MCNet: Meta-Clustering Learning Network for Micro-Expression Recognition

Ziqi Wang[a], **Wenwen Fu**[a], **Yue Zhang**[a], **Jiarui Li**[a], **Wenjuan Gong**[a,*], **Jordi Gonzàlez**[b]
[a]China University of Petroleum (East China), Qingdao Institute of Software, College of Computer Science and Technology, Computer Technology, Changjiang West Road, Qingdao, China, 266580
[b]Universitat Autònoma de Barcelona, Computer Vision Center, Edifici O, Bellaterra, Spain, 08193

Abstract. Facial micro-expressions are categorized into various types based on different criteria, and typically each major category is further divided into multiple subcategories of expressions. For traditional micro-expression recognition problems, multiple subcategories of the same emotions are indiscriminately learned and verified, leading to potential mis-classification, especially with negative emotions. To address the issue of intra-class variation in micro-expressions, we propose a novel meta-clustering learning network for micro-expression recognition called MCNet. This approach integrates the ideas of meta-learning and clustering, hierarchically clustering subcategories within a micro-expression class to generate multiple class centers for metric-based classification. The proposed method diverges from the common strategy of metric-based meta-learning algorithms, which typically use the mean feature of all samples within the same class as the class center. Furthermore, we incorporate transfer learning into the meta-learning process to jointly alleviate overfitting caused by the scarcity of micro-expression data. We conduct extensive comparative experiments based on the Leave-One-Subject-Out protocol on three widely used micro-expression datasets. The experimental results demonstrate the competitive performance and strong generalization ability of the proposed MCNet approach.

Keywords: meta-learning network, hierarchical clustering, few-shot classification, micro-expression recognition.

*Wenjuan Gong, wenjuangong@upc.edu.cn

1 Introduction

Facial expressions play a crucial role in conveying people's emotional states and psychological fluctuations, making them integral to interpersonal communication.[1] Therefore, facial expression recognition has consistently remained a prominent area of research in computer vision. Micro-expressions, characterized by their short duration and subtle facial muscle movements,[2] are often imperceptible to the naked eyes. However, they offer a fresh perspective for investigating human emotional states, as they can accurately reflect emotions, even when individuals deliberately conceal their true feelings. Moreover, micro-expression recognition holds immense value across various domains, including lie detection, psychological assessment, negotiation, and interrogation.

Despite the potential applications of micro-expression recognition, detecting and interpreting micro-expressions accurately with the naked eyes remains a formidable challenge. Even experts find it difficult to manually identify these subtle facial movements. To alleviate this challenge, psychologists have developed auxiliary tools such as the Micro Expression Training Tool (METT)[3] and the Facial Action Coding System (FACS)[4] to facilitate the manual identification of micro-expressions. However, manual micro-expression recognition still faces several obstacles. Firstly, the ability to discern micro-expressions is limited to extensively trained professionals, which incurs high costs. Secondly, even with intensive training, humans can only recognize approximately 47% of micro-expressions.[5] Moreover, human analysis of micro-expressions is time-consuming, expensive, and prone to errors. Consequently, the development of automated systems for micro-expression analysis has become imperative.

In recent years, the availability of several spontaneous micro-expression datasets[6–9] has sparked increased interest among researchers in the computer vision field to tackle the challenge of micro-expression recognition, resulting in the proposal of numerous recognition algorithms. Automatic micro-expression recognition primarily focuses on extracting salient and distinctive features from micro-expression clips. In one of the early works on micro-expression recognition, Polikovsky et al.[10] introduced the 3D Histogram of Oriented Gradients (3DHOG) descriptor to extract appearance-based micro-expression features. To enhance the compactness and robustness of the features under varying lighting conditions, Huang et al.[11] proposed the SpatioTemporal Completed Local Quantization Patterns (STCLQP) approach. This method extracts information encompassing sign, magnitude, and orientation differences, and efficiently applies vector quantization and codebook selection for each component in both spatial and temporal domains. Finally, the features from the three components are combined into one feature based on the feature subspace. Additionally,

dynamic micro-expression features based on optical flow have demonstrated notable performance in micro-expression recognition tasks.[12–15] Optical flow captures the motion patterns of objects or scenes in an image sequence by detecting pixel intensity changes between frames. For example, Liu et al.[14] proposed the Main Direction Mean Optical flow (MDMO) feature, which integrates the magnitude and direction of the primary optical flow vectors across 36 non-overlapping Regions of Interest (ROIs) in the face, effectively reducing the feature dimensionality.

Different from the aforementioned handcrafted feature extraction methods, deep learning approaches adopt the end-to-end manner by integrating feature extraction and classification. Patel et al.[16] were the first to introduce Convolutional Neural Networks (CNNs) to micro-expression recognition and mitigated overfitting due to the lack of sufficient micro-expression samples by using transfer learning and feature selection algorithms. Similarly, Kim et al.[17] employed a framework based on CNN and Long Short-Term Memory (LSTM) networks to learn spatial and temporal features of micro-expressions and incorporated expression states into the objective function to enhance category separability. However, both CNN and LSTM models require a substantial number of training samples, which poses a challenge due to the limited availability of micro-expression datasets and increases the risk of overfitting. To alleviate the issue of overfitting, many works[18–22] have adopted transfer learning and data augmentation. For example, Zhi et al.[21] employed transfer learning by pre-training a 3DCNN on the macro-expression dataset Oulu-CASIA[23] and transferring the learned knowledge to the micro-expression domain. Sun et al.[22] proposed a novel knowledge transfer method that extracts and transfers knowledge from facial action units for micro-expression recognition. Takalkar et al.[18] and Zhang et al.[24] expanded the dataset by combining and horizontally flipping the data, respectively.

Despite the advancements made in deep learning-based approaches for micro-expression recog-

nition, these works have not achieved remarkable results and even performed worse than some traditional methods. To address this issue, Peng et al.[25] introduced a Dual Temporal Scale Convolutional Neural Network (DTSCNN) that can handle micro-expression segments with different frame rates. They incorporated optical flow sequences extracted from the original micro-expression clips as inputs to the network, leading to a significant improvement in the overall recognition accuracy. Inspired by this success, subsequent works[26–30] have also adopted the strategy of extracting optical flow information as input to their CNN models. In addition to handling discriminative representations of micro-expression sequences from a dynamic feature perspective, deep learning methods based on attention mechanisms have also demonstrated success [31–34]. For instance, Zhou et al. [32] proposed the Dual-branch Attention Network (Dual-ATME), comprising Hand-crafted Attention Region Selection (HARS) and Automated Attention Region Selection (AARS). The HARS manually extracted features from the Region of Interest (ROI) using prior knowledge, while AARS automatically extracted hidden information from the sequence based on attention mechanisms. This method effectively learned micro-expression representations by the dual-scale features, and demonstrated that the influence of facial visual features and the visual relationships between facial ROIs in distinguishing micro-expressions. Therefore, Thuseethan et al. [35] introduced a novel end-to-end facial micro-expression detection framework named Deep3DCANN. This method employed a deep 3D convolutional neural network to learn spatiotemporal features from facial expression sequences and a deep artificial neural network to capture visual associations between different facial sub-regions. The combination of output features from both pipelines allowed for the identification of micro-expression changes on frame-level and showed excellent performance in multi-class classification tasks.

Furthermore, integrating deep features with handcrafted features in a dynamic facial micro-

expression recognition architecture has proven effective in enhancing performance. For example, Wang et al. [36] proposed the Local Cube Binary Pattern Spatial-Temporal Graph Convolutional Network (LCBP-STGCN), where STGCN consists of a Spatial Graph Convolutional Network (SGCN) for acquiring spatial information and a Temporal Convolutional Network (TCN) for capturing temporal information related to micro-expressions. This method introduced low-level spatio-temporal features obtained from Local Cube Binary Pattern, and then the spatio-temporal graph convolutional network was used to extract high-level features of micro-expression sequences. The combination of these structures preserved potential discriminative information in the sequence, improving the robustness of feature extraction. Although notable progress has been made by these methods in the field of micro-expression recognition, the limited availability of data samples and the risk of overfitting still pose challenges during the training process. Therefore, further improvements in the performance of micro-expression recognition can still be pursued.

To solve this problem, the meta-learning approaches have been proposed. Chelsea Finn et al. [37] introduced the idea of Model-Agnostic Meta-Learning (MAML). The goal of MAML is to enable the model to quickly adapt to new tasks through a small number of training steps and data. By performing gradient updates across multiple tasks, MAML finds initialized parameters, from which the model can achieve good generalization using a small amount of sample data within a few update steps. MAML, with its model-independent advantages, achieves the effect that can be used with any model structure, and has become an important tool in the field of meta-learning. Ravi et al.[38] proposed a novel meta-learning method by training an optimizer to enable it to quickly adapt to new tasks and achieve good performance in a few-shot learning task. This method is known as learning optimization (Learning to Optimize) or LSTM-meta-optimizer. The core idea is to let the model itself learn how to perform optimization, rather than relying on manually designed opti-

mization algorithms such as stochastic gradient descent (SGD). By using recurrent neural network to update the parameters of the model, the optimizer receives the information of the loss function during learning and outputs the parameter values to be changed. In the meta-learning phase on a specific task, the LSTM optimizer will decide how to do the next parameter update based on the past parameter update history and the gradient of the target function. In the new task, the LSTM optimizer can generate parameter updates when the model needs rapid adaptation, which significantly improves the performance of few-shot learning.

To mitigate the problem of limited data samples in micro-expression recognition, a meta-learning-based multi-model fusion network (Meta-MMFNet) was proposed in,[24] which was the first application of meta-learning in the field of micro-expression recognition. This method employed a Prototypical Network (PN) which represented each class by its average feature, and fused the transfer learning stream through a weighted sum approach. The Meta-MMFNet method achieves comparable performance to state-of-the-art methods even with a small number of training samples, thus alleviating the challenge of overfitting to some extent. However, it is worth noting that this metric-based meta-learning micro-expression recognition algorithm treats the mean feature as the class centroid and adopts the nearest centroid method for classifying unseen samples. From the perspective of micro-expression categories, certain classes may have more than one centroid under specific situations. For instance, the category of "negative" in the SMIC dataset comprises multiple sub-emotions such as "disgust", "repression" and "sadness", each with its own centroid. Moreover, individual differences among participants introduce additional factors unrelated to expressions, such as race and gender. Furthermore, due to variances in facial muscle movement habits, there exist substantial differences between different samples of the same expression category. Collectively, these factors contribute to a considerable intra-class distance within

micro-expressions.

To tackle this issue, we present a novel meta-clustering learning network named MCNet, drawing inspiration from hierarchical clustering method.[39] Hierarchical clustering is an algorithm that constructs a nested clustering tree with hierarchies by assessing the similarity between data points of various categories, including bottom-up merging and top-down splitting strategies. In this study, we focus on learning the bottom-up merging concept, where the two most similar data points are successively merged. We set the stopping criterion for clustering when the data points to be merged belong to different classes.

The proposed MCNet replaces the previous practice of utilizing a single mean class prototype (a centroid) by incorporating a hierarchical clustering module. This module employs multiple clustering prototypes to represent each category or sub-category. Specifically, when a category encompasses multiple distinct sub-categories, each with unique features, the hierarchical clustering module clusters micro-expression samples based on the feature differences among sub-categories, resulting in the formation of multiple centroids. This ensures a comprehensive modeling of intra-category differences, addressing the limitation of relying solely on average features as centroids that may inaccurately represent the distinctive characteristics of each category. Consequently, this enhances MCNet's ability to capture fine-grained features. The hierarchical clustering module facilitates the identification and separation of similar and dissimilar sub-groups within datasets, and the integration of multiple prototypes enhances MCNet's ability to capture the complexity and nuances of each category. This alleviates the discrepancies in characteristics among samples in sub-emotion clusters, thereby leading to more precise sentiment analysis. Additionally, MCNet introduces prior knowledge from both micro-expressions and macro-expressions, further enhancing recognition accuracy on unknown samples.

The main contributions of this study are as follows:

- This study identifies and addresses the mis-classification problem caused by the large intra-class differences in micro-expression samples. To tackle this, we proposes a novel meta-clustering learning network for micro-expression recognition, introducing a hierarchical clustering module that utilizes multiple clustering prototypes instead of a single mean prototype, breaking the simple strategy of using centroids as class prototypes in previous meta-learning methods.

- This study combines transfer learning, meta-learning and clustering principles, which utilizes the sufficient macro-expression data along with meta-learning to alleviate the the accuracy limitations caused by the scarcity of micro-expression data.

- The proposed method is evaluated on three widely-used micro-expression datasets, and the experimental results demonstrate highly competitive performance of the MCNet method.

2 The Meta-clustering Learning Network

Following the meta-learning formulation for micro-expression recognition,[24] we also adopt the N-way K-shot classification task of few-shot learning, where the dataset is typically divided into a meta-training set and meta-testing set. In meta-learning, the dataset is typically divided into a training (also called meta-training) set and a testing (also called meta-testing) set. In the training (also called meta-training) stage, N unique categories are randomly sampled from the meta-training set, from which K and M unique samples are randomly selected from each category to compose the support set Φ_S and query set Φ_Q, respectively. These support sets and query sets constitute distinct meta-tasks, allowing the model to be trained in each iteration. In the meta-testing stage, the

support set and query set are sampled from the testing set using a similar mechanism as the one used in the meta-training stage. By training on different meta-tasks, the model learns generalizable knowledge, enabling it to quickly adapt to new tasks.

Due to the limited availability of data in micro-expression classification, we make adjustments to the traditional meta-tasks using non-overlapping categories, according to the distribution of participants. Specifically, following the Leave-One-Subject-Out (LOSO) evaluation protocol, we set the query set Φ_Q^{test} of the meta-testing set Ω_{test} to include all samples from the participant set Z_{test} for testing. The support set Φ_S of the meta-training set Ω_{train} and the meta-testing set Ω_{test} are composed of samples from the remaining participant set Z_{train}, where $Z_{train} \cap Z_{test} = \varnothing$. This strategy ensures a comprehensive evaluation by sampling from different participants for training and testing, leading to robust and reliable results.

We use a metric-based prototypical network as the meta-learning model. The standard approach has some drawbacks when directly applied for micro-expression classification. The problem lies in that the query samples may be far away from the class prototype when the gap between the samples within the class is too large, thus affecting the classification accuracy of the model. It should be noted that a micro-expression class can be further divided into multiple sub-groups, and there are subtle variations among samples of the same class due to slight muscle movement differences. Therefore, it is difficult to achieve a favorable classification using only the center of samples. To address this issue, we incorporate hierarchical clustering, as illustrated in Fig. 1, in the proposed approach.

The proposed MCNet method aims to address the challenge of large intra-class differences in micro-expression recognition by integrating the idea of hierarchical clustering with a fusion network based on meta-learning. Meta-learning network is trained and tested on a per-task basis,

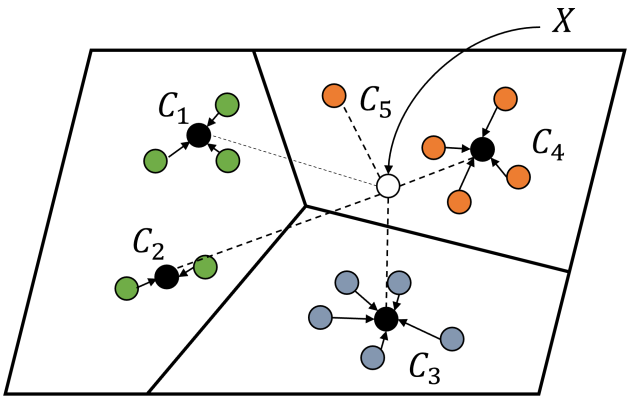


Fig 1 The process of predicting sample categories using the proposed meta-clustering method. Instead of using a single mean feature prototype for each class, the method determines the class prototypes by clustering the two nearest samples within the same class in the support set, resulting in N or more class prototypes, where N represents the number of classes in the support set. To predict the label of a query sample X during meta-testing, the meta-clustering method compares the distances between the sample and each class prototype and identifies the closest class prototype to the sample. The label of the test sample is obtained by mapping the nearest class prototype back to its original category.

where each task comprises a support set and a query set. For each micro-expression sample in meta-tasks, the fused motion feature obtained from the feature fusion module is fed into a pre-trained deep feature extraction module to acquire corresponding deep feature vector. In the metric-based meta-learning module, these feature vectors are utilized to generate multiple class prototypes using hierarchical clustering, allowing for the prediction of class membership for each sample in the query set.

The overall structure of the proposed method is shown in Fig. 2. The proposed method mainly consists of three components: the feature processing module, the meta-learning module and the meta-clustering module. To effectively capture motion features and reduce redundancy in micro-expression clips, we fuse the optical flow features and frame difference features, which are then processed by a deep feature extraction network to obtain comprehensive and discriminative deep features. To tackle the problem of limited micro-expression data, we introduce transfer learning during the pre-training phase and obtain two independent feature extraction networks. Subse-

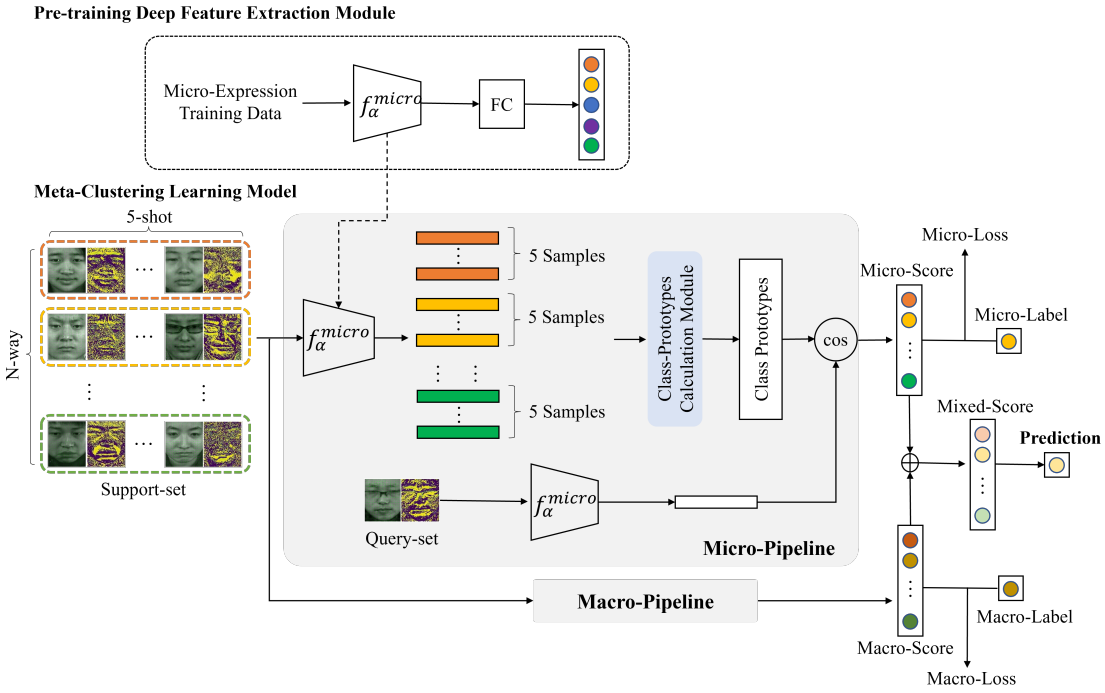


Fig 2 The overall architecture of the meta-clustering learning network. This method comprises of two main stages: pre-training the deep feature extraction module and learning the meta-clustering classification model. In the pre-training stage, two independent deep feature extraction networks based on ResNet18 are trained on the micro-expression dataset Ω_{train} and the macro-expression dataset CK+. In the meta-clustering stage, the fused dynamic features are fed into the micro-pipeline and macro-pipeline and the deep features of micro-expressions are then extracted in each pipeline. The cosine distance between the query samples and the class prototypes obtained from the class prototype calculation module is computed, yielding the micro-score and macro-score, respectively. Based on these two score vectors, we optimize the micro-pipeline and macro-pipeline separately using cross-entropy loss. during the validation process, we combine the score vectors generated from both pipelines to merge the micro- and macro-predictions to predict the final micro-expression category.

quently, the metric-based meta-learning module learns micro-expression classification on a task-based level under two pathways: the micro-pipeline and the macro-pipeline. In the testing stage, the meta-clustering module generates multiple class prototypes for each task to address the challenge of mis-classification caused by intra-class variations. Furthermore, to fully utilize priors from both micro-expressions and macro-expressions, we fuse the two predictions in the last step, further enhancing the recognition accuracy.

2.0.1 *Feature processing module*

Due to subtle variations and brief durations of micro-expression motions, micro-expression sequences often contain redundant information. Previous studies[15,40] have demonstrated that extracting features from the onset and apex frames of a video are more effective. Therefore, frame difference features and optical flow features are extracted from the two frames and then concatenated to obtain the fused feature x_i. Based on this, we represent the support set and the query set of the meta-task as $\Phi_S = \left\{\left(x_s^1, y_s^1\right), ..., \left(x_s^i, y_s^i\right), ..., \left(x_s^{NK}, y_s^{NK}\right)\right\}$ and $\Phi_Q = \left\{\left(x_q^1, y_q^1\right), ..., \left(x_q^i, y_q^i\right), ..., \left(x_q^{NM}, y_q^{NM}\right)\right\}$, where y_i represents the ground-truth label of the i-th sample and M denotes the number of samples in each category of the query set. And x_s^i represents the dynamic feature of the i-th micro-expression sequence in the support set, while x_q^i is the dynamic micro-expression feature of the i-th sample in the query set.

The fused feature serves as input to the pre-training deep feature extraction module for further processing. Collecting micro-expressions is a time-consuming and labor-intensive process, and manual annotations from experts are required, resulting in a severe lack of available data. The success of previous works[41,42] have shown the feasibility of leveraging the abundance of macro-expressions to recognize micro-expressions. To address the data scarcity issue, we introduce the idea of transfer learning into the deep feature extraction module. In this module, we use ResNet-18 as the backbone network and pre-train it on both the micro-expression training set and the macro-expression dataset, respectively. This process yields two trained deep feature extraction networks denoted as f_α^{micro} and f_β^{macro}, where α and β represent the network parameters pre-trained on the corresponding datasets. These two networks are trained independently and have no interference with each other, only being weighted and fused in the final recognition stage to achieve better

predictions.

2.0.2 Meta-learning module

During the training of the meta-learning model, each meta-task consists of a support set with N classes, where each class is sampled with K samples, and a query set with randomly selected M samples from each class. For each task, the meta-learning module processes the fused feature vectors extracted for each sample using the previous module. Specifically, the fused features are processed by networks $f^*, * = \{micro, macro\}$ to adapt to micro-expression meta-tasks based on the priors from micro-expressions and macro-expressions, respectively. The updated feature matrices of a meta-task for each pipeline are represented as $\Psi_S = \left[\psi_s^1, ..., \psi_s^i, ..., \psi_s^{NK}\right]$ and $\Psi_Q = \left[\psi_q^1, ..., \psi_q^i, ..., \psi_q^{NM}\right]$.

To optimize the prototype networks using stochastic gradient descent mechanism, we calculate the average feature vector for each class c as the representation of the corresponding class prototype r_c during the meta-training phase of each data stream. Then, we compute the cosine distance between each query sample feature ψ_q^i and each class prototype to obtain the predicted probabilities of the i-th sample belonging to each class. The prediction is denoted as $V_i = \left[v_i^1, ..., v_i^c, ..., v_i^C\right]$, where C represents the total number of classes in the support set. The predicted probability is formulated as follows:

$$v_i^c = softmax(cos(\psi_q^i, r_c)), \tag{1}$$

where $cos(\cdot)$ calculates the cosine distance between vectors, and $softmax(\cdot)$ represents the activation function. Next, we update the meta-learning processes for the micro-pipeline and the

macro-pipeline using cross-entropy loss, respectively:

$$L = -\sum_{i=1}^{N\times M}\sum_{c=1}^{C} y_q^{i,c}\log v_i^c, \tag{2}$$

where $y_q^{i,c}$ represents the ground-truth label indicating whether the i-th query sample belongs to class c, and v_i^c represents the predicted probability of the i-th query sample belonging to class c.

2.0.3 Meta-clustering module

In the meta-testing stage of a standard prototypical network based solution, the feature vectors of the samples in the testing task are exacted from the deep feature extraction network. Then, the mean feature of all samples of each category in the support set is computed as the corresponding class prototype:[24]

$$r_c = \frac{1}{|S_c|}\sum_{x\in S_c} f^*(x), * = \{micro, macro\}, \tag{3}$$

where S_c is the sample cluster of class c in the support set Ψ_S. After obtaining the class prototypes, the distances between the feature vectors of the query samples and class prototypes are calculated for classification.

To address the issue of large intra-class distances and mitigate prediction errors, we propose an improved clustering method in the testing phase. Specifically, we adopt a hierarchical clustering approach, which aggregates feature vectors from the bottom up based on their similarity. Fig. 1 illustrates the process of the proposed hierarchical clustering module. The module generates a number of prototypes equal to or greater than the number of classes in the support set, which are then used for distance metric calculations. Taking the micro-pipeline for example, each sample in the support set is initially considered as a cluster, i.e., treated as a prototype. For all the feature

vectors in the Ψ_S^{test}, the distances between each pair of vectors are computed, and the two most similar micro-expression data are sequentially identified and merged:

$$x_{new} = \frac{f_\alpha^{micro}(x_i) \times w_i + f_\alpha^{micro}(x_j) \times w_j}{w_i + w_j}$$
$$w_{new} = w_i + w_j \tag{4}$$

where $f_\alpha^{micro}(\cdot)$ is the deep feature extraction network pre-trained for micro-pipeline and α represents the parameter set of this neural network. And, the two feature vectors x_i and x_j from the sample cluster of category c have the same ground-truth class label and the closest distance among all pairs of vectors in the support set. Each vector is assigned a weight, w_i and w_j, respectively. And the aggregated new vector, x_{new}, is obtained by calculating the weighted average of the two vectors. The weight of the new vector is denoted as w_{new}. In this study, initial weights of all samples are 1.

The above process is iterated, where the closest two clusters are merged until two clusters from different classes are encountered, indicating the end of clustering. Each cluster is treated as a separate class, and the distances between the generated class prototypes and the query samples are metricized. There is a mapping relationship between the newly generated classes and the original classes in the support set. The label of the test sample is obtained by mapping the nearest class prototype back to its original category.

Additionally, knowledge from macro-expressions domain can provide valuable insights for micro-expression recognition. To leverage this prior, we utilize a feature extraction network, f_β^{macro}, which has been pre-trained on macro-expressions, and transfer it to the micro-expression domain. In the metric-based meta-learning module, we calculate two vectors of cosine similar-

ities through both the macro-pipeline and micro-pipeline, referred to as macro-score and micro-score, respectively. During the testing process, we further integrate the prior knowledge from both micro-expressions and macro-expressions by combining the outputs of these two data streams in a weighted sum, which defines the final metric distances between samples and the class prototypes. The process is formulated as follows:

$$\begin{aligned} H^i_{total} &= H^i_{micro} + \gamma H^i_{macro}, \\ P_i &= softmax(H^i_{total}), \end{aligned} \tag{5}$$

where γ is a hyperparameter that balances the weight of micro-expression and macro-expression prior knowledge, H^i_{micro} and H^i_{macro} represent the similarity scores obtained from the micro-pipeline and macro-pipeline, respectively, and P_i represents the probabilities of the i-th query sample matching each prototype. The predicted category is the prototype with the highest probability. This fusion of micro- and macro-priors enhances the model's classification performance and helps to improve the accuracy of micro-expression recognition.

3 Experimental results

In this study, we conducted rich experiments on common micro-expressions datasets (including SMIC, CASME, and CASME II datasets) to validate the effectiveness of the metaclustering learning model for micro-expressions classification. The SMIC dataset consists of 164 micro-expression videos from 16 subjects. The database categorizes micro-expressions into three categories, âĂIJpositiveâĂİ, âĂIJnegativeâĂİ, and âĂIJsurpriseâĂİ. The SMIC database includes micro-expression image sequences captured at 25 frames per second using a camera, as well as infrared image sequences. The CASME dataset comprises of 195 video sequences of micro-expressions

captured in natural situations from 26 participants. It covers seven main types of expressions, i.e., âĂIJhappinessâĂİ, âĂIJdisgustâĂİ, and others. The classification task in this experiment mainly focused on four expression types, i.e., âĂIJdisgustâĂİ, âĂIJsurpriseâĂİ, âĂIJrepressionâĂİ, and âĂIJtenseâĂİ. The CASME II dataset contains 256 micro-expression events collected from 35 Asian participants who were asked to watch specific videos to elicit micro-expression. CASME II has the advantages of high resolution, high sampling rate, large sample size, and detailed labeling. The classification task in this experiment mainly focused on four expression types, i.e., âĂIJhappinessâĂİ, âĂIJdisgustâĂİ, âĂIJsurpriseâĂİ, and âĂIJrepressionâĂİ.

Following the LOSO protocol, we compared the effectiveness of the MCNet with several other state-of-the-art methods, utilizing prediction accuracy and F1-score as the evaluation metrics. The results clearly demonstrated that the proposed MCNet method achieved competitive performance. In the experimental process, the model was initially trained using meta-learning methods with specific task settings. Each meta-task is divided into a support set and a query set, with the support set used for model training and the query set used for evaluating its performance. In the assessment of the model's performance, accuracy was used as the primary performance metric. Accuracy represents the proportion of correct classifications by the model for a given task.

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}, \tag{6}$$

where TP stands for True Positives, TN stands for True Negatives, FP stands for False Positives, and FN stands for False Negatives. Additionally, a confusion matrix was employed to provide a more detailed breakdown of the classification results. The confusion matrix reveals the model's classification performance for each category, including true positives, true negatives, false posi-

tives, and false negatives. In addition to accuracy and the confusion matrix, we also employ the F1 score as an evaluation metric.

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, \\ F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall}, \tag{7}$$

where P represents Precision, R represents Recall, and F1 represents the F1 Score. The F1 score combines precision and recall into a weighted average, providing a comprehensive assessment of the model's performance. Precision represents the proportion of true positives among samples predicted as positive, while recall represents the proportion of true positives among all actual positive samples. In the experimental results, a detailed report of accuracy, confusion matrix, and F1 scores was provided for the relevant test tasks, allowing for a comprehensive understanding of the model's performance across tasks and evaluating its strengths and weaknesses.

3.1 Experimental setup

For the meta-learning sampling process for each dataset, we set $K = 5$, and the value of N is determined based on the number of classes in the dataset. For example, for the SMIC dataset, we sampled 3-way 5-shot tasks, and for the CASME and CASME II datasets, we sampled 4-way 5-shot tasks. In the pre-processing phase of the deep feature extraction network, we initialized the parameters with ResNet18 pre-trained on the ImageNet dataset, and trained the model for 50 epochs with a batch size of 128, using an NVIDIA GeForce RTX 3090 GPU. During the meta-learning phase, we kept the learning rate fixed at 0.01 and used a batch size of 8 meta-tasks. For the SMIC dataset, where the vertex frames were not annotated, we used the middle frame of each micro-expression sequence as the vertex frame. For fusing the priors of micro- and macro-

expressions, the hyperparameter γ was learned independently for each dataset. In this study, we randomly selected a subject set Z_{val} from the dataset Z_{total} to learn the optimal γ, which maximized the prediction accuracy on the samples from Z_{val}. The value of γ was chosen between 0.1 and 1 with a step size of 0.1. Notably, once the subject was picked for a dataset, the samples of Z_{val} were not involved in the accuracy validation during the testing phase, i.e., $Z_{total} - Z_{val} = Z_{train} + Z_{test}$.

3.2 Results and analysis

To validate the effectiveness of the proposed meta-clustering network, we conducted comprehensive comparisons with the state-of-the-art methods using LOSO protocol. The methods of comparison follow various evaluation protocols, including LOSO and Leave-One-Video-Out (LOVO). LOSO and LOVO are two cross-validation strategies that are commonly used to evaluate the performance of machine learning models, especially in sentiment analysis, facial expression recognition, or other human behavior analysis LOSO is a method for cross-validation that is particularly useful when dealing with datasets involving multiple participants.The LOVO strategy, on the other hand, involves using the data from one video at a time as a test set and the data from all the remaining videos as a training set in a dataset containing multiple video samples. Again, this step is repeated once for each video sample to ensure that each sample has a chance to be used as test data. In this paper, LOSO is used because LOSO uses all the videos of a participant at a time as the test set, and the model does not learn the micro-expression distribution of that participant, so LOSO validation better reflects the model's adaptability to unknown objects.

The comparison results on the SMIC, CASME, and CASME II datasets were listed in Tables 1, 2, and 3, respectively. To enhance the clarity of the comparison results, we only retained works that employed the same classification criteria as MCNet, excluding those methods evaluat-

Table 1 Performance comparisons with the existing methods on SMIC dataset

Method	Protocol	Task	Accuracy(%)	F1-score
Baseline[6]	LOSO	Negative,Positive,Surprise	45.70	0.46
LBP-SIP[43]	LOSO	Negative,Positive,Surprise	42.12	0.42
Selective[16]	LOSO	Negative,Positive,Surprise	53.60	-
STCLQP[11]	LOSO	Negative,Positive,Surprise	64.02	0.63
FHOFO[13]	LOSO	Negative,Positive,Surprise	51.83	0.52
VGGMag[44]	LOSO	Negative,Positive,Surprise	59.75	0.58
Bi-WOOF[15]	LOSO	Negative,Positive,Surprise	62.80	-
3D-FCNN[28]	LOVO	Negative,Positive,Surprise	55.49	-
SSSN[26]	LOSO	Negative,Positive,Surprise	63.41	0.63
DSSN[26]	LOSO	Negative,Positive,Surprise	63.41	0.65
STRCN-A[45]	LOSO	Negative,Positive,Surprise	53.10	0.51
SVM+revised HOOF[46]	LOSO	Negative,Positive,Surprise	60.67	-
LGCconD[47]	LOSO	Negative,Positive,Surprise	63.41	0.62
FR[48]	LOSO	Negative,Positive,Surprise	57.90	-
Meta-MMFNet[24]	LOSO	Negative,Positive,Surprise	63.13	-
Dual-ATME[32]	LOSO	Negative,Positive,Surprise	64.60	-
SVM+LCBP-STGCN[36]	LOSO	Negative,Positive,Surprise	75.51	0.74
MFVAN[33]	LOSO	Negative,Positive,Surprise	**79.87**	**0.8009**
Ours	LOSO	Negative,Positive,Surprise	65.63	0.65

Table 2 Performance comparisons with the existing methods on CASME dataset

Method	Protocol	Task	Accuracy(%)	F1-score
Baseline[6]	LOSO	Disgust,Surprise,Repression,Tense	40.35	0.26
LBP-SIP[43]	LOSO	Disgust,Surprise,Repression,Tense	36.84	0.33
STCLQP[11]	LOSO	Disgust,Surprise,Repression,Tense	57.31	0.50
FHOFO[13]	LOSO	Disgust,Surprise,Repression,Tense	65.99	0.54
VGGMag[44]	LOSO	Disgust,Surprise,Repression,Tense	60.23	0.58
3D-FCNN[28]	LOVO	Disgust,Surprise,Repression,Tense	54.44	-
LGCconD[47]	LOSO	Disgust,Surprise,Repression,Tense	57.31	0.54
LGCcon[47]	LOSO	Disgust,Surprise,Repression,Tense	60.82	0.60
Meta-MMFNet[24]	LOSO	Disgust,Surprise,Repression,Tense	69.59	-
GPT-4V[49]	-	Disgust,Surprise,Repression,Tense	36.93	-
SVM+LCBP-STGCN[36]	LOSO	Disgust,Surprise,Repression,Tense	**81.26**	**0.77**
Ours	LOSO	Disgust,Surprise,Repression,Tense	70.27	0.70

ing using different sets of emotions from the proposed method. From these tables, it was able to be observed that the proposed MCNet achieves competitive recognition performance compared to baselines and other existing methods. Specifically, the prediction accuracy reached 65.63%, and the F1-score reached 0.65 on the SMIC dataset, showing 1.03% improvement over the accuracy

Table 3 Performance comparisons with the existing methods on CASME II dataset

Method	Protocol	Task	Accuracy(%)	F1-score
DTCM[50]	LOSO	Happiness,Surprise,Disgust,Repression	72.06	-
Meta-MMFNet[24]	LOSO	Happiness,Surprise,Disgust,Repression	80.95	-
Ours	LOSO	Happiness,Surprise,Disgust,Repression	**81.63**	**0.81**

(64.60%) achieved by the latest Dual-ATME method.[33] For the CASME dataset, the prediction accuracy was 70.27%, and the F1-score was 0.7, indicating an improvement of 0.68% in accuracy over the Meta-MMFNet method [24] and an increase of 0.1 in F1-score over the LGCconD method [47]. Moreover, on the CASME II dataset, the prediction accuracy achieved 81.63%, and the F1-score achieved 0.81, which represents an accuracy improvement of 0.68% over the performance of the Meta-MMFNet.

–While the MCNet method on SMIC dataset was less effective than the optimal MFVAN method [33], it achieved significantly higher recognition scores than MFVAN method on the CASME II dataset. Additionally, on the CASME dataset, the SVM+LCBP-STGCN method [36] outperformed the proposed method in accuracy and F1-score by 10.99% and 0.07, respectively. However, the proposed MCNet method achieved the best recognition results for the four classification tasks of "happiness", "surprise", "disgust", and "repression" on the CASME II dataset. The evaluation results across all three datasets demonstrated that the micro-expression recognition method proposed in this study exhibited good generalization ability and achieved competitive performance on these datasets.

The results also indicated that the proposed MCNet method achieved the most promising performance on the CASME II dataset. We argued that this outcome is attributed to the balanced distribution of the four categories in the CASME II dataset. In contrast, the SMIC and CASME datasets suffered from severe class imbalance issues, which negatively impact the recognition performance. Additionally, the absence of annotated apex frames in the SMIC dataset poses a challenge. The

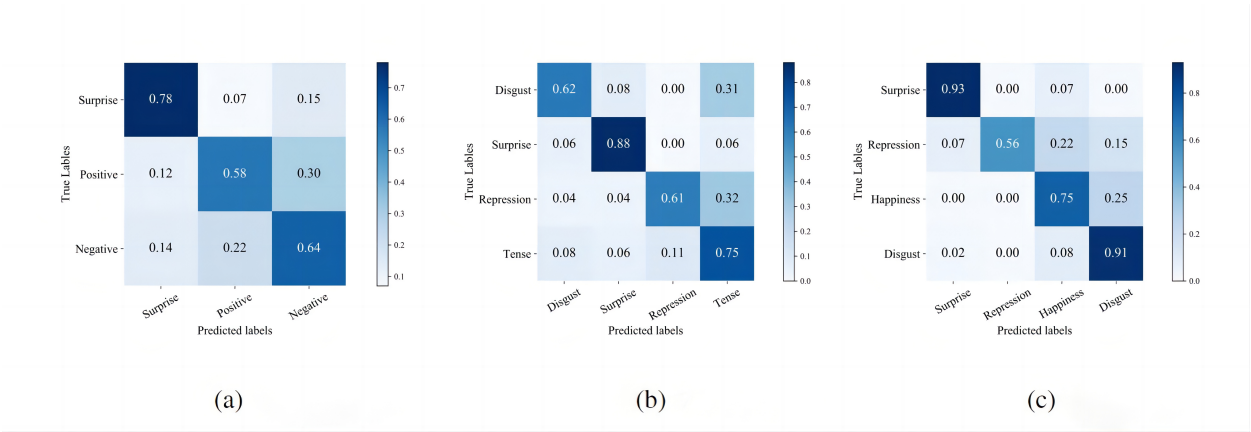


Fig 3 Confusion matrices of the proposed method on three validation datasets. (a) Confusion matrix on the SMIC dataset. (b) Confusion matrix on the CASME dataset. (c) Confusion matrix on the CASME II dataset.

apex frame, which exhibited the greatest variation in the micro-expression sequence, was crucial for determining the type and strength of micro-expressions. Although the middle frame between the onset and offset frames was used as an approximation, it introduced some errors, leading to a relatively lower recognition rate for micro-expressions in the SMIC dataset.

To further analyze the performance of the proposed method, we visualized the confusion matrices for each dataset in Fig. 3. These matrices provided insights into the recognition accuracy of each micro-expression category and the probability of mis-classification into other categories. In these figures, the horizontal axis represented the predicted class labels, while the vertical axis represented the true class labels. From the confusion matrices, it was evident that the "surprise" expression consistently achieves the highest recognition rate in all three datasets, reaching 78%, 88%, and 93% on the SMIC, CASME, and CASME II datasets, respectively. On the other hand, mis-classifications often occured between positive and negative emotions. For example, the "surprise" and "positive" expressions in the SMIC dataset were easily mis-predicted as "negative", primarily because the majority of samples belong to the negative category.

4 Conclusion

This study proposed a novel micro-expression recognition method based on the meta-clustering network, aiming to address the challenge of mis-classification caused by large intra-class distances in micro-expressions. Different from the standard centroid-based strategy used in previous meta-learning methods, the proposed method effectively took into account the variations within each class and introduced a hierarchical clustering module that produced multiple prototypes instead of the mean prototypes. This effectively reduced the impact of intra-class differences in micro-expression samples. Additionally, The proposed approach not only integrated two dynamic features but also extracted deep feature representations of micro-expressions, enabling a more comprehensive and robust understanding of the facial expressions. To alleviate the problem of limited micro-expression datasets, the proposed MCNet method leveraged transfer learning by pre-training on a macro-expression dataset with ample data and then adapting to the micro-expression target domain. Moreover, the fused priors of micro-expression and macro-expression in the meta-testing phase further contributed to improving the classification accuracy of micro-expressions. Comprehensive experiments were conducted on multiple datasets, including the SMIC, CASME, and CASME II datasets, and the comparative results highlighted the effectiveness of the proposed model in the field of micro-expression recognition.

In future works, the proposed method holds potential for enhancement. The current methodology involves pre-training deep feature extraction networks, f_{α}^{micro} and f_{β}^{macro}, on micro-expression and CK+ datasets, respectively, as an attempt to integrate prior knowledge of micro-expressions and macro-expressions. This approach hinders the end-to-end training of the proposed MCNet method. Additionally, the challenge of collecting and annotating micro-expression training data

persists, leading to issues of data insufficiency and class imbalance in recognition tasks. Although the proposed MCNet, employing few-shot learning, mitigates the problem of limited amount of data to some extent, its accuracy remains to be further improved. Furthermore, the high similarity among sub-classes in negative micro-expressions contributes to an elevated mis-classification rate. In the forthcoming research, we will introduce information on facial action unit combinations and leverage attention mechanisms to discern more nuanced negative micro-expression classes, aiming for further enhancement in recognition performance.

Data Used In This Manuscript

Availability of data used in this study is limited, which were used under licence for this study. The "SMIC" data that support the findings of the study are available on request at https://www.oulu.fi/en/university/facu and-units/faculty-information-technology-and-electrical-engineering/center-machine-vision-and-signal-analysis, the "CASME" are available on request at http://casme.psych.ac.cn/casme/c1, and the CASME II are available on request at http://casme.psych.ac.cn/casme/c2

Acknowledgment

The authors acknowledgment the support by the Natural Science Foundation of Shandong Province under the project No. ZR2023MF041, the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF) under Grants PID2020-120311RB-I00 funded by MCIN/AEI/10.13039/501100011033.

References

1 R. Arya, J. Singh, and A. Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Computer Science Review* **40**, 100399 (2021).

2 W.-J. Yan, Q. Wu, J. Liang, *et al.*, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior* **37**, 217–230 (2013).

3 T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," *British journal of clinical psychology* **45**(4), 579–583 (2006).

4 P. Ekman and E. L. Rosenberg, Eds., *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, New York (2005).

5 M. Frank, M. Herbasz, K. Sinuk, *et al.*, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *The annual meeting of the international communication association. Sheraton New York, New York City*, 1–35 (2009).

6 X. Li, T. Pfister, X. Huang, *et al.*, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, 1–6 (2013).

7 W.-J. Yan, Q. Wu, Y.-J. Liu, *et al.*, "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)*, 1–7 (2013).

8 W.-J. Yan, X. Li, S.-J. Wang, *et al.*, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one* **9**(1), e86041 (2014).

9 A. K. Davison, C. Lansley, N. Costen, *et al.*, "Samm: A spontaneous micro-facial movement dataset," *IEEE transactions on affective computing* **9**(1), 116–129 (2016).

10 S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," *IET Conference Proceedings* , 16–16(1) (2009).

11 X. Huang, G. Zhao, X. Hong, *et al.*, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing* **175**, 564–578 (2016).

12 C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, 214–223 (2007).

13 S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Transactions on Affective Computing* **10**(3), 394–406 (2017).

14 Y.-J. Liu, J.-K. Zhang, W.-J. Yan, *et al.*, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing* **7**(4), 299–310 (2015).

15 S.-T. Liong, J. See, K. Wong, *et al.*, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication* **62**, 82–92 (2018).

16 D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *2016 23rd international conference on pattern recognition (ICPR)*, 2258–2263 (2016).

17 D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 24th ACM international conference on Multimedia*, 382–386 (2016).

18 M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *2017 international conference on digital image computing: techniques and applications (DICTA)*, 1–7 (2017).

19 M. Peng, Z. Wu, Z. Zhang, *et al.*, "From macro to micro expression recognition: Deep learn-

ing on small datasets using transfer learning," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 657–661 (2018).

20 N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–7 (2019).

21 R. Zhi, H. Xu, M. Wan, *et al.*, "Combining 3d convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," *IEICE TRANSACTIONS on Information and Systems* **102**(5), 1054–1064 (2019).

22 B. Sun, S. Cao, D. Li, *et al.*, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Transactions on Affective Computing* **13**(2), 1037–1043 (2020).

23 G. Zhao, X. Huang, M. Taini, *et al.*, "Facial expression recognition from near-infrared videos," *Image and vision computing* **29**(9), 607–619 (2011).

24 W. Gong, Y. Zhang, W. Wang, *et al.*, "Meta-mmfnet: Meta-learning based multi-model fusion network for micro-expression recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **20**(39), 1–20 (2022).

25 M. Peng, C. Wang, T. Chen, *et al.*, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in psychology* **8**, 1745 (2017).

26 H.-Q. Khor, J. See, S.-T. Liong, *et al.*, "Dual-stream shallow networks for facial micro-expression recognition," in *2019 IEEE international conference on image processing (ICIP)*, 36–40 (2019).

27 H.-Q. Khor, J. See, R. C. W. Phan, *et al.*, "Enriched long-term recurrent convolutional net-

work for facial micro-expression recognition," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 667–674 (2018).

28 J. Li, Y. Wang, J. See, *et al.*, "Micro-expression recognition based on 3d flow convolutional neural network," *Pattern Analysis and Applications* **22**, 1331–1339 (2019).

29 Y. S. Gan, S.-T. Liong, W.-C. Yau, *et al.*, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication* **74**, 129–139 (2019).

30 S. Zhao, H. Tao, Y. Zhang, *et al.*, "A two-stage 3d cnn based learning method for spontaneous micro-expression recognition," *Neurocomputing* **448**, 276–289 (2021).

31 M. Tang, M. Ling, J. Tang, *et al.*, "A micro-expression recognition algorithm based on feature enhancement and attention mechanisms," *Virtual Reality* **27**, 2405–2416 (2023).

32 H. Zhou, S. Huang, J. Li, *et al.*, "Dual-atme: Dual-branch attention network for micro-expression recognition," *Entropy* **25**(3), 460 (2023).

33 H. Pan, H. Yang, L. Xie, *et al.*, "Multi-scale fusion visual attention network for facial micro-expression recognition," *Frontiers in Neuroscience* **17** (2023).

34 D. Hao, M. Zhu, C. Zhang, *et al.*, "A lightweight attention-based network for micro-expression recognition," *Multimedia Tools and Applications* , 1–22 (2023).

35 S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep3dcann: A deep 3dcnn-ann framework for spontaneous micro-expression recognition," *Information Sciences* **630**, 341–355 (2023).

36 Y. Wang, J. Han, and Z. Guo, "Lcbp-stgcn: A local cube binary pattern spatial temporal graph convolutional network for micro-expression recognition," *Journal of Intelligent & Fuzzy Systems* **44**(2), 1601–1611 (2023).

37 C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 1126–1135 (2017).

38 S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," (2016).

39 F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(1), 86–97 (2012).

40 H. Lu, K. Kpalma, and J. Ronsin, "Motion descriptors for micro-expression recognition," *Signal Processing: Image Communication* **67**, 108–117 (2018).

41 C. Wang, M. Peng, T. Bi, *et al.*, "Micro-attention for micro-expression recognition," *Neurocomputing* **410**, 354–362 (2020).

42 X. Jia, X. Ben, H. Yuan, *et al.*, "Macro-to-micro transformation model for micro-expression recognition," *Journal of computational science* **25**, 289–297 (2018).

43 Y. Wang, J. See, R. C.-W. Phan, *et al.*, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12*, 525–537 (2015).

44 Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 3094–3098 (2018).

45 Z. Xia, X. Hong, X. Gao, *et al.*, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia* **22**(3), 626–640 (2020).

46 Q. Li, J. Yu, T. Kurihara, *et al.*, "Deep convolutional neural network with optical flow for facial micro-expression recognition," *Journal of Circuits, Systems and Computers* **29**(01), 2050006 (2020).

47 Y. Li, X. Huang, and G. Zhao, "Joint local and global information learning with single apex frame detection for micro-expression recognition," *IEEE Transactions on Image Processing* **30**, 249–263 (2020).

48 L. Zhou, Q. Mao, X. Huang, *et al.*, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition* **122**, 108275 (2022).

49 Z. Lian, L. Sun, H. Sun, *et al.*, "Gpt-4v with emotion: A zero-shot benchmark for multimodal emotion understanding," *arXiv preprint arXiv:2312.04293* (2023).

50 Z. Lu, Z. Luo, H. Zheng, *et al.*, "A delaunay-based temporal coding model for micro-expression recognition," in *Computer Vision-ACCV 2014 Workshops: Singapore, Singapore, November 1-2, 2014, Revised Selected Papers, Part II 12*, 698–711 (2015).

Biography

Ziqi Wang received the B.S. degree from Shandong University of Science and Technology, China, in 2021. He is currently pursuing the master's degree with the Qingdao Institute of software, college of computer science and technology, China University of Petroleum (East China), China. His research interests include micro-expression recognition and multimedia information fusion processing.

Wenwen Fu received the B.E. degree from University of Jinan, China, in 2020. She is currently pursuing the master's degree with the Qingdao Institute of Software, College of Computer Science

and Technology, China University of Petroleum (East China), China. Her research interests include micro-expression recognition and cross-modality facial expression generation.

Yue Zhang received her Bachelor of Engineering degree from Jinan University, China in 2020 and her Master of Engineering degree from China University of Petroleum (East China) with 2023. Her research interest is micro-expression recognition.

JiaRui Li received the B.E. degree from Shandong University of Science and Technology, China, in 2023. He is currently pursuing the master's degree with the Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), China. His research interests include Multimodal fusion (images, text, etc.) and Cross-modal alignment tasks.

Wenjuan Gong received the Ph.D. (cum laude) degree from the Autonomous University of Barcelona, in 2013. She was a Postdoctoral Research Assistant with Oxford Brookes University, in 2014. She is currently an associate Professor with the China University of Petroleum (East China). She participated in the European Project Consolider Ingenio 2010 and the EPSRC Project Tensorial Modeling of Dynamical Systems for Gait and Activity Recognition. She has led the CCF-Tencent Funds and the Natural Science Foundation of China of Shandong Province. Her research interests include computer vision and machine learning.

Jordi Gonzàlez received the Ph.D. degree in computer engineering from the Universitat Autònoma de Barcelona (UAB), in 2004. He is currently a Professor in computer science with the Department of Computer Science, UAB. He is also a Research Fellow with Computer Vision Center, where he has cofounded three spin-offs, namely Cloud Size Services, Visual Tagging, and Care Respite, and the Image Sequence Evaluation (ISE Lab) Research Group. His research interest includes machine learning techniques for the computational interpretation of social images, or visual hermeneutics.

List of Figures

1 The process of predicting sample categories using the proposed meta-clustering method. Instead of using a single mean feature prototype for each class, the method determines the class prototypes by clustering the two nearest samples within the same class in the support set, resulting in N or more class prototypes, where N represents the number of classes in the support set. To predict the label of a query sample X during meta-testing, the meta-clustering method compares the distances between the sample and each class prototype and identifies the closest class prototype to the sample. The label of the test sample is obtained by mapping the nearest class prototype back to its original category.

2 The overall architecture of the meta-clustering learning network. This method comprises of two main stages: pre-training the deep feature extraction module and learning the meta-clustering classification model. In the pre-training stage, two independent deep feature extraction networks based on ResNet18 are trained on the micro-expression dataset Ω_{train} and the macro-expression dataset CK+. In the meta-clustering stage, the fused dynamic features are fed into the micro-pipeline and macro-pipeline and the deep features of micro-expressions are then extracted in each pipeline. The cosine distance between the query samples and the class prototypes obtained from the class prototype calculation module is computed, yielding the micro-score and macro-score, respectively. Based on these two score vectors, we optimize the micro-pipeline and macro-pipeline separately using cross-entropy loss. during the validation process, we combine the score vectors generated from both pipelines to merge the micro- and macro-predictions to predict the final micro-expression category.

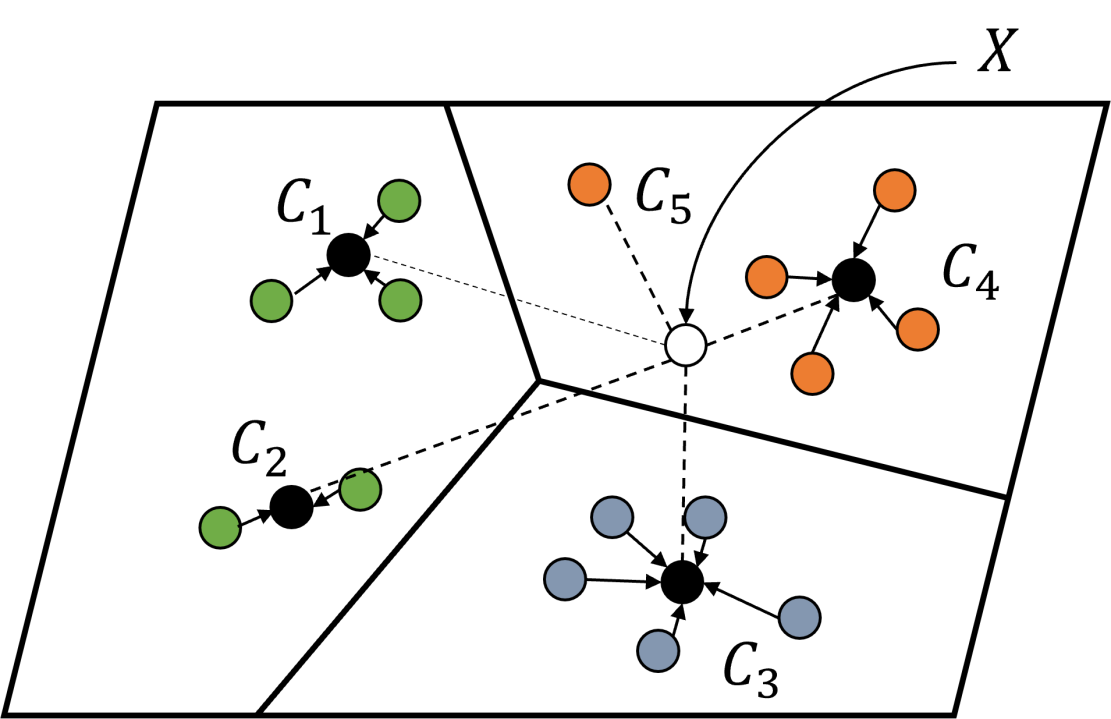
3 Confusion matrices of the proposed method on three validation datasets. (a) Confusion matrix on the SMIC dataset. (b) Confusion matrix on the CASME dataset. (c) Confusion matrix on the CASME II dataset.

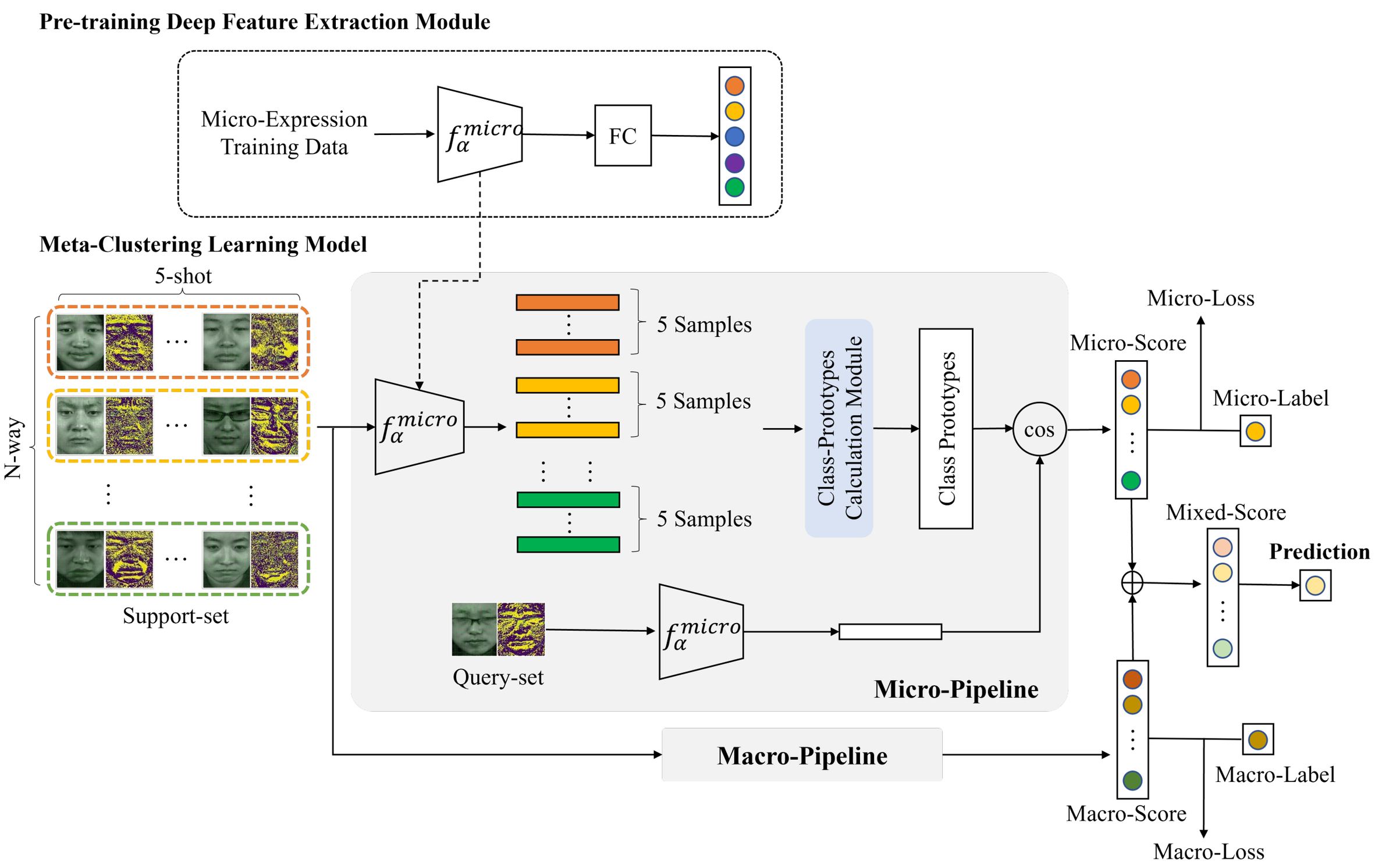
List of Tables

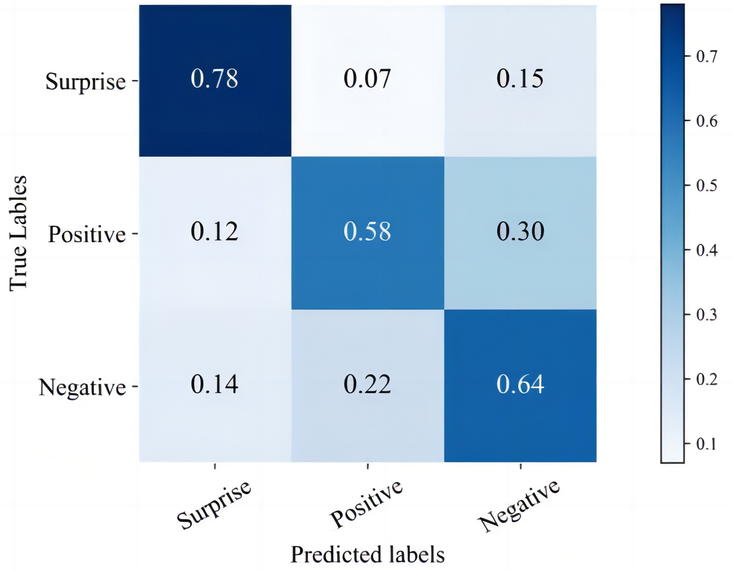
1 Performance comparisons with the existing methods on SMIC dataset

2 Performance comparisons with the existing methods on CASME dataset

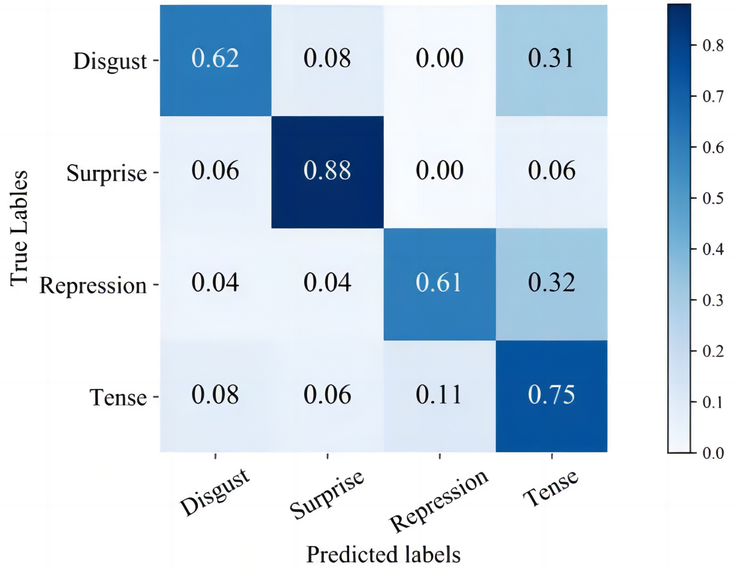
3 Performance comparisons with the existing methods on CASME II dataset


X
C_1
C_2
C_3
C_4
C_5

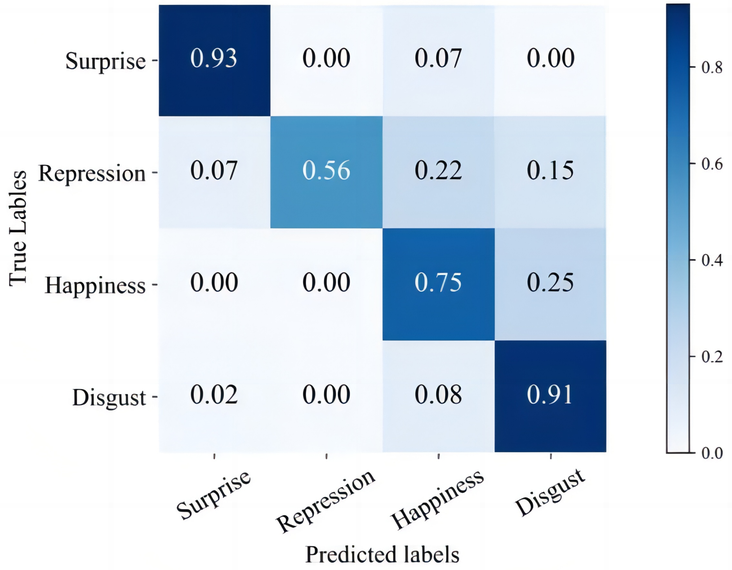

Pre-training Deep Feature Extraction Module
Micro-Expression Training Data
f_α^{micro}
FC
Meta-Clustering Learning Model
5-shot
N-way
Support-set
5 Samples
5 Samples
5 Samples
Class-Prototypes Calculation Module
Class Prototypes
cos
Query-set
Micro-Pipeline
Macro-Pipeline
Micro-Loss
Micro-Score
Micro-Label
Mixed-Score
Prediction
Macro-Label
Macro-Score
Macro-Loss


Surprise
Positive
Negative
True Lables
0.78
0.07
0.15
0.12
0.58
0.30
0.14
0.22
0.64
Surprise
Positive
Negative
Predicted labels
0.7
0.6
0.5
0.4
0.3
0.2
0.1

(a)


Disgust
Surprise
Repression
Tense
True Lables
0.62
0.08
0.00
0.31
0.06
0.88
0.00
0.06
0.04
0.04
0.61
0.32
0.08
0.06
0.11
0.75
Disgust
Surprise
Repression
Tense
Predicted labels
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(b)


Surprise
Repression
Happiness
Disgust
True Lables
0.93
0.00
0.07
0.00
0.07
0.56
0.22
0.15
0.00
0.00
0.75
0.25
0.02
0.00
0.08
0.91
Surprise
Repression
Happiness
Disgust
Predicted labels
0.8
0.6
0.4
0.2
0.0

(c)