{N} \sum_{i=1}^{N} (y_i - S_i)^2 \tag{20}$$

V. Results

A. Database Setting

In collaboration with a pediatric hospital, Guangzhou Women and Children's Medical Center, we have collected the Multi-modal Postoperative Pain database (MPP) from postoperative children. To promote data collection, a smartphone application was created to enable the easy uploading of videos capturing children's facial and vocal expressions. Additionally, a web-based platform was established to allow multiple doctors to log in and independently assess the FLACC scores of video samples and generate the average labels of scores.

Until now, our database contains 701 valid video samples with accompanying audio from various patients, which we call the *MPP* Database. Their labels are averaged by several independent FLACC scores, evaluated by 1 to 4 medical doctors. Some scores are based on fewer than 4 doctors due to the absence of some doctors.

The FLACC scale assesses five behavioral components (Face, Legs, Activity, Cry, Consolability) to estimate pain intensity, with each component scored from 0 to 2, summing to a total score ranging from 0 to 10. Although the FLACC has potential observer bias, it was chosen for this study due to its relevance and established use in pediatric pain assessment [5].

A detailed analysis of the FLACC scale's reliability and consistency in our database is provided in Table II. It shows that multiple FLACC assessments are reliable and consistent because all average Intraclass Correlation Coefficients and Cronbach's alpha values exceed 0.85 (ICC,$\alpha > 0.85$) with a significance level of $p < 0.001$.

The average FLACC labels of each sample follow the distribution statically shown in Fig. 4. As illustrated, over 40% of samples fall within the 0-1 score range. It shows that the label distribution is imbalanced, with a greater number of non-painful samples than painful ones. We have constructed this database, referred to as the *MPP* Database, for the purpose of multi-modal postoperative pain assessment in children.

More details of the *MPP* Database are presented in Table III.

TABLE II
THE FLACC SCALE RATINGS OF THE *MPP* DATABASE

Raters	Samples	Proportion	ICC	α
1	63	9.0%	-	-
2	278	39.7%	0.865	0.886
3	295	42.0%	0.904	0.923
4	65	9.3%	0.925	0.942

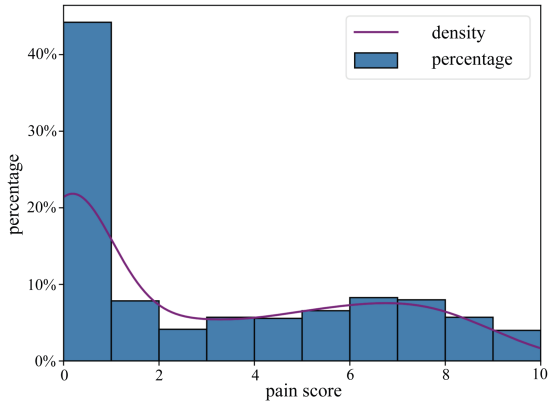


Fig. 4. The distribution of pain scores in the *MPP* Database. The blue bars represent the percentage of the samples falling within specific score ranges, while the purple line depicts the Kernel Density Estimate (KDE) plot based on the overall distribution of pain scores.

TABLE III
DETAILS OF THE *MPP* DATABASE

Characteristic	Value
Number of patients	701
Sex, female/male	204/497
Age, year	4 (3-6 [0-15])
Body weight, kg	17 (14-23 [5-71])
Video time, s	59 (51-60 [2-95])
Pain label	1.5 (0-6 [0-10])
Surgery categories	
-ENT	306 (43.6%)
-Urology	187 (26.7%)
-General	23 (3.3%)
-Orthopedic	48 (6.8%)
-Thoracic	11 (1.6%)
-Oral	58 (8.3%)
-other	68 (9.7%)

Data presented as median (interquartile range [total range]) or number (proportion). ENT = Ear, Nose, and Throat.

B. Evaluation Metrics

We treated the pain assessment as a regression task to approximate the mean score. Similar to the Mean Squared Error (MSE) we adopted as the loss function for the model training shown in (3), for a regression task, two important evaluation metrics are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) measuring the error of the predicted

score compared to the ground truth:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^{pre}| \tag{21}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^{pre})^2} \tag{22}$$

where N is the number of samples, y_i is the ground truth, and y_i^{pre} is the predicted score measured as a result for the i^{th} sample.

In addition to the aforementioned metric, it is crucial to evaluate the similarity of the linear distribution between predicted scores and ground truth values. The Pearson Correlation Coefficient (PCC) is an effective metric for assessing the linear correlation between these two sets of data:

$$PCC = \frac{cov(Y, Y^{pre})}{\sigma_Y \sigma_{Y^{pre}}} \tag{23}$$

Another measure called Concordance Correlation Coefficient (CCC) [35] can be used to evaluate both correlation and consistency between them:

$$CCC = \frac{2cov(Y, Y^{pre})}{\sigma_Y^2 + \sigma_{Y^{pre}}^2 + (\mu_Y - \mu_Y^{pre})^2} \tag{24}$$

where $cov(Y, Y^{pre})$ represents the covariance of the set of ground truths Y and the set of predicted scores Y^{pre}, μ and σ represent the mean value and the variance.

Following the evaluation of research [36], we adopted these four evaluation metrics including MAE, RMSE, PCC, and CCC, to assess the performance of a model in predicting pain scores for this regression task. A better model will have lower MAE and RMSE values and higher PCC and CCC values.

C. Basic Experiment Configurations

Experiments are conducted within the Python environment, utilizing the OpenCV [37] and librosa [32] packages for input processing, and the PyTorch [38] and scikit-learn [39] packages for model construction.

From the *MPP* Database, we preprocessed the facial images and vocal spectrograms by cropping and resizing them to dimensions of 224 × 224 × 3 and 128 × 87 × 1, respectively. We used 5-fold cross-validation to evaluate our models on this database.

Our network comprised a ResNet18 frame extractor and a BiLSTM temporal extractor. By the way, some cross-attention blocks were embedded in our BiLSTM [40]. The projections with all W and b were implemented by MLPs. The feature dimension D was set to 128, as determined by the initial settings of our ResNet18 and BiLSTM extractors. The prototype dimension P was varied between $2^4 = 16$ and $2^8 = 256$ in the sensitivity analysis.

The deep learning algorithm for sample clustering based on features iterated for 1000 epochs, then the prototypes were fixed after 20 epochs of training. The number of experts in our framework ranged from 1 to 6, and these experts were trained for 100 epochs. We employed the Adam optimization with a learning rate of 0.001 and L2 regularization weight decay of 0.0001 to drive deep learning.

D. Comparative Experiments

The pain score obtained through (16) serves as the prediction of our method for a single modality. This method is denoted as *MMF-F* for facial modality, or *MMF-V* for vocal modality. The final fusion score obtained through (19) is the final prediction of our method for multiple modalities, and this method is denoted as *MMF-FV*.

To evaluate our overall *MMF*, we compared it with other pain assessment methods across different modalities. Furthermore, the main strategies in our framework contain sample clustering and multi-expert fusion, we also conducted comparative experiments to analyze the effectiveness of these strategies in the model.

1) Performance of Pain Assessment: Based on our collected *MPP* Database, the *MMF* was employed for several modalities to finish the pain assessment as a regression task, compared with other classical algorithms as follows:

- *Pseudo-Inverse (Pinv) [41]:* This algorithm calculates the least-squares optimal linear mapping matrix by computing the pseudo-inverse from feature inputs to ground truths. The mapping matrix is obtained using training samples and subsequently applied for the prediction of testing samples.
- *Random Forest (RF) [42]:* This algorithm builds multiple decision trees on various samples by randomly selecting prior attributes, using an averaging strategy to aggregate the results. In our study, each RF model consists of 100 trees. For node splitting, the Mean Squared Error is used.
- *Support Vector Machines (SVM) [43]:* This algorithm decides on an optimal hyperplane to divide two categories for classification or to fit the data points for regression. It is also called Support Vector Regression (SVR) when used for regression. In this study, the SVMs utilize the Radial Basis Function (RBF) kernel to enable nonlinear mapping.
- *XGBoost [44]:* This algorithm iteratively generates a new tree model to fit the residuals between the predictions of the previous iteration's tree and the actual values. The predictions from all trees are weighted to provide XGBoost's final output. In this study, the XGBoost models consist of 100 trees over 100 iterations, aiming to minimize the Mean Squared Error in regression tasks.
- *Convolutional Recurrent Neural Network (CRNN) [45]:* This is a combination of a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). The CNN of ResNet18 is applied as the feature extractor of a frame, while the RNN of BiLSTM is applied as the temporal feature extractor of the entire sequence comprised of previous frames. It is suitable for facial expressions (F) and vocal expressions (V) in a video.
- *Transformer [48]:* Similar to RNN, it is also applied as the temporal feature extractor of the entire sequence.
- *Separated Spatial and Temporal 3D Convolutions with Gating (S3D-G) [46]:* Traditional 3D convolutions handle both spatial and temporal information in videos

TABLE IV
DIFFERENT **PAIN ASSESSMENT METHODS** FOR REGRESSION ON THE *MPP* DATABASE FOR DIFFERENT MODALITIES

modality	method	MAE↓	RMSE↓	PCC↑	CCC↑
F	Pinv [41]	1.8859	2.3228	0.6749	0.6085
	RF [42]	1.9704	2.5001	0.6034	0.5590
	SVR [43]	1.8432	2.3531	0.6671	0.6099
	XGboost [44]	1.9802	2.5867	0.5785	0.5470
	CRNN [45]	1.3771	1.9435	0.7866	0.7776
	S3D-G [46]	1.4723	2.0716	0.7576	0.7508
	PET [47]	1.3381	1.9633	0.7882	0.7821
	MMF	1.2714	1.8565	0.8105	0.8063
V	Pinv	1.5962	2.0750	0.7566	0.7474
	RF	1.4272	1.7586	0.8435	0.7848
	SVR	1.4889	1.8700	0.7984	0.7617
	XGboost	1.4860	1.8843	0.7970	0.7672
	CRNN	1.4017	1.9151	0.7902	0.7727
	S3D-G	1.4359	2.1356	0.7560	0.7406
	PET	1.3950	2.0500	0.7811	0.7615
	MMF	1.2946	1.9441	0.8004	0.7850
F+V	Pinv	1.6478	2.1086	0.7534	0.7500
	RF	1.3688	1.7298	0.8372	0.7967
	SVR	1.4903	1.8710	0.7983	0.7611
	XGboost	1.4083	1.8419	0.8147	0.7889
	CRNN	1.1078	1.5013	0.8773	0.8696
	S3D-G	1.1196	1.5281	0.8729	0.8632
	PET	1.1403	1.5943	0.8623	0.8593
	MMF	**1.0399**	**1.4587**	**0.8880**	**0.8875**

F = face, V = voice. ↓ indicates a better model has a lower value, ↑ indicates a better model has a higher value.

simultaneously. S3D-G separates 3D convolutions into spatial and temporal convolutions and introduces a gating mechanism for channel attention. In this study, we directly applied the S3D-G to frame sequences for F and V.

- *PET (Pain Estimate Transformer) [47]:* This method embeds a Bottleneck Attention Module (BAM) [49] into the ResNet, then uses a Transformer encoder to process temporal features and predict the pain level. Similar to the S3D-G, we also directly applied the PET to frame sequences for F and V.

For the features input in Pinv, RF, SVR, and XGboost, different modalities have their distinct manual extraction methods:

- *Face (F):* Manual facial features are extracted using Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) features, producing 30-dimensional histograms for three planes [19].
- *Voice (V):* Manual vocal features are extracted using their Mel-Frequency Cepstral Coefficients (MFCC) features and transformed into 156-dimensional statistical measures such as mean, median, and standard deviation [50].

The feature inputs for other deep learning methods such as CRNN, Transformer, S3D-G, PET, and our *MMF* can be automatically extracted by models.

The results of the comparative experiment on the *MPP* Database are presented in Table IV. For single F and multiple modalities $F+V$, our *MMF* is the best model for all metrics. However, for V, the *MMF* doesn't perform the best on metrics other than MAE, indicating that it improves facial expressions more effectively than voice. Additionally, compared to any single-modal model or other multi-modal fusion methods, *MMF* for multiple modalities $F+V$ demonstrates superior performance in regressive pain assessment.

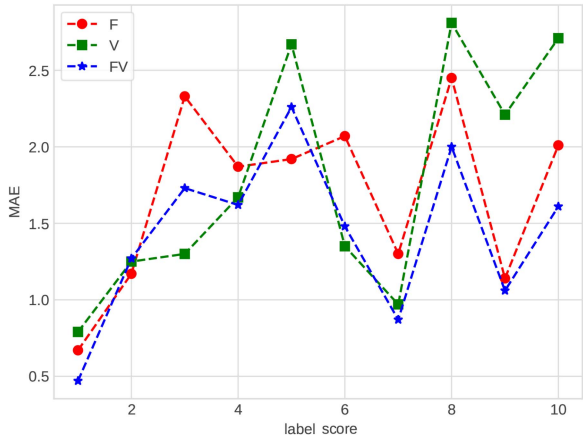


Fig. 5. MAE performance of *MMF* on the *MPP* Database within each specific score range for different modalities.

Moreover, using the Intraclass Correlation Coefficient (ICC), 95% confidence interval (CI), and its significance level to compare the baseline CRNN and our *MMF*, we found that the CRNN achieves an ICC of 0.931 (95% CI: [0.899,0.952], $p < 0.001$), while the *MMF* achieves an ICC of 0.941 (95% CI: [0.941,0.959], $p < 0.001$). These results show that the *MMF* provides higher consistency and reliability between the label and predicted scores, indicating a statistically significant improvement.

As shown in Fig. 5, *MMF* performs better at lower score labels compared to higher score labels due to an imbalance caused by the large presence of non-painful samples. However, the multi-modal framework alleviates the MAE in the high score range. Except for a few low score ranges, combining both facial and vocal modalities results in better performance across different score labels.

2) Effectiveness of Clustering: In addition to our proposed drag-push clustering method, which operates on our feature space, there exist several alternative clustering strategies that can be employed as substitutes. These include:

- *k-means [51]:* The k-means algorithm is a widely-used clustering method that partitions data into K clusters by minimizing the within-cluster sum of squares. It is a classical machine-learning technique that can also be applied to cluster samples based on features without the need for gradient-based parameter optimization.
- *Labels:* In terms of the regression task, it can also cluster samples based on their continuous labels. This approach sorts the training samples in ascending order of their score labels to divide them into K clusters. It encourages multiple experts to concentrate on distinct levels of pain.

We conducted experiments to evaluate the effectiveness of these different clustering strategies based on the *MPP* Database across various modalities. As shown in Table V, neither the k-means nor labels-based approaches are as effective as our proposed method. The k-means approach only updates centers iteratively, without pushing different sets of feature vectors away as our loss function does. While the sample clusters formed

TABLE V
DIFFERENT **CLUSTER STRATEGIES** IN *MMF* ON THE *MPP* DATABASE FOR DIFFERENT MODALITIES

modality	cluster	MAE↓	RMSE↓	PCC↑	CCC↑
F	k-means [51]	1.3636	1.9407	0.7928	0.7894
	labels	1.3536	1.9544	0.7949	0.7917
	MMF	1.2714	1.8565	0.8105	0.8063
V	k-means	1.3883	2.0344	0.7821	0.7760
	labels	1.5861	2.1549	0.7582	0.7474
	MMF	1.2946	1.9411	0.8004	0.7850
F+V	k-means	1.1117	1.5069	0.8764	0.8719
	labels	1.0881	1.4915	0.8791	0.8752
	MMF	**1.0399**	**1.4587**	**0.8880**	**0.8875**

TABLE VI
DIFFERENT **MULTI-EXPERTS METHODS** ON THE *MPP* DATABASE FOR DIFFERENT MODALITIES

modality	method	MAE↓	RMSE↓	PCC↑	CCC↑
F	MoE [24]	1.3438	1.9227	0.7975	0.7846
	SMoE [25]	1.3601	1.9737	0.7896	0.7877
	MMF	1.2714	1.8565	0.8105	0.8063
V	MoE	1.3249	1.9595	0.8001	0.7748
	SMoE	1.3157	1.9213	0.7953	0.7843
	MMF	1.2946	1.9411	0.8004	0.7850
F+V	MoE	1.1056	1.4813	0.8804	0.8733
	SMoE	1.1368	1.5275	0.8723	0.8627
	MMF	**1.0399**	**1.4587**	**0.8880**	**0.8875**

by labels are the same for both modalities, it doesn't consider the different feature distributions across different modalities. Overall, these results indicate that our proposed method is more effective than other cluster strategies.

3) Effectiveness of Multi-Expert Fusion: In contrast to conventional approaches that utilize the entire sample set for training multiple experts, we proposed our multi-experts framework based on the distinct regions from the whole database. We conducted the comparative experiments based on the *MPP* Database, using the following multi-expert methods:

- *MoE [24]:* This approach employs a gating network to automatically generate learning weights for aggregating features from multiple experts based on the entire database. To be fair, we set the number of experts ranges from 2 to 6 to get the best performance of MoE.
- *SMoE [25]:* It indicates the Sparse MoE. This method builds upon the MoE framework, utilizing a gating network to assign learning weights. Although increasing expert capacity, it employs a topK sampling strategy to limit the number of activated experts. Restricted by the small amount of our database, the number of experts for each modality is set to 10. Various topK samplings are tested to get the best performance of SMoE for fairness, where K ranges from 2 to 6.

Table VI compares the performance of various multi-expert methods. It suggests that our multi-expert fusion of *MMF* is more accurate in pain assessment compared to MoE and SMoE, although SMoE possesses a larger capacity of experts.

E. Ablation Studies

MMF for multiple modalities (*MMF-FV*) comprises *MMF-F* and *MMF-V*. Each of these uni-modal parts can be further divided into multiple components, including:

TABLE VII
ABLATION OF *MMF* FOR **FACE** MODALITY

MMF-F with			metric			
clt	*pro*	*fsn*	MAE↓	RMSE↓	PCC↑	CCC↑
			1.3771	1.9435	0.7866	0.7776
		✓	1.3667	1.8565	0.8066	0.7945
✓		✓	1.3721	1.8714	0.8017	0.7889
	✓	✓	1.3331	1.9283	0.7953	0.7919
✓			1.4832	1.9958	0.7769	0.7306
✓	✓	✓	**1.2714**	**1.8561**	**0.8105**	**0.8063**

clt = clustering optimization, pro = prototype, fsn = confidence-based fusion. ✓indicates *MMF-F* is with this component, while a blank indicates it does not.

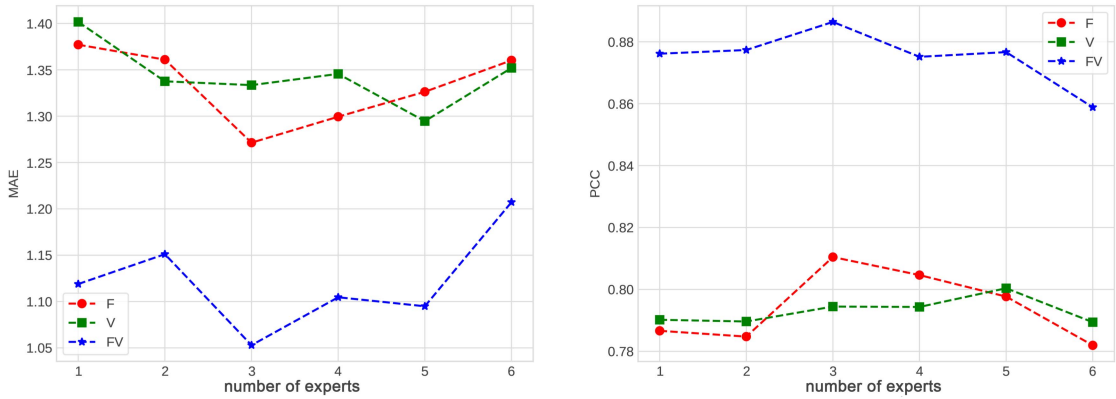
TABLE VIII
ABLATION OF *MMF* FOR **VOICE** MODALITY

MMF-V with			metric			
clt	*pro*	*fsn*	MAE↓	RMSE↓	PCC↑	CCC↑
			1.4017	1.9151	0.7902	0.7727
		✓	1.3462	2.0305	0.7893	0.7751
✓		✓	1.3760	2.0884	0.7726	0.7628
	✓	✓	1.3318	1.9466	0.7925	0.7843
✓			1.3385	**1.9029**	0.7964	0.7722
✓	✓	✓	**1.2946**	1.9441	**0.8004**	**0.7850**

- *clt:* This component represents our sample clustering optimization based on the feature space. When removed, the loss function in (5) is not activated, but the initial random K sample clusters are still retained according to (4). This means that sets of feature vectors are still formed by the nearest features, but their assignment is not optimized for minimum inner distance.
- *pro:* This component represents the transformation from the feature space to the prototype space. When removed, the loss function in (12) is not activated, and confidence computation is performed directly on the feature space without convergence of sets of feature vectors.
- *fsn:* The confidence-based fusion component. For the uni-modal part, it represents the fusion of results from different experts according to (16). For the multi-modal framework, it represents the fusion of features from different experts according to (18). When removed, confidence computation in (15) is not activated, and the previous *pro* component also loses its significance for confidence computation. However, the clustering result of *clt* remains, allowing for the training of different experts. In this case, the mean value of results or features from multiple experts replaces the weighted-sum value by confidence.

To evaluate the individual contributions of different components and modalities to the overall performance of *MMF*, we conducted a series of ablation experiments. In these experiments, we removed different components or uni-modal parts from the framework and measured their impact on performance using various metrics.

Tables VII and VIII present the results of ablation experiments conducted on the *MPP* for facial (*MMF-F*) and vocal (*MMF-V*) modalities, respectively. These experiments involved the removal of various components of the uni-modal part, including clt, pro, and fsn. The results indicate that the removal of any of these components leads to a decrease in performance. However, the removal of the clt component has a relatively



(a) (b)

Fig. 6. Different results with different **numbers of experts**. (a) Mean Absolute Error (MAE); (b) Pearson Correlation Coefficient (PCC).

TABLE IX
ABLATION OF *MMF* FOR MULTIPLE MODALITIES

MMF-FV	MAE↓	RMSE↓	PCC↑	CCC↑
without *V*	1.2714	1.8561	0.8105	0.8063
without *F*	1.2946	1.9441	0.8004	0.7850
without *clt+pro*	1.1754	1.5331	0.8722	0.8618
without *fsn*	1.1261	1.5109	0.8773	0.8727
with all	**1.0399**	**1.4587**	**0.8880**	**0.8875**

low impact on performance due to the residual effect of the initial clusters based on the nearest features. This suggests that even in the absence of optimization via our clustering algorithm, multiple experts trained on distinct data domains remain crucial for *MMF* performance. Furthermore, the ablation results suggest that the removal of the fsn component has a smaller impact on performance for the vocal modality compared to the facial modality. Nonetheless, removing the fsn component still has a negative impact on other performance metrics apart from RMSE for the vocal modality. Overall, these results suggest that all components (clt, pro, and fsn) are essential for achieving optimal performance in both facial and vocal modalities.

Table IX presents the results of ablation experiments conducted on the *MPP* for multiple modalities (*MMF-FV*). These experiments involved the removal of various components or modalities from the framework. The results indicate that removing any uni-modal part has a larger impact on performance compared to removing clt, pro, or fsn components. Moreover, confidence computation based on clt and pro is more important than fsn. Overall, these results suggest that both modalities (F and V) are essential for achieving optimal performance in the *MMF-FV* framework. The smaller components also have an impact on performance, but their removal has a smaller impact compared to removing either modality.

F. Sensitivity Analysis

1) The Number of Experts: The number of experts K in the uni-modal part is a crucial hyperparameter of the *MMF*. As illustrated in Fig. 6, we conducted experiments with different numbers of experts K, ranging from 1 to 6, to analyze the performance of MAE and PCC for both *MMF-F* and *MMF-V*, as well as their combination *MMF-FV*. The performance of *MMF-FV* is consistently better than either *MMF-F* or *MMF-V* alone across all numbers of experts. Additionally, as the number of experts increases, their performances generally improve and then decline. This suggests that there is an optimal number of experts for the *MMF*, and adding more experts may overfit the smaller amount of training data. As a result, adopting 3 experts is the best choice for facial expression while choosing 5 experts is optimal for vocal expression. This suggests that the optimal number of experts may vary depending on the modality due to different feature distributions. As *MMF-F* has a higher effective improvement than *MMF-V*, the best result of *MMF-FV* combined by using the same number of experts for each uni-modal part is achieved with 3 experts.

To determine the optimal combination of experts for the uni-modal part with varying values of K, we conducted an evaluation of the MAE performance of the *MMF-FV* model using different combinations of experts for the F and V modalities. The results are illustrated in Fig. 7, where deeper shades of blue correspond to better performance in terms of MAE and PCC. The performance of either the facial or vocal model alone is inferior to that of the multi-modal framework. The lowest MAE of 1.040 is achieved when combining 5 experts for the facial modality and 3 experts for the vocal modality, as indicated by the deepest shade of blue. PCC also achieves its highest value at 0.888 simultaneously. However, increasing the number of experts beyond this point may result in overfitting and diminished performance.

2) The Dimension of Prototypes: Another critical hyperparameter of the *MMF* is the dimension of the prototypes. Similar to the previous experiments, we also conducted the sensitivity analysis on this hyperparameter, exploring prototype dimensions P ranging from $2^4 = 16$ to $2^8 = 256$. We evaluated the impact on the performance metrics, MAE and PCC, for *MMF-F*, *MMF-V*, and their combined *MMF-FV*. As illustrated in Fig. 8, *MMF-FV* outperforms both *MMF-F* and *MMF-V* across all prototype dimensions. Notably, *MMF-FV* achieves the best performance on both MAE and PCC at a prototype dimension

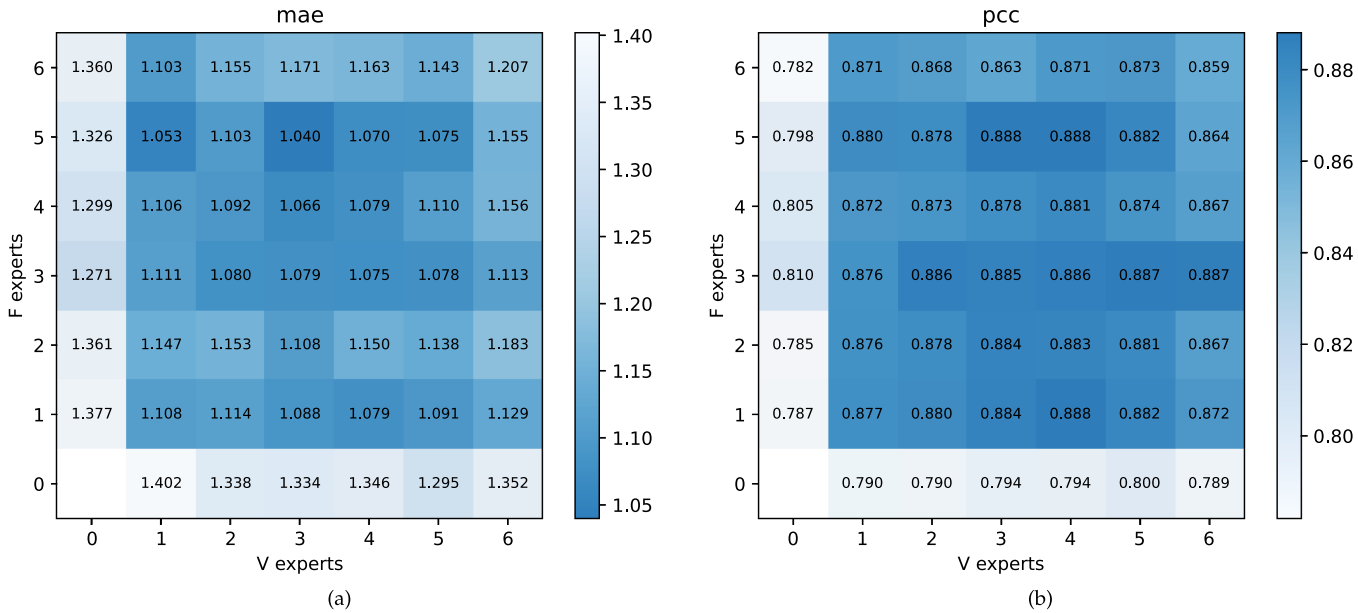


Fig. 7. Different results of *MMF-FV* with different combinations of **numbers of experts** for different modalities. Deeper shades of blue indicate better performance. (a) Mean Absolute Error (MAE); (b) Pearson Correlation Coefficient (PCC).

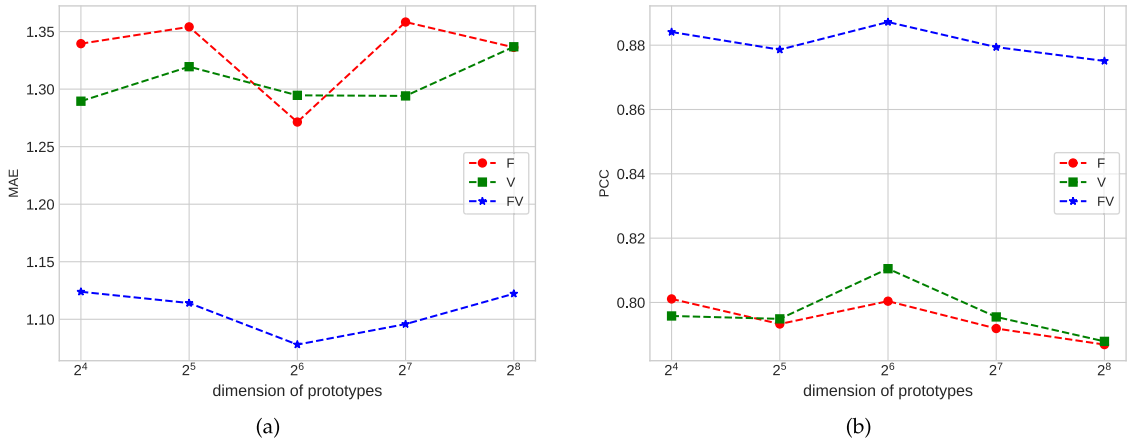


Fig. 8. Different results with different **dimensions of prototypes**. (a) Mean Absolute Error (MAE); (b) Pearson Correlation Coefficient (PCC).

of $2^6 = 64$, which we have identified as the optimal setting for this hyperparameter.

G. Case Studies

In Fig. 9, the feature space is depicted with ground truth and sample cluster results based on the feature space and prototype space of our proposed method. It is an instance of *MMF-F* with 3 experts. The ground truth reveals that a wider space is occupied by unpainful samples, which are colored blue. *MMF* employed deep learning to search for optimal centers to group similar feature representations together. The sample cluster results of *MMF* were similar to those obtained using k-means based on features, but the performance listed in Table V shows that our method outperforms k-means. Additionally, each set of feature vectors converged to an elliptical shape as a Gaussian distribution in prototype space, satisfying a rational assumption for confidence computation. Overall, these results indicate that our clustering algorithm is a promising method for grouping samples based on feature distribution.

For the interpretability of the model, we used Grad-CAM++[52], a technique for visualizing the regions in an image that were most relevant for a deep neural network's prediction, to analyze the facial features that contributed to pain assessment. Fig. 10 shows some examples of Grad-CAM++ heatmaps for different models. We can observe that the baseline model, which used a single pipeline, mainly focused on the areas around the eyes and mouth. These are known to be expressive in pain. However, the first expert model of *MMF-F*, which was trained on a subset of data with higher pain scores, paid more attention to

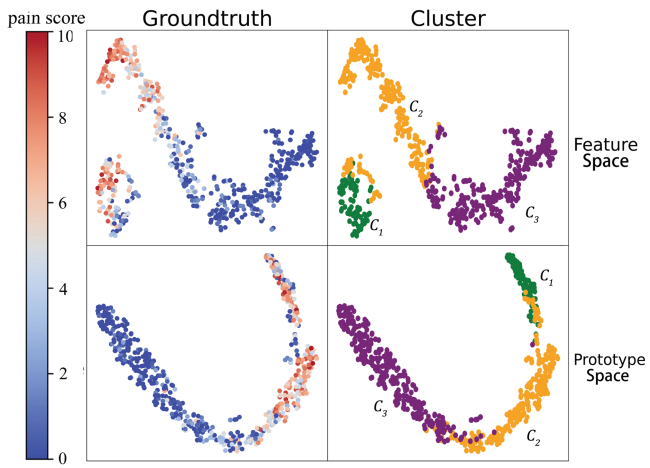


Fig. 9. T-SNE plots of the feature vectors, including the ground truth and sample clusters based on feature space and prototype space of the *MMF* for modality *F*. In the first column, the ground truth is denoted by cool and warm colors, with warmer colors indicating a higher level of pain. In the second column, the feature vectors are clustered and displayed in different colors, such as green, orange, and purple.

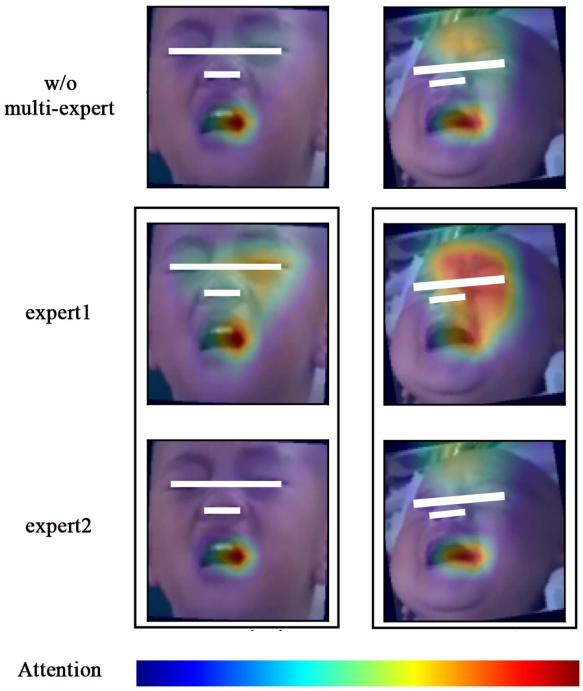


Fig. 10. Grad-CAM++ Heatmaps for facial expressions of two samples. Each sample is a randomly selected frame from a video that is processed by a model of CNN and RNN to extract pain features. The first row shows the results of a single-pipeline model without experts, and the second and third rows show the results of multiple expert models of *MMF*. The facial regions with warmer colors indicate higher attention from the model, which means they contribute more to the pain assessment.

the glabellar wrinkles (the vertical lines between the eyebrows), the closed eyes, and the nasolabial folds (the creases from the nose to the mouth corners). The second expert model maintained attention on the pull-down corners of the mouth and the glabellar wrinkles in the second sample. These features are consistent with the Facial Action Coding System (FACS) [53], which defines the facial muscle movements that correspond to different emotions. The result indicates that different experts of *MMF* are able to capture various facial expression components and identify more informative regions than the single pipeline.

H. Computational Cost

Due to the complexity of our multi-step training process, defining a precise training cost is difficult. However, we can provide specific details regarding the testing process in our multi-expert multi-modal pain assessment using the *MMF*.

With a sampling rate of 1 frame per second, the *MMF* averages 0.535 seconds to predict a 60-second video sample consisting of 60 facial frames and corresponding vocal frames, which equates to approximately 0.009 seconds per frame. The results highlight the efficiency of the *MMF* in processing multi-modal data.

VI. Conclusion

In conclusion, our proposed *MMF* demonstrated superior performance in pain assessment compared to other methods. By employing multiple experts for each modality, our framework effectively addresses the challenges of individual variation of pain expressions and imbalanced label distribution. The proposed framework has potential applications in clinics, where it can provide objective and efficient pain monitoring for postoperative children who cannot communicate their pain verbally. In the future, we plan to extend our framework to include more modalities, such as biological metrics and body movements, and to further improve its performance and interpretability in pain assessment.

Acknowledgment

The collection of the *MPP* database received approval from the Review Board of Guangzhou Women and Children's Medical Center, under the Institutional Review Board (IRB) 044A01. It was registered with the Chinese Clinical Trial Registry, bearing the ID ChiCTR2200058967, with Bilian Li as the Principal Investigator. The registration was completed on April 21, 2022, prior to the enrollment of the first patient. The data supporting the findings of this study are available from the corresponding author upon request.

Written informed consent was secured from the guardians of all subjects who were recruited for this study before any surgical procedures were carried out. This study followed the ethical guidelines set forth by the revised Declaration of Helsinki of the World Medical Association.

References

[1] S. Dolin, J. Cashman, and J. Bland, "Effectiveness of acute postoperative pain management: I evidence from published data," *Brit. J. Anaesth.*, vol. 89, no. 3, pp. 409–423, 2002.

[2] C. Schnakers, C. Chatelle, S. Majerus, O. Gosseries, M. De Val, and S. Laureys, "Assessment and detection of pain in noncommunicative severely brain-injured patients," *Expert Rev. Neurotherapeutics*, vol. 10, no. 11, pp. 1725–1731, 2010.

[3] L. J. DeLoach, M. S. Higgins, A. B. Caplan, and J. L. Stiff, "The visual analog scale in the immediate postoperative period: Intrasubject variability and correlation with a numeric scale," *Anesth. Analg.*, vol. 86, no. 1, pp. 102–106, 1998.

[4] G. Garra, A. J. Singer, A. Domingo, and H. C. Thode Jr, "The Wong-Baker pain faces scale measures pain, not fear," *Pediatr. Emerg. Care*, vol. 29, no. 1, pp. 17–20, 2013.

[5] T. Voepel-Lewis, J. R. Shayevitz, and S. Malviya, "The FLACC: A behavioral scale for scoring postoperative pain in young children," *Pediatr. Nurs.*, vol. 23, no. 3, pp. 293–297, 1997.

[6] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011, pp. 57–64.

[7] S. Walter et al., "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *Proc. IEEE Int. Conf. Cybern.*, 2013, pp. 128–131.

[8] Y. Abdulaziz and S. M. S. Ahmad, "Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients," in *Proc. Int. Conf. Inf. Retrieval Knowl. Manage.*, 2010, pp. 260–263.

[9] V. K. Mittal, "Discriminating the infant cry sounds due to pain vs. discomfort towards assisted clinical diagnosis," in *Proc. 7th Workshop Speech Lang. Process. Assistive Technol.*, 2016, pp. 7–42.

[10] L. S. Franck, D. Ridout, R. Howard, J. Peters, and J. W. Honour, "A comparison of pain measures in newborn infants after cardiac surgery," *PAIN*, vol. 152, no. 8, pp. 1758–1765, 2011.

[11] D. J. Mellor and K. J. Stafford, "Physiological and behavioural assessment of pain in ruminants: Principles and caveats," *Alternatives Lab. Animals*, vol. 32, no. 1_suppl, pp. 267–271, 2004.

[12] T. H. Wideman, R. R. Edwards, D. M. Walton, M. O. Martel, A. Hudon, and D. A. Seminowicz, "The multimodal assessment model of pain: A novel framework for further integrating the subjective pain experience within research and practice," *Clin. J. Pain*, vol. 35, no. 3, pp. 212–221, 2019.

[13] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 530–552, Jan.–Mar. 2022.

[14] R. B. Fillingim, "Individual differences in pain: Understanding the mosaic that makes pain personal," *Pain*, vol. 158, no. Suppl 1, pp. S11–S18, 2017.

[15] L. A. Jeni, J. F. Cohn, and F. De L. Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 245–251.

[16] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "SVM classification of neonatal facial images of pain," in *Proc. 6th Int. Workshop Fuzzy Log. Appl.*, Crema, Italy, 2006, pp. 121–128.

[17] D. Harrison et al., "Too many crying babies: A systematic review of pain management practices during immunizations on YouTube," *BMC Pediatrics*, vol. 14, pp. 1–8, 2014.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[19] P. Thiam et al., "Multi-modal pain intensity recognition based on the *SenseEmotion*database," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 743–760, Jul.–Sep. 2019.

[20] M. Velana et al., "The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system," in *Proc. 4th IAPR Workshop Multimodal Pattern Recognit. Social Signals Hum.-Comput. Interact.*, Cancun, Mexico, 2017, pp. 127–139.

[21] R. Klette, A. Koschan, and K. Schluns, *Computer Vision - Three-Dimensional Data From Images*. Singapore: Springer, 1998.

[22] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 4148–4153.

[23] M. Kächele et al., "Adaptive confidence learning for the personalization of pain intensity estimation systems," *Evolving Syst.*, vol. 8, pp. 71–83, 2017.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[25] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, 2017, pp. 1–19.

[26] D. Lepikhin et al., "Gshard: Scaling giant models with conditional computation and automatic sharding," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–23.

[27] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with LIMoE: The language-image mixture of experts," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 9564–9576.

[28] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

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

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK) : A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 94–101.

[31] N. Sato and Y. Obuchi, "Emotion recognition using MEL-frequency cepstral coefficients," *Inf. Media Technol.*, vol. 2, no. 3, pp. 835–848, 2007.

[32] B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. SciPy*, 2015, pp. 18–24.

[33] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5–6, pp. 602–610, 2005.

[34] J. Lv, K. Liu, and S. He, "Differentiated learning for multi-modal domain adaptation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1322–1330.

[35] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[36] J. O. Egede et al., "Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2020, pp. 849–856.

[37] G. Bradski, "The openCV library," *Dr Dobb's J.: Softw. Tools Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[38] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[39] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[40] A. Pandey and D. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1374–1385, 2022.

[41] F. Schwenker, C. Dietrich, C. Thiel, and G. Palm, "Learning of decision fusion mappings for pattern recognition," *Int. J. Artif. Intell. Mach. Learn.*, vol. 6, pp. 17–21, 2006.

[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[43] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.

[44] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[45] P. Rodriguez et al., "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3314–3324, May 2022.

[46] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[47] H. Xu and M. Liu, "A deep attention transformer network for pain estimation with facial expression video," in *Proc. 15th Chin. Conf. Biometric Recognit.*, 2021, pp. 112–119.

[48] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[49] J. Park, S. Woo, J. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf. 2018*, Newcastle, U.K., 2018, p. 147.

[50] N. Sriraam and S. Tejaswini, "Infant cry detection and pain scale assessment: A pilot study," *Int. J. Biomed. Clin. Eng.*, vol. 3, no. 1, pp. 42–51, 2014.

[51] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. Roy. Stat. Society. Ser. C (Appl. Stat.)*, vol. 28, no. 1, pp. 100–108, 1979.

[52] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[53] P. Ekman and W. V. Friesen, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.*, 1978.

Zequan Liang received the BS degree in computer science and engineering in 2021 from the South China University of Technology, Guangzhou, China, where he is currently working toward the MS degree in computer science. His research interests include multimodal fusion, data mining, and artificial intelligence.

Hao Luo received the BS degree in anesthesiology in 2018. He is currently working toward the MS degree in anesthesiology with Guangzhou Medical University, Guangzhou, China. His research interests include medium and deep sedation therapy for children and children's pain diagnosis and treatment.

Xi Chen received the BEng (with Hons.) degree in electrical and electronic engineering from University College London, London, U.K., in 2023. He is currently interning as a research assistant with the South China University of Technology, Guangzhou, China. His research interests include multi-modal data fusion, artificial intelligence, and integrated systems.

Zhipeng Zhong received the bachelor's degree in anesthesiology in 2022. He is currently working toward the master's degree in anesthesiology with Guangzhou Medical University, Guangzhou, China. His research interests include pediatric anesthesia, thoracic surgical anesthesia, postoperative pain management, and nasal-brain-targeted medication.

Cheng Fan received the BS degree in anesthesiology in 2019. He is currently working toward the MS degree in anesthesiology with Guangzhou Medical University, Guangzhou, China. His research interests mainly include pediatric anesthesia and pediatric pain diagnosis and treatment.

Xingrong Song received the master's degree in anesthesiology from Sun Yat-sen University Medical College, Guangzhou, China. He is currently the chief physician and the head with the Department of Anesthesiology, Guangzhou Women's and Children's Medical Center, affiliated with Guangzhou Medical University. He has authored or coauthored more than ten papers in his research interests which include pediatric anesthesia and neuroprotection. He has also undertaken several research projects related to pediatric anesthesia.

Bilian Li received the MS degree in anesthesiology from Guangzhou Medical University, Guangzhou, China. She is currently a chief physician with the Department of Anaesthesiology, Guangzhou Women's and Children's Medical Center, and also a postgraduate supervisor. She has authored or coauthored more than 20 SCI articles, one of which has been Lancet Journal. Her research interests include medium and deep sedation therapy for children, transnasal-brain targeted drug delivery, and treatment of children with neurological disorders.

Jianming Lv (Member, IEEE) received the BS degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2002, and the PhD degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing, China, in 2008. He is currently a professor with the South China University of Technology, Guangzhou, China. He has authored or coauthored more than 60 papers in some prestigious journals and conferences including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Affective Computing*, CVPR, ACM MM, and CIKM. His research interests include artificial intelligence, machine learning, and computer vision.